Quality Assurance 344

ECSA Graduate Attribute 4 – Data Analysis, Statistical Process Control, and Service Optimisation

A Report by Ella Potterton

Student Number 26960176

24 October 2025

Table of Contents

Introduction

The following report combines the work completed for the module Quality Assurance 344 to meet the requirements of the Engineering Council of South Africa's Graduate Attribute 4 (GA4). These requirements require students to conduct investigations, analyse data and design experiments. Explorative analysis should be conducted through understanding data and visual representations of this data.

Firstly, a basic data analysis conducted on data describing the performance of a technology company, followed by a Statistical Process Control and a process capability analysis assessing the company's delivery process. Next, process risk quantification was performed before discrepancies in the original datasets were corrected and reanalysed. It was then discovered how statistical methods can be applied to optimise profit gained by a company. Hypothesis testing was conducted to identify potential areas for process improvement based on relationships in the data. Finally, the reliability of a process was investigated using a statistical approach.

1. Data Analysis

This section covers basic data analysis for a technology company that sells software, Cloud subscriptions, laptops, monitors, keyboards, and computer mice. This analysis has been done to assist upper management in making informed business decisions pertaining to costs, profitability, efficiency, and customer requirements.

The investigation was approached systematically and in line with quality assurance principles, ensuring data was used effectively to extract insights regarding business performance and opportunities for improvement. To achieve this, the raw data was loaded an inspected, summary statistics were computed, missing values were identified and handled, data was filtered and sub-set, visualizations were created, and key relationships between variables were explored. The analysis was conducted through RStudio, and the findings are documented in the report.

1.1. Data Loading and Inspection

Four datasets were loaded for the analysis. The columns of each dataset represent variables that describe each instance, or row, of the dataset. These variables can be classified as either numeric (if they are of type integer or numeric) or as categorical (if they

are of type character). A table describing all the variables found in the datasets is given below:

Variable Name	Variable Type
CustomerID	Character
Gender	Character
Age	Integer
Income	Numeric
City	Character
ProductID	Character
Category	Character
Description	Character
SellingPrice	Numeric
Markup	Numeric
Quantity	Integer
orderDay	Integer
orderMonth	Integer
orderYear	Integer
pickHours	Numeric
deliveryHours	Numeric

It is evident that the files are linked by the variables customer ID and product ID.

The customer_data.csv file contains information pertaining to individual customers. The dataset contains 5000 rows, with each row representing an individual customer. There are no missing values, nor are there any duplicate customer IDs present.

The products_data.csv file contains information pertaining to individual products. The dataset only contains 60 rows, with each row representing an individual product. There are no missing values or duplicate product IDs present.

The products_Headoffice.csv file appears to mirror the products_data dataset. However, this dataset contains 360 rows and 250 duplicate IDs. The differing sizes between the datasets suggest that products_data represents a subset of the company's entire product catalogue, whereas products_Headoffice represents the company's entire product catalogue. Upon closer inspection, the repeated product IDs have the prefix 'NA', suggesting an error occurred when recording the IDs of these products in the catalogue.

The sales2022and2023.csv file contains information pertaining to individual sales made by the company over the years 2022 and 2023. The dataset contains 100000 rows, , no missing values and 95000 duplicate IDs. Upon further investigation, it is evident that these duplicate IDs are a result of repeat customers, or a product with the same product ID being bought multiple times. They do not reflect errors in the data, and must therefore remain in the dataset.

1.2. Summary Statistics

Summary statistics were computed for each dataset. Numeric variables are described by there mean, standard deviation, range, skewness, and Kurtis is. Categorical variables are described by the percentages indicating the portion of the dataset made up by their values. Before further discussion of the summary statistics, it is important to note a small or low variance or standard deviation cannot be explicitly defined, as it is dependent on the type of data being considered (Bobbitt, 2021). Furthermore, skewness values with a magnitude between 0 and 0.5 are almost symmetrical, show moderate skew between 0.5 and 1 show, and show significant skew for magnitudes greater than 1 (Menon, 2022). Additionally, Menon (2022) also highlights that negative skewness values show data is skewed to the left, whereas positive skewness values show data is skewed to the right.

The first dataset that was investigated further was customer_data. It was found that the customer age had a mean of 51.55 years of age and a median of 51 years of age, a value close to the mean. However, the age of the customers vary considerably as the standard deviation was calculated as 21.22 years, while values ranged from 16 years old to 105 years old. Customer age had a skewness value of 0.2 and was thus slightly skewed to the right. This indicated that the data had a longer tail to older ages which pulled the mean up slightly. Conversely, the income variable was slightly skewed to the left, having a skewness

value of -0.21. This led to the income mean being pulled down slightly, having a value of $80 797. The income values ranged from a minimum of $5 000 to a maximum of $14 000. Both of these numeric variables had a negative kurtosis, thus it can be concluded that the distributions are relatively flat. This suggests fewer outliers than what would be predicted in a typical normal distribution. The customer base is also most perfectly balanced, with 48.6% of the customers representing females and 47% representing males. The remaining 4.4% fell into the ‘other’ category. Furthermore, city distribution shows a relatively even spread of customers across main cities, with Chicago, Houston, and Los Angeles accounting for 14.5% each, and Miami making up 12.9%. Smaller cities are underrepresented.

When summarising products_data, it was discovered that selling prices are heavily skewed to the right, with a skewness value of 1.43. This skew can be further seen by the mean of $4 495.59, which is much greater than the median of $794.19, thus it is clear that a few very expensive products pull the value of the mean right up. Selling prices vary greatly, ranging from $350.45 to $19 725.18 with a standard deviation of $6 503.77. Product markup is much more consistent, seen by its standard deviation of only 6%, and it’s smaller range of 10.13% to 29.84%. Furthermore, markup is approximately symmetric, with a skewness value of only -0.036 and it’s mean of 20.46% being very close in value to the median of 20.33%. Product IDs are all unique, but they are not relevant for analysis since they are only identifiers. The dataset includes six equally represented categories: Cloud Subscription, Keyboard, Laptop, Monitor, Mouse, and Software. Each category accounts for 16.7% of the products, ensuring balanced representation across types.

Since products_data is presumably a subset derived from products_Headoffice, it can be expected that their summary statistics should be identical, if not they should at least show great similarities. These similarities were confirmed. The mean selling price was $4 410.96 and the mean markup is 20.39%. Once again, this dataset illustrates a large variability in selling price, with a standard deviation of $6 463.82. Markup remains relatively stable, with a standard deviation of 5.67%. The minimum selling price is $290.52 and the maximum is $22 420.14, with a median of $797.22, which is much lower than the mean. The markup ranges from 10.06% to 30%, with a median of 20.58%. Selling price is right-skewed (1.53) while markup is nearly symmetrical (-0.048). Identically to the products_data, this dataset includes six equally represented categories: Cloud Subscription, Keyboard, Laptop, Monitor, Mouse, and Software. Each category accounts for 16.7% of the products, ensuring balanced representation across types. However, since all the summary statistics

are not identical between the two datasets, it is possible that discrepancies exist between them. These should be further investigated.

Finally, the summary statistics for sales2022and2023 were computed. As previously mentioned, the date of the sale was split into four separate variables that described the time, day, month, and year of sale, respectively. As expected, the order year range covered the years 2022 to 2023, the order month range covered the months 1 to 12, or January to December, the order days ranged from 1 to 31, and the order time ranged from 1 to 23. Approximately half the orders are from 2022 while the other half are from 2023. There is a small skewness value of 0.15 and a standard deviation of 0.5, indicating that the dataset was balanced across the years. Additionally, orders are evenly spread across the months of the year. The mean month is 6.45, which lies close to the median month 6. There is an insignificant skewness value of 0.007 and a small standard deviation of 3.28 months. Similarly, the day of the month orders are placed on was roughly uniform, with an insignificant skewness value of 0.003. The mean is day 15, which once again is close in value to the median day 15.5. Once again, there is a small standard deviation for this variable, calculated as 8.65 days. The order time was also discovered to be roughly symmetric with a mean of 12.93 and a median of 13. The standard deviation was calculated to be 5.5, and the skewness value was -0.23. This indicates the data was skewed slightly to the left, thus there were more orders in the afternoons and evenings than in the mornings. Furthermore, the order quantity data is skewed to the right, with a skewness value of 1.04 indicating that most orders are small in size. However, the standard deviation of 13.76 indicated that the order quantities varied greatly. This was further highlighted by the data having a mean order quantity of 13.5, but a median order quantity of 6. The smallest order quantity was one product, most likely to an individual end customer, while the largest order quantity was 50 products, most likely to another store where the products were resold, or a company that required numerous technological devices. This data demonstrates that the technology company likely focuses on sales to individual customers. Picking times are skewed to the right with a value of 0.74, with the average picking time for orders calculated at 14.7 hours. The picking times range from 0.43 hours to 45.06 hours and a large standard deviation of 10.39 hours indicated that the picking times have a high variability. Delivery times also vary greatly, with a standard deviation for the variable calculated to be 10. The data is slightly left-skewed with a skewness value of -0.47, which demonstrated that there are a greater number of deliveries that are faster than the mean time of 17.48 hours that there are deliveries that are slower.

This idea is enhanced by the fact that the mean delivery time is less than the median delivery time of 19.55 hours, which is because a larger number of lower values pull it down.

1.3. Data Filtering and Subsetting

Data was grouped in three main ways: sales by year, sales by income group, and sales by product category. Additionally, a calculation to calculate revenue (Quantity * SellingPrice) was included during the grouping of sales by product category. This derived variable give the company insight into the amount of income each of their products generate. Furthermore, log-transformations were applied to selling price and revenue, due to their right-skewed distributions and large variations. This makes it easier to identify trends in the data.

Sales by year allow for the detection of trends. The rise or fall of sales can be identified, as well as seasonal patterns. This aids management in determining where their resources should be focused.

Sales by income group highlight purchasing behaviour differences amongst the company's customers. It indicates what the company should emphasize to their customers, depending on their income. It also indicates what products attract greater revenue for the company and must thus be focused on.

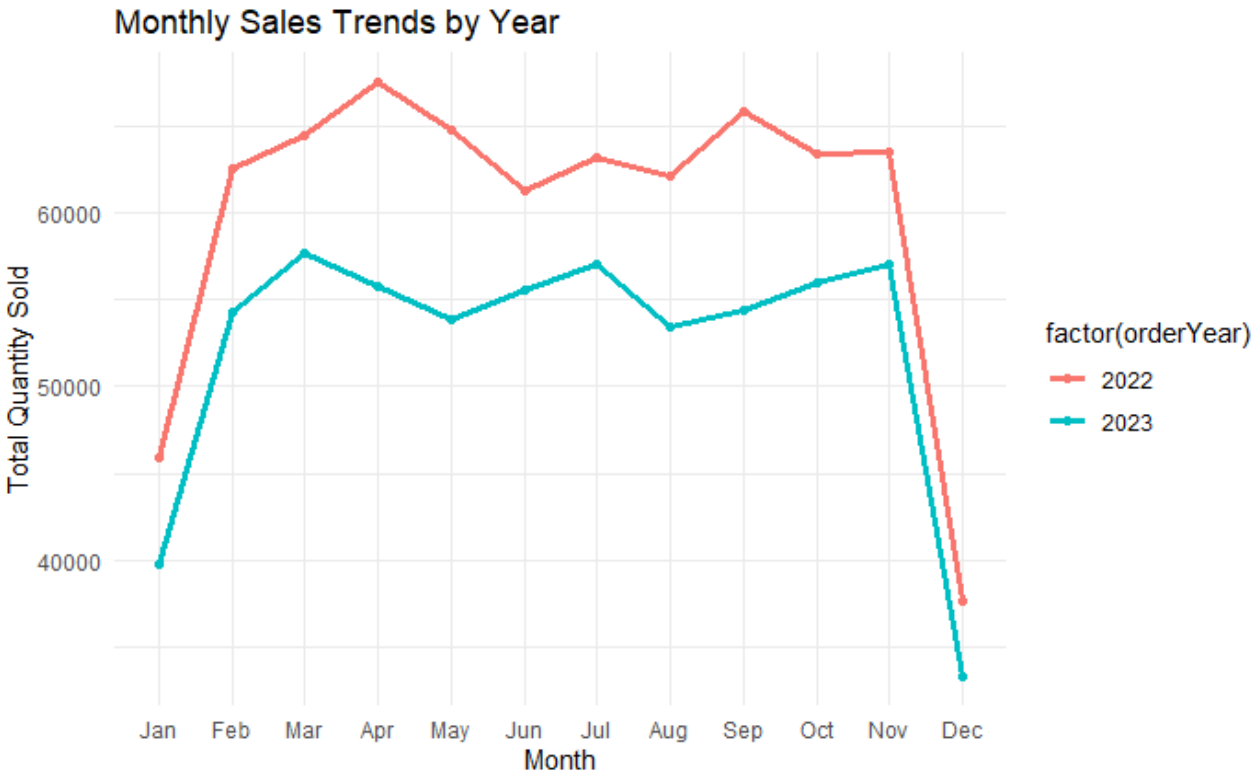
Sales by product category can be used to identify best sellers and revenue contribution. It also enables further investigation into the variation of delivery times between product categories.

These groupings are useful for future visual analysis.

1.4. Data Visualisation

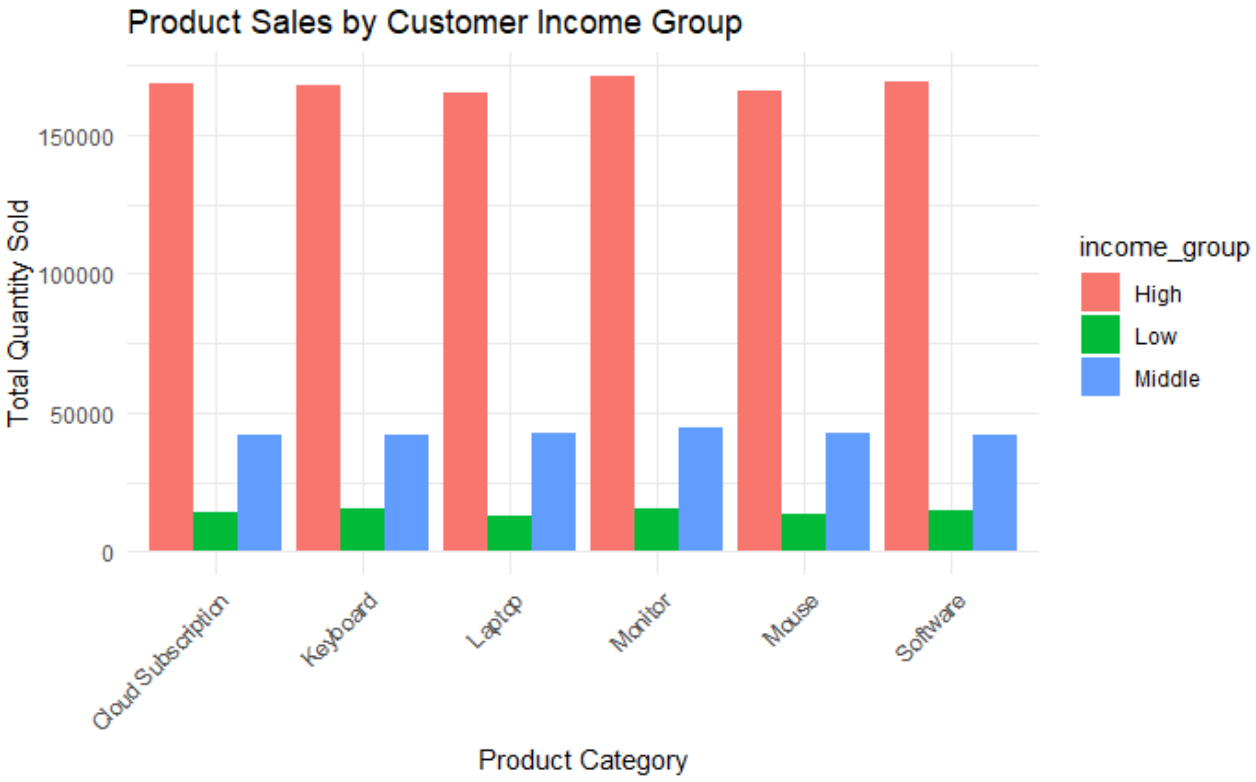
Data visualisation is an important component of data analysis. It allows data to be understood more thoroughly, uncovering patterns and trends within the data that may be hidden when viewing numerical statistics alone. These visual representations highlight information that can be used to enhance business performance.

The quantity of products sold is analysed for each month across 2022 and 2023.



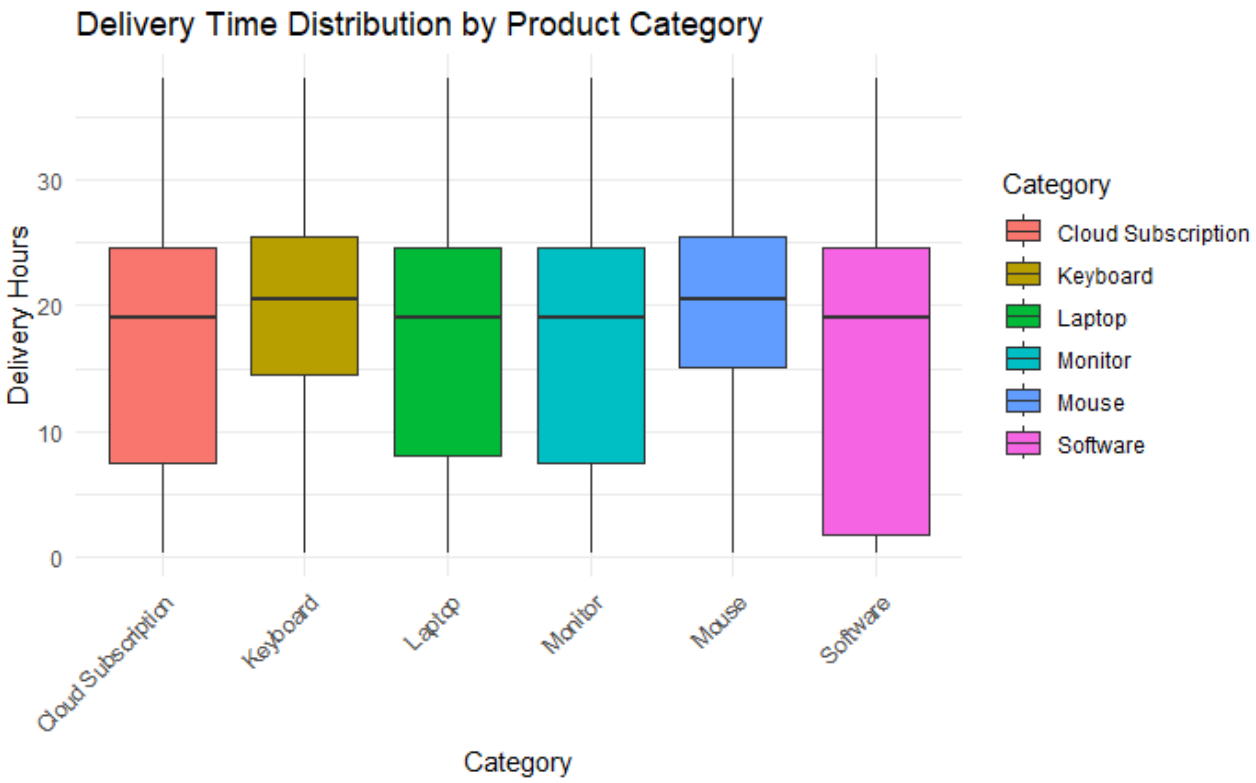
A similar pattern in the quantity of products sold each month can be seen in both year 2022 and year 2023, indicating seasonality that can be forecasted. Both years sell low quantities in December and January, with December seeing the lowest quantities sold. The quantities sold in 2022 peak in April, while the peak in 2023 is in March. Overall, the quantities sold decreased from 2022 to 2023. Previous years should be analysed to see if this decrease is consistent, as it could indicate a gradual decline in the company's success. Potential reasons for the decrease seen between the quantities sold over the two years include an economic crisis faced by the customers, poor marketing strategies, or a lack in updating products, which creates a lack of continuous demand.

Sales were grouped by income and product category and analysed.



The bar chart illustrates that the majority of the company's products are sold to customers within the high income group (greater than or equal to $60 000 a month) since they dominate sales across all categories. Their smallest customer group fall in the low income (less than $30 000 a month). This suggested that the products sold by the company are high in value and therefore typically not accessible to everyone. The quantities sold to each income group remain consistent across product categories.

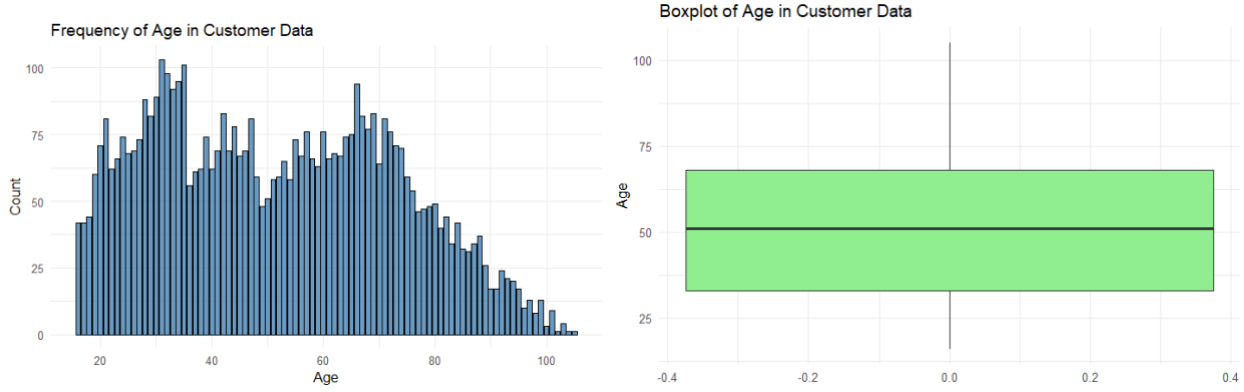
Boxplots showing the distribution of delivery times across product categories can be used to assess consistency and identify potential differences in performance.



These box plots show that all product categories have similar median delivery times of approximately 20 hours, however the variation of delivery times per product category is not consistent. This implies that issues with the delivery process may be related to the category of product being delivered. Different techniques may be used in delivering products from different categories, which results in this evident variability. If this is the case, the company should consider standardising their delivery process to achieve more consistent delivery times across product categories. The ‘software’ category has the most variability in delivery times, seen by the fact that it has the largest interquartile range. This should be investigated, since software typically does not have to be physically delivered, thus the delivery times are expected to be fast and consistent. The ‘mouse’ category has the least variability in delivery times, seen by the fact that it has the smallest interquartile range. No extreme outliers are evident, implying that most of the delivery times fall within the expected range across all categories.

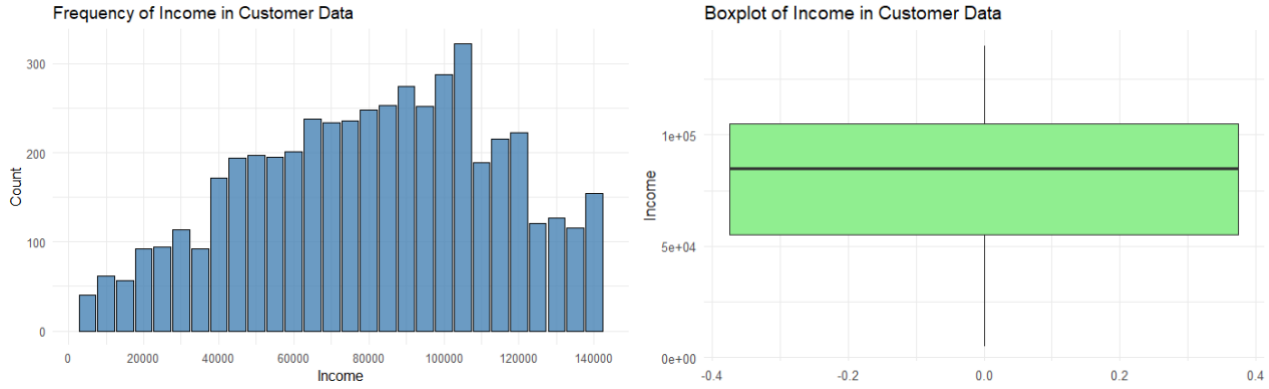
The following charts help visualise the distribution of the variables in each dataset. These can be used to confirm the conclusions drawn from the calculated summary statistics.

Age:



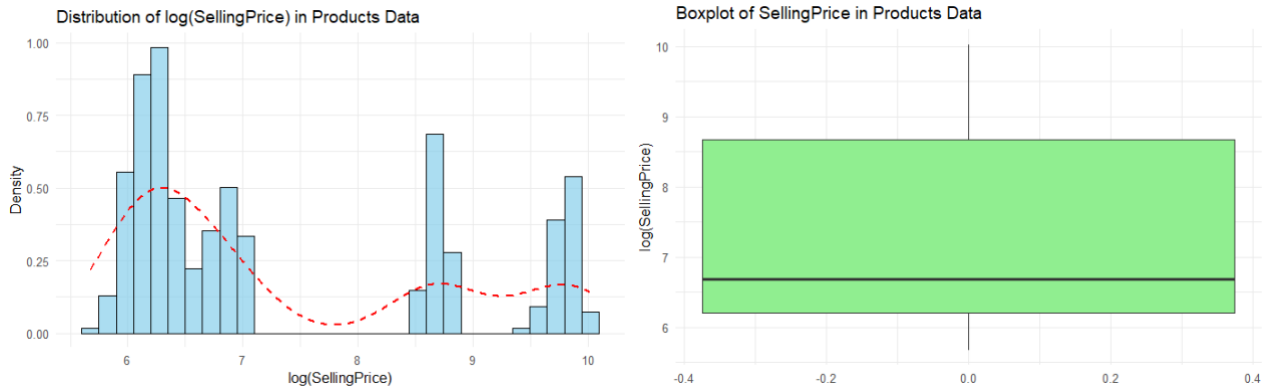
These visuals confirm that the age data has a greater spread of older values, skewing it slightly to the right. This is illustrated by the greater range of ages above the median age in the frequency plot, and backed up with the longer tail at the top of the boxplot. The population being analysed contains a lot of younger customers and fewer older customers, refining the target market of the company.

Income:



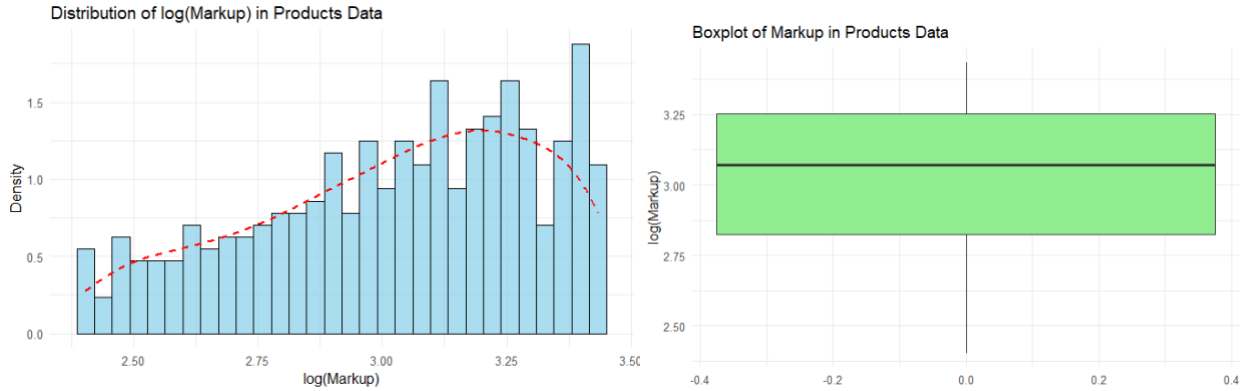
It is evident that the income data is skewed slightly to the left. The frequency plot has a greater spread of lower values and the boxplot also shows a greater range of values below the median. This suggests that the company's customers are higher-income individuals.

Selling price:



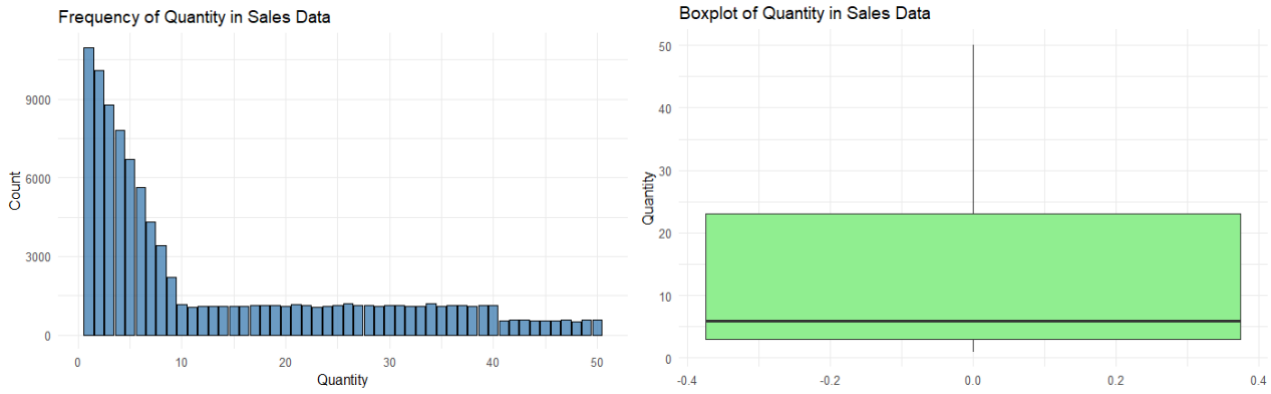
A log transformation was applied to the selling price variable due to its highly skewed distribution. This makes it easier to visualise and interpret patterns by eliminating the very long tail on the high end, effectively compressing the scale of values and reducing the impact of large outliers. The visualisation highlights that selling price does in fact have a significant skew to the right as there is a clustering of lower values and fewer much higher values. Most products have lower prices; however, the company sells a few high-end products.

Markup:



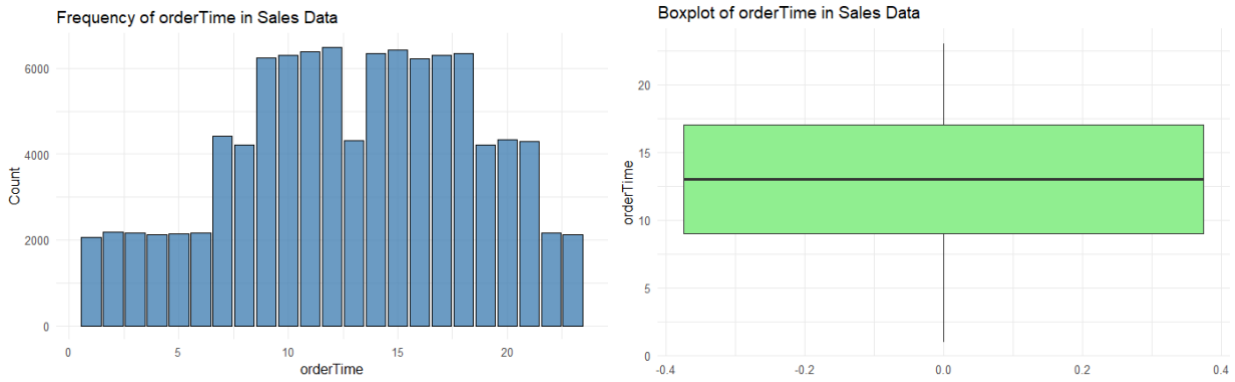
Similar to selling price, a log transformation was applied to the markup variable. However, this is simply because the log transformation was applied to all continuous numeric variables in the sales2022and2023 dataset. It does not change the distribution of these variables, but rather makes highly skewed variables' values more comparable. Thus, the log transformation was not removed from the markup variable. However, these plots do show that the data is slightly skewed to the left, but is almost symmetrical. The company typically has high-margin products, but a few products have much lower markups.

Quantity:



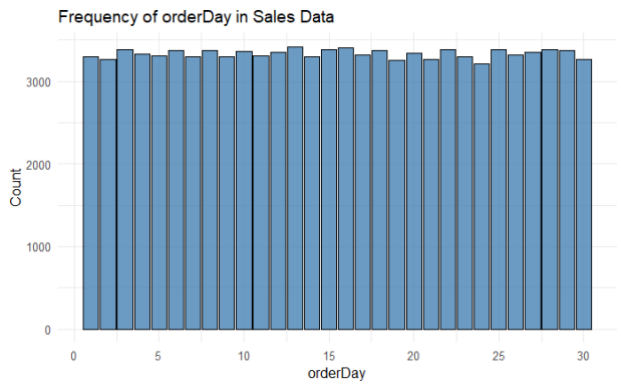
Quantity is significantly skewed to the right. The visuals portray a much larger range between the median and the maximum quantity. It is evident that most customers purchase products in small quantities, however a few customers purchase products in bulk. These could be organizations such as schools or businesses, or it could be smaller companies buying products to resell them.

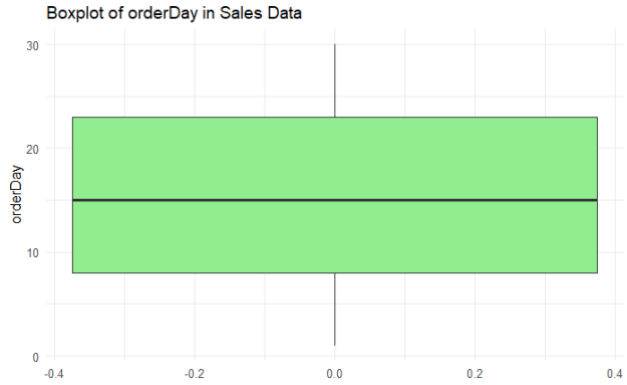
Order time:



The data is slightly skewed to the left, with fewer orders being placed in the early hours of the morning. Other than that, the number of orders in an hour remains relatively consistent. There is a dip in order number around midday, which is more than likely due to the fact that most customers and possibly employees are on a lunch break. If fewer employees are working during this period, the productivity of the company can be expected to drop, thus it makes sense for fewer orders to be processed.

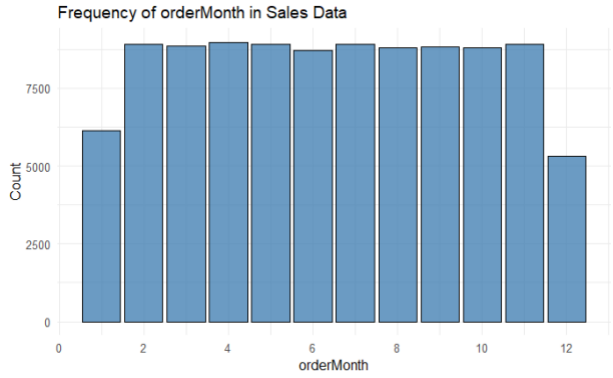
Order day:

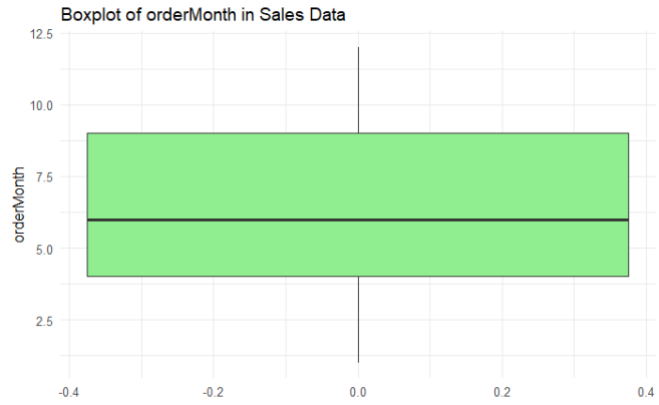




The order day data is uniformly distributed. No days appear to be more popular order days. This suggests that the company operates every day of the week, including weekends.

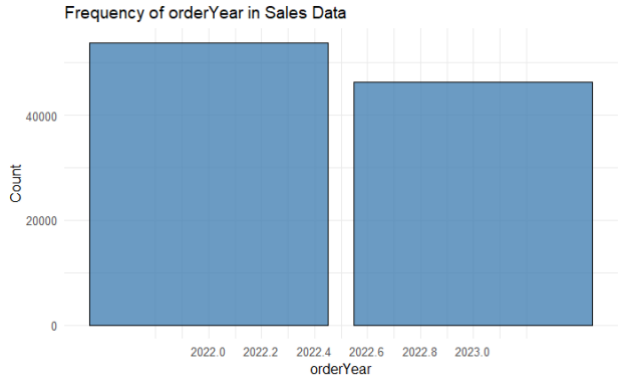
Order month:





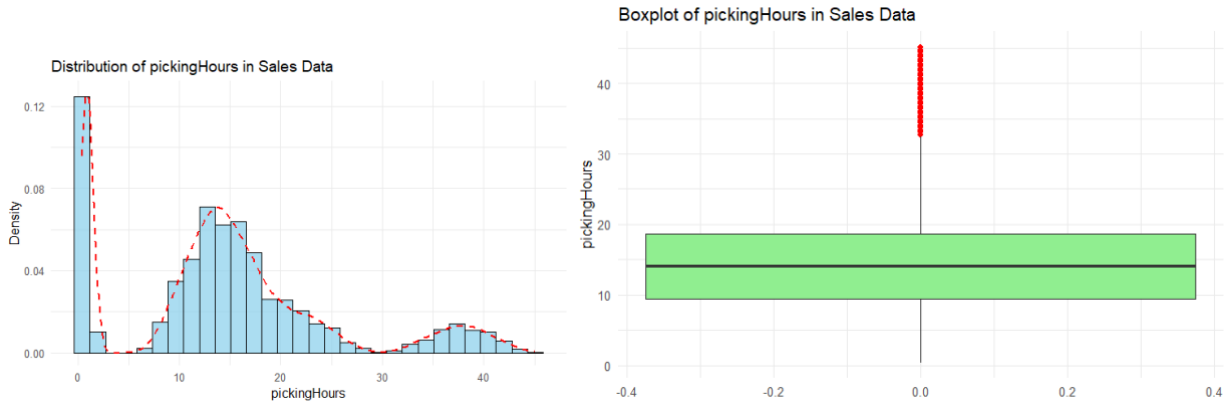
The visuals of order month data confirm the consistency of orders throughout the year, apart from January and December. This information correlates well with the line plot illustrating the quantity ordered per month.

Order year:



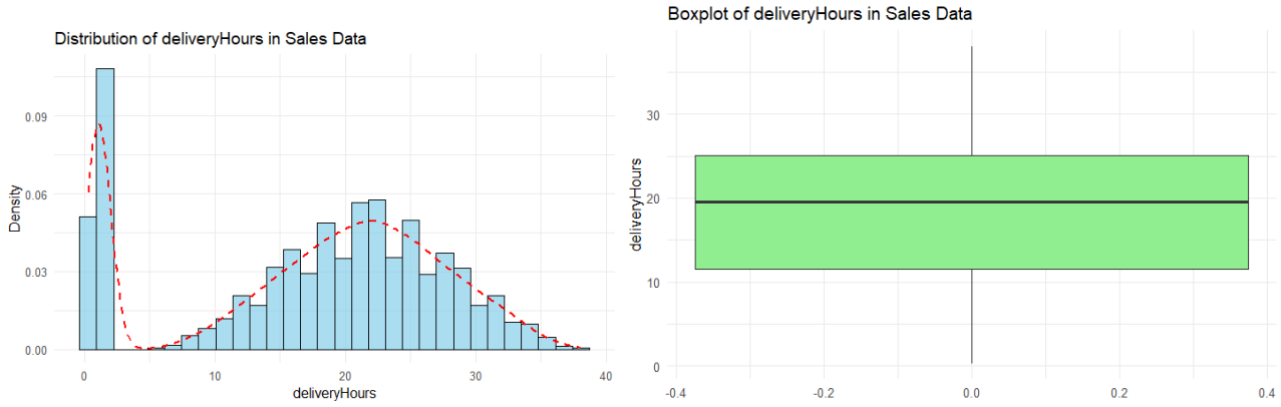
Similar to the line plot illustrating the quantity ordered per month, this visual emphasizes the fact that orders drop from year 2022 to year 2023. This drop should be investigated in order to find a way in which the company can be improved. Greater orders lead to greater profit, and thus success of the company.

Picking hours:



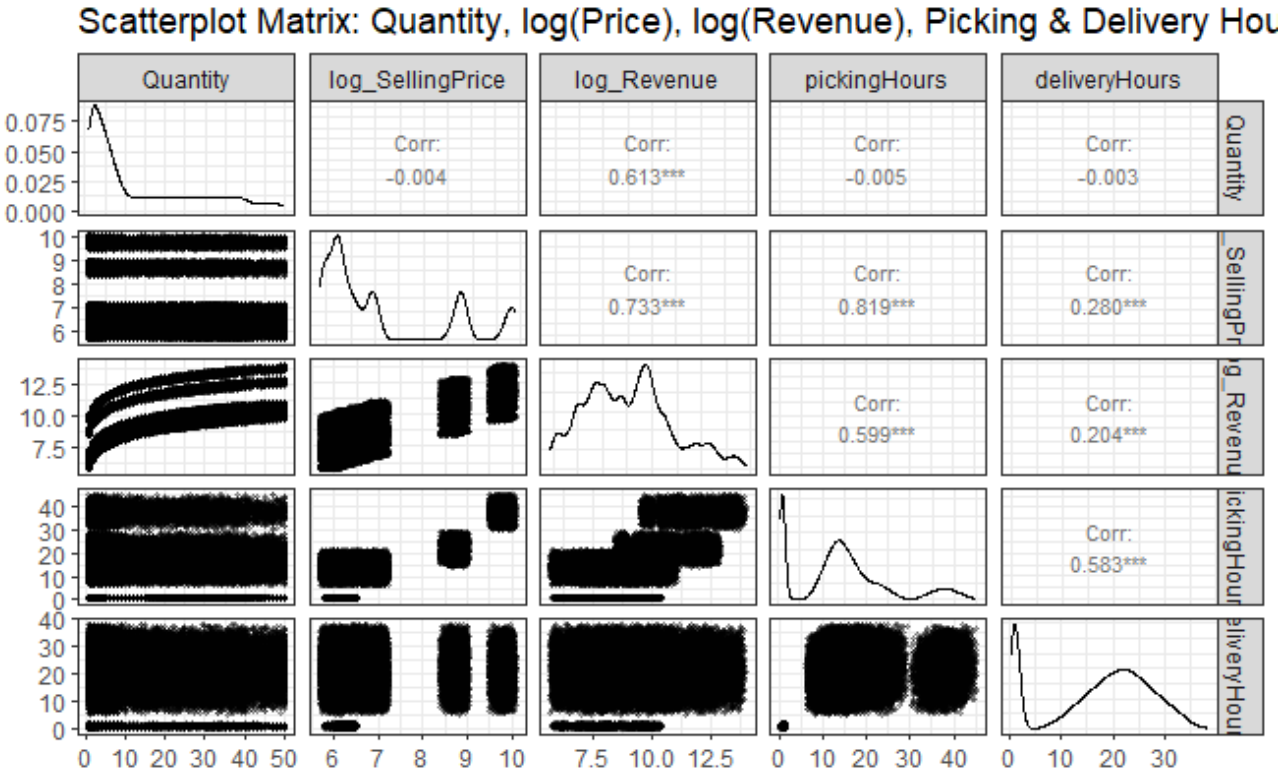
These visuals confirm that picking orders are skewed to the right. In fact, the red portion of the boxplot indicates outliers in the data. These extremely high picking hour values bring the mean up, causing the mean to be an unreliable measure of central tendency. More products have a much quicker process of removing them from inventory, but a few products appear to have inefficiencies in this process, causing it to take significantly longer.

Delivery hours:



It is evident that the delivery hours are slightly skewed to the left. The spike in the frequency of short delivery times could possibly be due to a different process being used to deliver a certain product category to customers.

1.5. Exploring Relationships

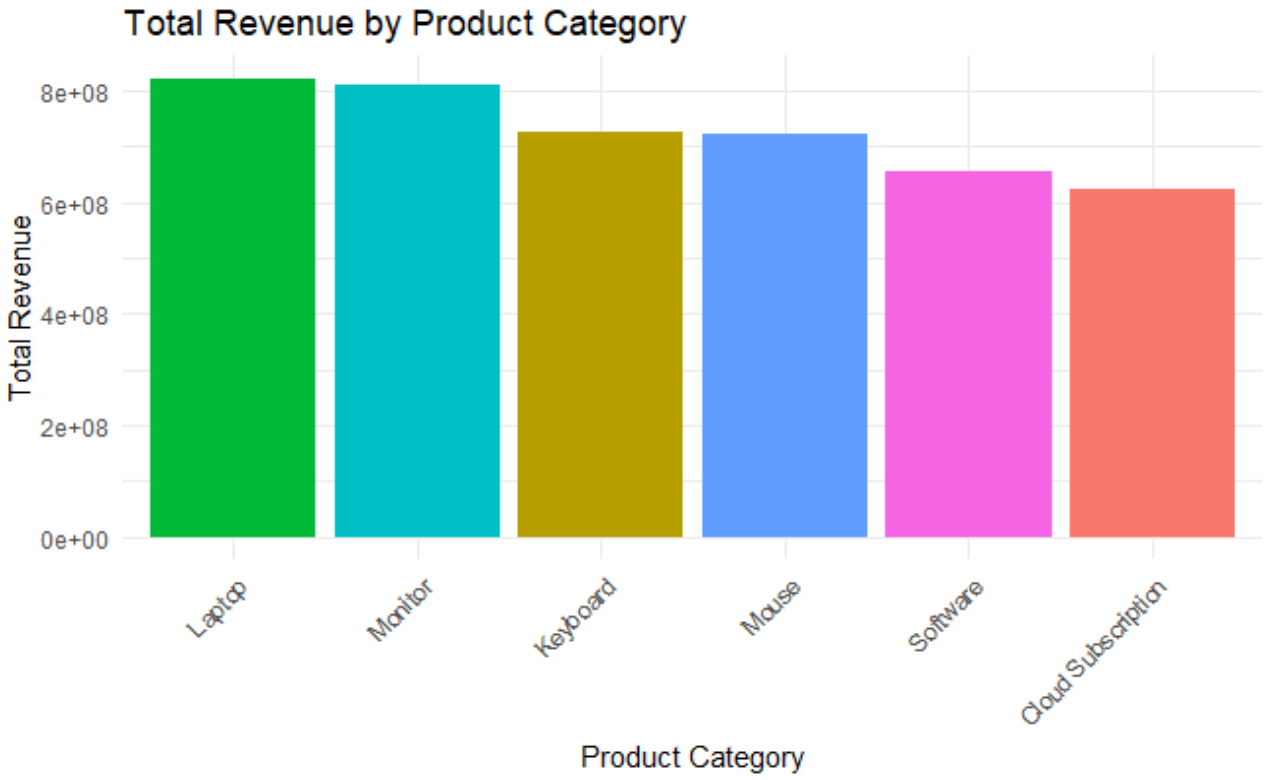


A scatterplot matrix was generated. Along with numeric correlations, the scatterplot matrix also provides a visual representation of the correlation between these variables, confirming the interpretations given in the table below:

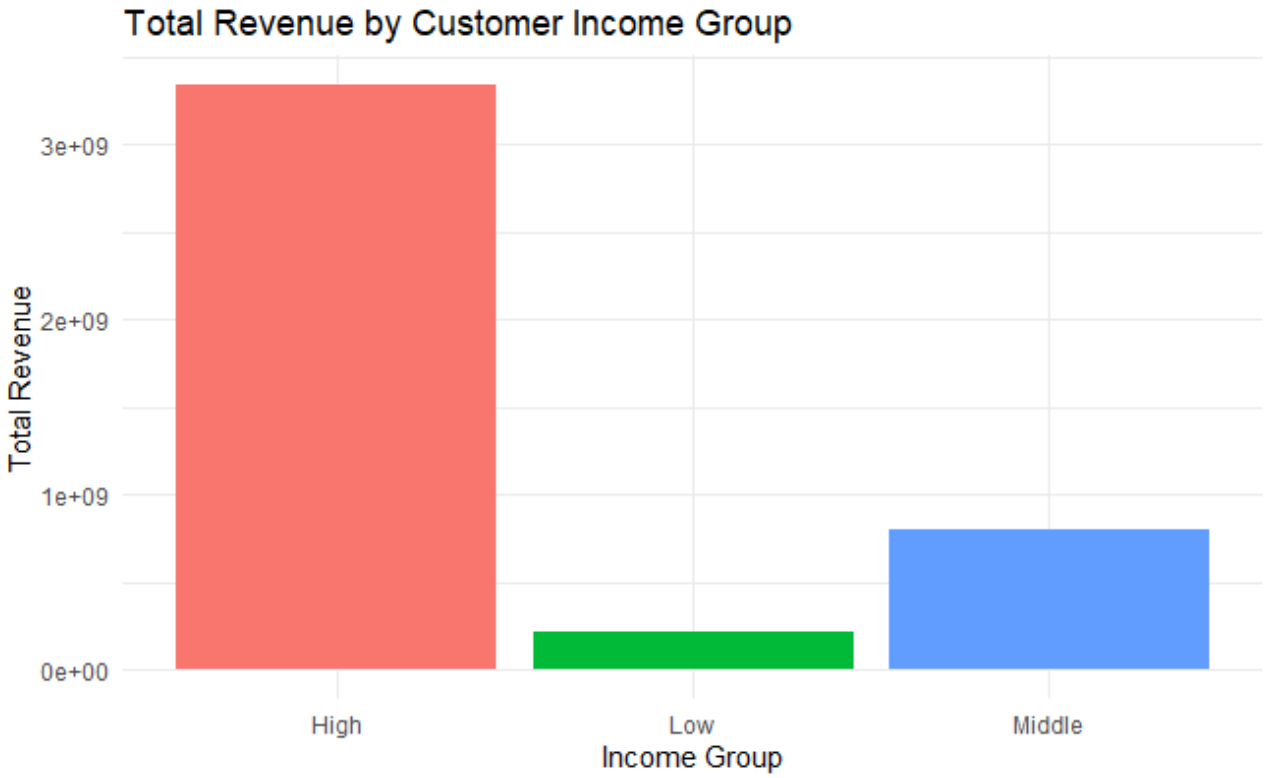
Relationship	Correlation	Interpretation
Quantity ↔ log_Revenue	0.613	Strong positive correlation, indicating higher quantities result in higher generated revenues.
log_SellingPrice ↔ log_Revenue	0.733	Strong positive correlation, indicating higher selling prices result in higher generated revenues.
log_SellingPrice ↔ pickingHours	0.819	Very strong positive correlation, demonstrating more expensive products take longer to pick.
pickingHours ↔ log_Revenue	0.599	Moderate positive correlation, suggesting that higher revenue orders may require greater picking time.
deliveryHours ↔ log_Revenue	0.204	Weak positive correlation, suggesting that high-value orders might have longer deliver times.
Quantity ↔ pickingHours	≈0	No correlation, demonstrating picking time is not influenced by the number of products ordered.
Quantity ↔ deliveryHours	≈0	No correlation, demonstrating delivery time is not influenced by the

Relationship	Correlation	Interpretation
		number of products ordered.

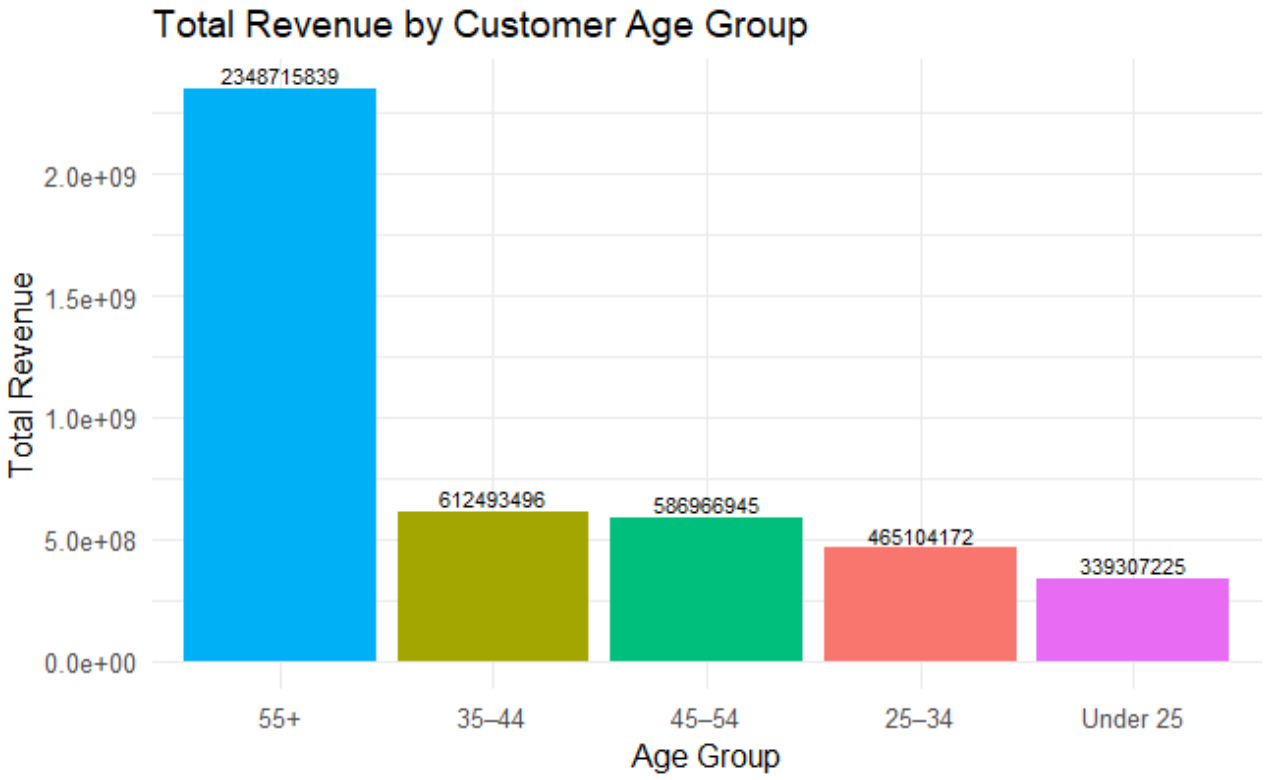
The following charts were created to help management identify areas that are most profitable to them, as well as to help them identify areas in which they can expand.



The chart depicting the total revenue per product category indicates that all of the company's products generate approximately the same amount of revenue, with laptops generating the most and cloud subscriptions generating the least. The fairly consistent amount of revenue across product types suggests that the company does not specialise in a single product, but rather spreads their resources evenly across all categories. A potential method to boost the company's success would be to focus their resources on their high-value products, such as laptops and monitors. While this may lead to a decrease in the revenue generated by the other product categories, the company may see greater profits by selling a higher quantity of products with much larger selling prices.



As expected from the products sales by income group chart, the majority of the company's revenue comes from customers falling into the high income group. As previously mentioned, the reason for this is most likely due to the fact that the company sells expensive products, most of which only high earners can afford. This is further incentive to focus on selling more high-value products, since their main customer market will be able to afford them.



This chart demonstrates that customers older than 55 years generate the most amount of revenue. This is likely due to the fact that there is a much greater frequency of customers above the age of 55 than there are in each of the age groups, as seen by the frequency of age in customer data chart. Customers younger than 25 years generate the least amount of revenue, which could be due to the fact that this age group has the lowest number of customers in it. Additionally, younger customers probably earn lower incomes and are thus unable to purchase large quantities of high-value products. This could explain why the revenue generated appears to increase with age.

1.6. Discussion

Basic data analysis enables greater understanding of business data, allowing for better informed decisions regarding company performance to be made. It illustrated that the target market of the company is younger individuals with high incomes. It also highlighted potential room for improvements, such as standardizing delivery processes and focusing resources on higher value products. However, before the improvements are implemented, discrepancies between the datasets should be corrected and further investigations should be conducted. This includes understanding the delivery process, or processes, of the business on a qualitative level, not just a quantitative level.

2. Statistical Process Control

This section applies Statistical Process Control (SPC) to the delivery-time data for each product type. This monitors whether the delivery-time process for each product type remains stable and consistent over time, or whether special-cause variation occurs. Establishing stability provides management with a basis for investigating potential sources of variation and implementing corrective actions to improve the overall delivery process.

Control charts were constructed to define the acceptable limits of variation in delivery time. Then, the delivery-time process for each product type is monitored further to detect where it may be out-of-control. Once process stability was evaluated, a capability analysis was conducted to assess how well the delivery process performance meets customer specifications. Finally, samples that violate specific SPC rules were identified to highlight where intervention may be required.

2.1. Data Preparation

A new dataset, sales2026and2027, is taken under consideration in this section. It contains the same information as the sales2023and2024 dataset; however, its values are expected values that the company has forecasted.

Before the SPC was carried out, the data source products_data.csv was merged with the data source sales2026and2027.csv. This allowed a ‘product category’ column to be added to the sales2026and2027 dataset according to the matching product IDs in both datasets, which was used to group the data by product type.

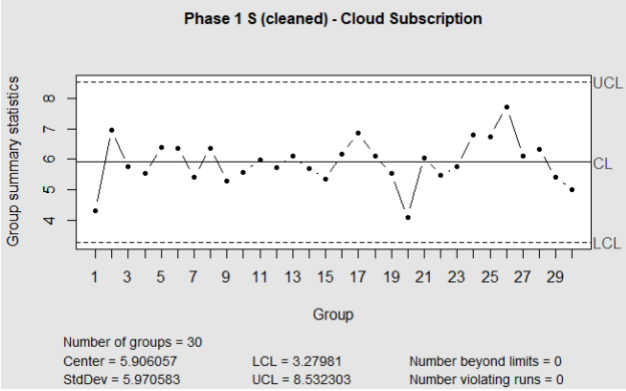
The data was also sorted chronologically by year, month day, and order time, which allowed samples to be taken for each product type at different points in time.

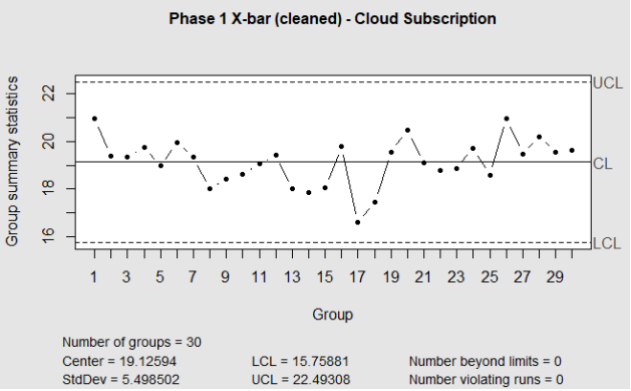
Delivery times were then grouped into subsamples of 24 deliveries per product categories, with only full samples (samples containing 24 deliveries) considered.

2.2. Phase 1 – Initialisation

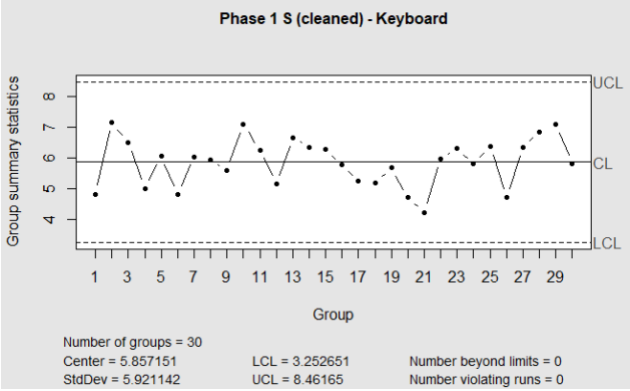
This phase develops control charts that illustrate what is considered to be ‘normal’ variation in the delivery process for each product type. Baseline control limits (centerline, ±1σ, ±2σ, and ±3σ) were established for each product type using the first 30 subgroups of 24 deliveries for each product. These values are given below each of the control charts. No categories had any subgroups exceeding +- 3 sigma, thus recalculation of control limits was not needed. The phase 1 control charts were displayed as follows:

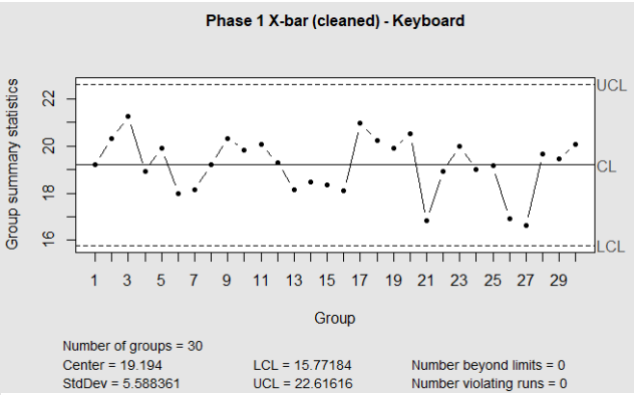
Cloud subscription:



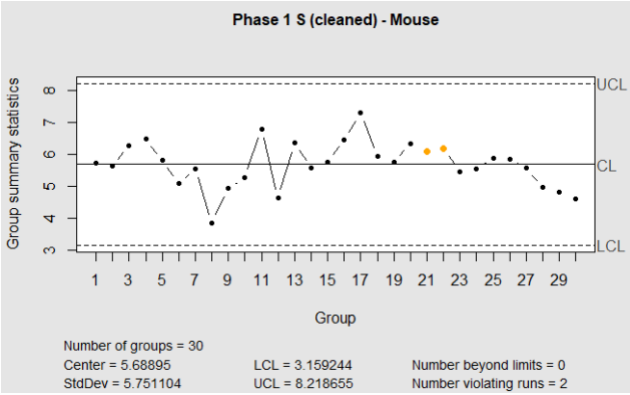


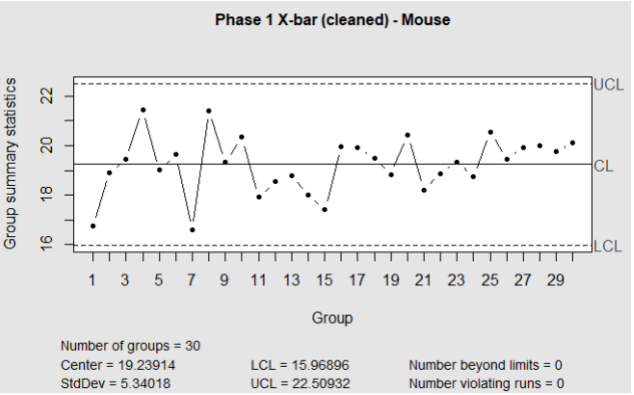
Keyboard:



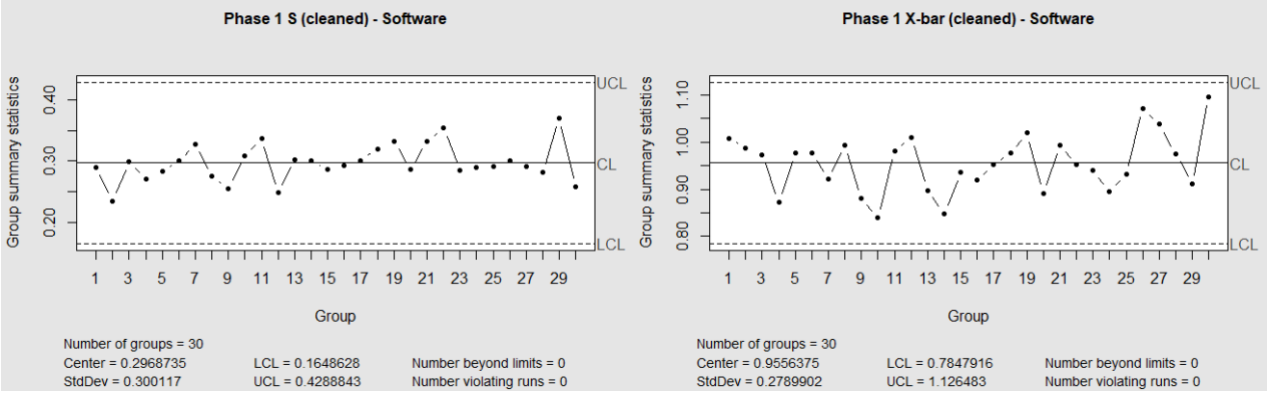


Mouse:

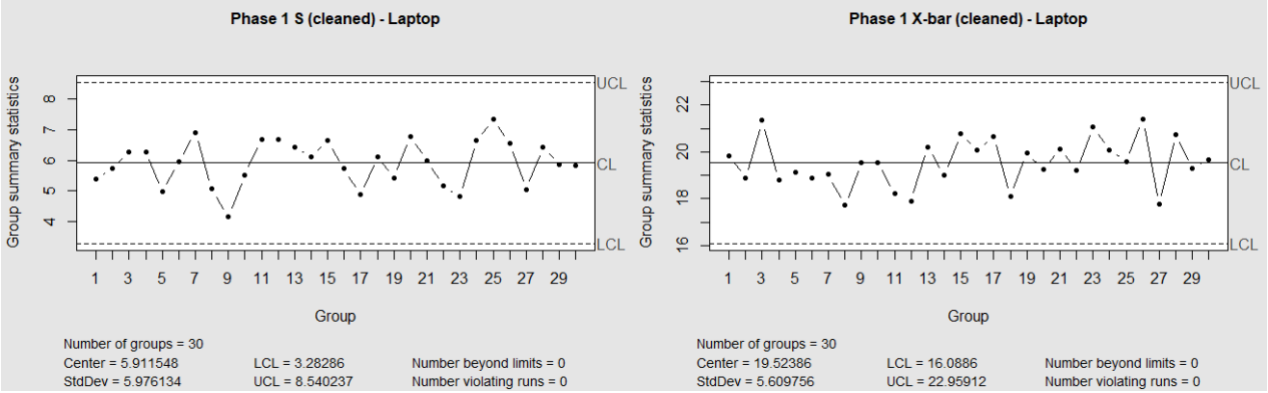




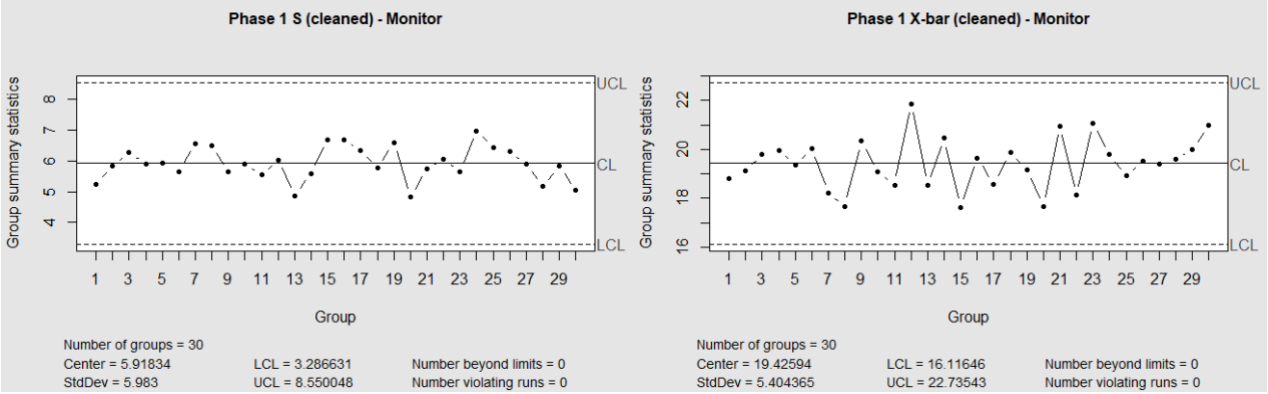
Software:



Laptop:



Monitor:



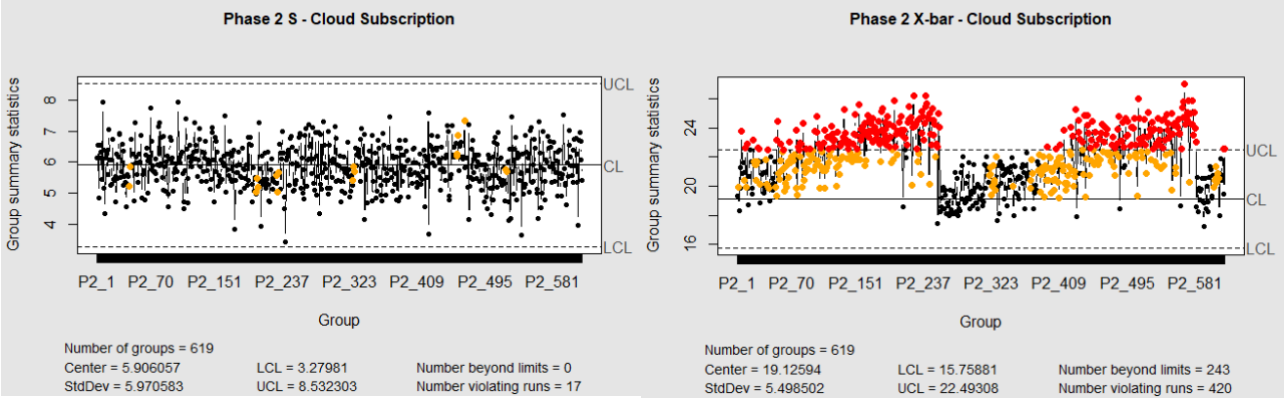
2.3. Phase 2 – Monitoring

Phase 2 was carried out to simulate real-time monitoring of delivery times using the remaining samples. This phase assesses whether the processes remain within the acceptable limits, which are the same as those calculated in phase 1 (their values are displayed below each of the charts). If they do, it can be said that the processes are stable and behave predictably. However, if the processes fall outside the control limits, variation

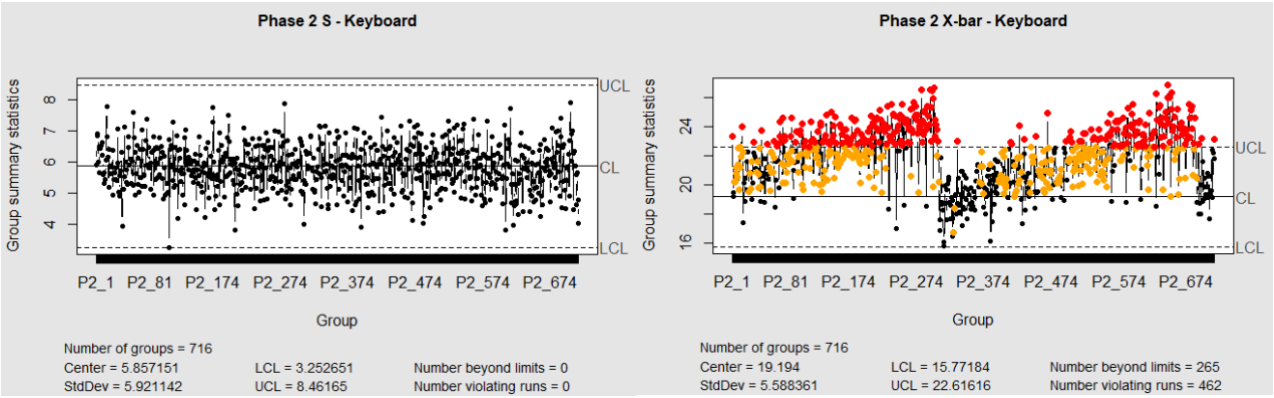
is present, and it should be further investigated to uncover any unusual occurrences in the process. These occurrences could affect quality, efficiency, and/or reliability.

Although these samples are labelled as starting from 1, the first sample is the first sample taken after the samples taken for initialization (in other words, sample 1 in phase 2 is sample 31 overall, sample 2 in phase 2 is sample 32 overall, and so on and so forth).

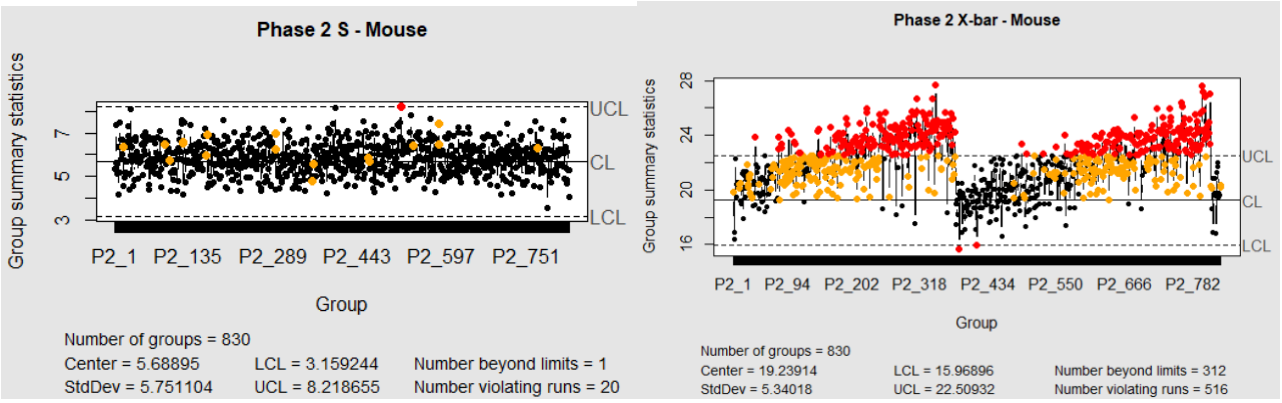
Cloud subscription:



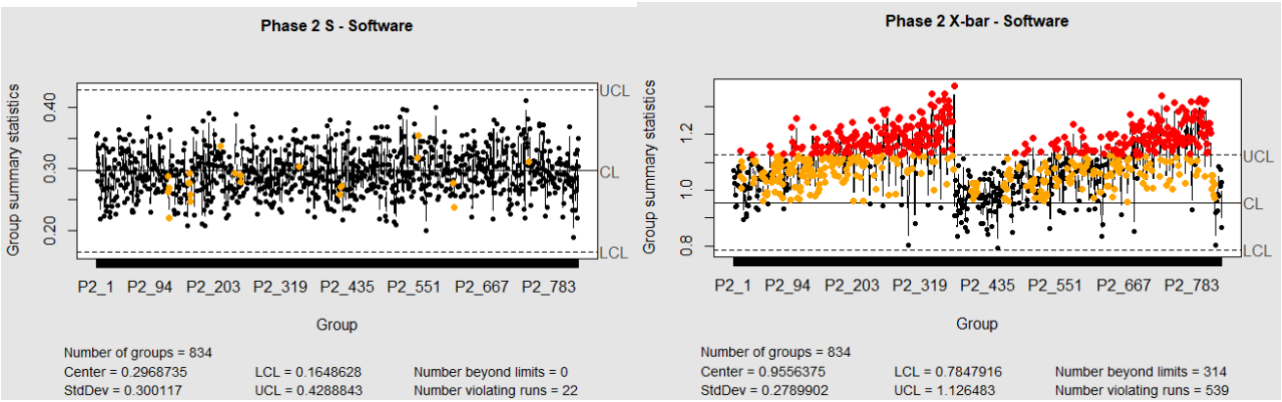
Keyboard:



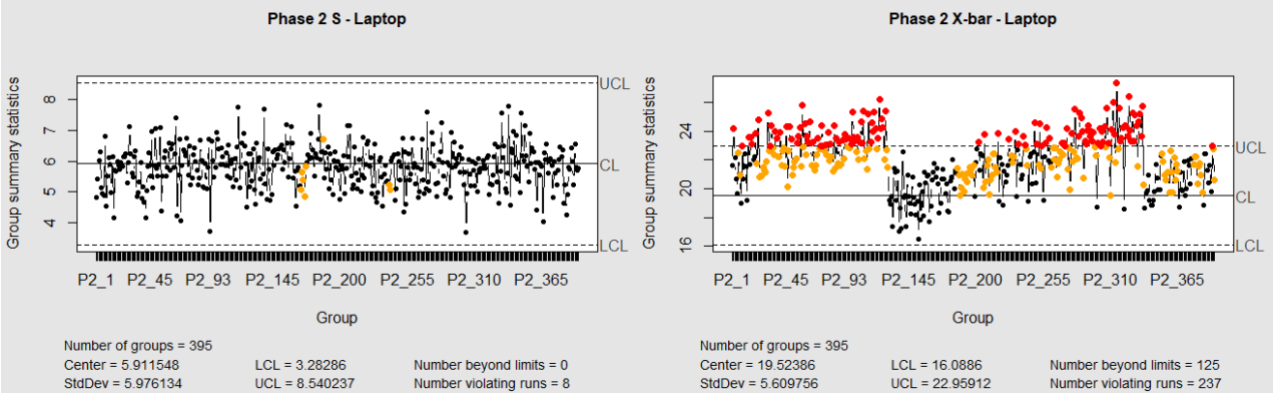
Mouse:



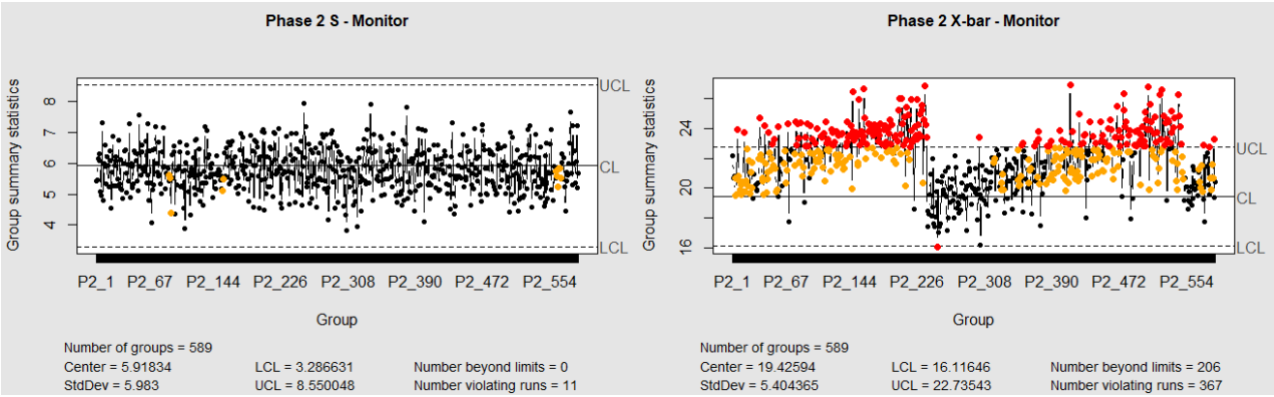
Software:



Laptop:



Monitor:



The analysis of these charts followed the standard SPC sequence: since it shows variation within the data, the s chart was evaluated first, with the x-bar chart evaluated afterwards to assess shifts in the process mean. This should prevent conclusions being draw on the x-bar chart if the s chart is out of control, since this makes x-bar values unreliable.

Five out of six of the products have phase 2 s charts that are in control, illustrating a stable delivery process for these products. However, the ‘mouse’ product has one subgroup (sample 522) that is out of control. This demonstrates special cause variation that could be

attributed to network congestion, or a change in service demand. However, apart from the single subgroup, the variability of the delivery times for ‘mouse’ products is in control. It can be concluded that the out-of-control point was more than likely due to an uncontrollable external influence rather than the company’s actual operational performance.

The x-bar charts for all product types display the same trend. There is an initial period in which the process means are in control. However, as time passes, the subgroup means gradually rise and exceed the upper control limit, forming a steady upward diagonal trend. Approximately halfway through the samples, there is a drop that brings the means back within the limits, followed by a repetition of the described trend. This suggests a cyclical cause affecting the process mean. Since this pattern is seen for all product types, it is not likely that the cause for out-of-control mean delivery times is due to the type of product being delivered, but rather a time-related or operational issue. It is also evident that the process mean itself is unstable. Even though the mean appears in control for certain periods, it is shifting systematically over time. This periodic behaviour could be due to accumulating delays, scheduling inefficiencies, or changes in workload, followed by a system reset or clearing.

2.4. Process Capability

The process capability was analysed to determine if the delivery process for each product type met customer requirements. The maximum acceptable delivery time, or upper specification limit (USL) was specified as 32 hours, and the minimum acceptable delivery time, or lower specification limit (LSL) was specified as 0 hours. The calculated process capability indices are displayed in the table below:

Product	Cp	Cpu	Cpl	Cpk
Cloud Subscription	0.898	0.717	1.079	0.717
Keyboard	0.917	0.729	1.105	0.729
Mouse	0.915	0.727	1.104	0.727
Software	18.166	35.247	1.084	1.084
Laptop	0.895	0.693	1.097	0.693

Monitor	0.889	0.700	1.079	0.700

The calculated indices suggest that the delivery process is only capable of delivering products of 'software' type within the specifications, since its Cp value is greater that 1.33 and its Cpk value is greater than 1. This is likely due to automation of the delivery process, since software is not a physical product. Since the Cpk is less than 1.33, it is clear that there is a chance of producing deliveries out of specification due to shifts in the process mean. However, the rest of the product categories do not have a delivery process capable of meeting the specifications, as all of their Cp values are lower than 1.33 and their Cpk values are less than 1. These products do not have the potential capacity to meet specification limits, nor are the delivery times consistently within the limits. This is expected after the analysis of the control charts demonstrating special cause variation in the delivery process means.

2.5. Control Issues Analysis

Potential control issues were analysed for each of the product types. This examines the stability and control of the various product categories. These issues were defined by the following rules:

- Rule A: A single s sample is above the upper 3 sigma limit.
- Rule B: The largest number of samples within the negative and positive 1 sigma limits.
- Rule C: 4 consecutive X-bar samples are above the positive 2 sigma control limits.

Since each of the product types have a different number of samples, a percentage of the samples satisfying the rules has been calculated for fair comparison. The results are displayed in the table below:

Product	Rule A (Sample/s)	Rule B (total)	Rule B (%)	Rule C (total)	Rule C (%)	Rule C (first 3 samples)	Rule C (last 3 samples)
Cloud subscription	None	19	3.069	283	45.719	23, 24, 25	582, 583, 584
Keyboard	None	22	3.073	304	42.458	73, 74, 75	689, 690, 691

Mouse	522	14	1.687	348	41.928	95, 96, 97	813, 814, 815
Software	None	21	2.518	340	40.767	103, 104, 105	819, 820, 821
Laptop	None	19	4.810	152	38.481	20, 21, 22	336, 337, 338
Monitor	None	34	5.772	254	43.124	72, 73, 74	549, 550, 551

As previously discussed, the product category ‘mouse’ is the only category that violates Rule A. Since only a single sample is identified by the rule, it should be investigated to see if it is a potential outlier.

Rule B highlights stable processes with little variation as they indicate long stretches where the samples remain within one sigma of the centerline. The longest consecutive samples falling in this region ranged from 14 (Mouse) to 34 (Monitor). Monitor products have the highest percentage of samples in the +-1 sigma range, while mouse products have the lowest percentage. However, all of the products have a low percentage of samples falling within, demonstrating the variability in the process as previously discussed. While this could demonstrate that the delivery process is responsive to changes in demand, investigations should be done to determine if a more stable and efficient process might be more profitable for the business.

Rule C indicates a shift in the process trend. This shift is relatively consistent across product types, as seen by the similar percentage of samples that satisfy the rule across product categories. This agrees with the previous findings of a consistent upward trend in process mean values for all product types.

2.6. Discussion

Analysing the delivery process across product categories highlighted the fact that the company may be experiencing operational inefficiencies. These create an increase in the mean delivery time, and should be investigated to bring the process back into control. In doing so, it should also address the current incapabilities of the delivery process for most product types to meet customer specifications.

3. Process Risk and Data Correction

The focus of this section is on evaluating the reliability and accuracy of the company's product delivery process.

The type I error, also known as the manufacturer's error, and the type II error, also known as the consumer's error, are calculated. These errors represent false alarms and undetected shifts in the process, respectively.

The analysis begins with calculating theoretical probabilities for SPC rules A, B, and C, which were described in the SPC section. Then, a practical example of a bottle-filling process is examined and the probability of failing to calculate a shift in the process is calculated.

The last part of this section does not relate to process monitoring error calculations. Rather, the discrepancies between the products_data file and the products_Headoffice file from section one were rectified based on additional information provided by head office.

3.1. Type I Error

A type I error is the probability that a process chart indicates the process is out of control even though it was stable. These probabilities are theoretical values and thus hold for all the delivery processes.

The null hypothesis (H_o) is the assumption that the process that the process is in control and centered on the centerline calculated during phase 1 of the statistical process control. The alternative hypothesis (H_a) is the assumption that the process is out of control. The type I error is the probability that H_a is determined to be true, even though H_0 is actually true. The results were as follows:

A: The probability that it is concluded that 1 sample is above the 3 sigma line when it is actually in control.

The probability of a type I error is 0.00135. This very low probability indicates that a single extreme point is rare if the process is actually in control.

B: The probability that seven consecutive subgroups are concluded to be within +- 1 sigma even when they are not.

The probability of a type I error is 0.0691, with is relatively low since seven consecutive subgroups are concluded to be within +- 1 sigma only occurs naturally roughly 7% of the time. If this occurs, may indicate unusually consistent performance, which is unlikely by

C: The probability that four consecutive subgroups are concluded to be above 2 sigma when this is not actually the case.

The probability of a type I error is 0.000000268, which is extremely small. This suggests that if a pattern of four subgroups are above the upper 2 sigma limit, this is more than likely not a false alarm. It thus should be investigated, since this rule indictes a shift in the process mean.

3.2. Type II Error

A type II error is the probability that a process is not investigated even though it is out of control. This occurs when instability is not detected in the process. In statistical terms, H_0 is concluded to be true even though H_a is actually true.

The type II error was calculated for a bottle filling process that has an X-bar chart with a centerline of 25.05 liters, a lower control limit of 25.011 liters and an upper control limit of 25.089 liters. The sigma value, or standard deviation, for this process is currently 0.013 liters. The process has shifted to an average of 25.028 liters, thus the x-bar chart has a new sigma value of 0.017. If this shift is not identified, the process would be concluded to be in control, and a type II error would have been made. The probability of making this error is 0.8412.

The type II error indicates there is a very high chance that this small shift in the process average goes unnoticed. This is due to the wider variability, which increase difficulty in determining small deviations. Additionally, the power of this test is only 0.1588. This low power indicates that the charts are not sensitive to small shifts.

3.3. Data Correction

In the data analysis conducted in section one, there were errors in the products_Headoffice file. The file contained repeated errors in ProductID, SellingPrice, and Markup for rows 11–60 of each type. These were corrected by repeating the first 10 correct values per type, and the productID column was updated to match the category.

It is important to note that transaction level data, including variables such as quantities sold, dates, delivery times, and customer information, remain unchanged. Therefore, a re-analysis was only conducted on information related to product data.

The reanalysis showed that the summary statistics for the products_data dataset remained almost unchanged, with only markup value’s standard deviation increasing slightly to

6.07% and the skewness value decreasing slightly to -0.0037. However, these changes can be said to be insignificant.

A greater change occurred in the products_Headoffice dataset once the discrepancies were corrected. The mean selling price increased to $4 493.59, which is exactly the same as the products_data mean, thus indicating consistency, while the standard deviation increased to $6 458.32. The median selling price decreased slightly to $794.19, which is the same as the mean value for products_data, while the skewness value decreased to 1.46. This illustrates that the data remains skewed to the right, with a few high-end products amongst a greater number of inexpensive ones. The range of this dataset also matches the range of products_data, with a minimum selling price of $350.45 and a maximum selling price of $19 725.18. The Head Office and subset price data now align almost perfectly. This consistency confirms that the correction successfully synchronised product pricing across both datasets. The slightly lower standard deviation in the Head Office data reflects the greater variety of mid-range products that balance the extreme values. Similarly, the mean markup percentage increased to 20.46%, matching the mean markup of the products_data dataset. A slightly lower standard deviation of 6.07% suggests that this dataset has more uniform markups overall compared to the subset of the product catalogue, while its mean of 20.34% has decreased from the original dataset, making it identical to the products_data dataset. The markup range of 10.13% to 29.84% matches the subset dataset's range. Since markup values are consistent across the subset and Head Office data, uniform pricing strategies can be identified.

The close agreement between the two datasets indicates that the correction process successfully synchronized product details between local and central records. Minor differences in standard deviation likely reflect sample-size effects rather than true discrepancies. This confirms that subsequent analyses based on the corrected local dataset can be considered consistent with corporate pricing information.

The following tables illustrate the changes in the average variable values for each product type once the discrepancies were addressed:

Category	Old Average Price ($)	New Average Price ($)	Change (new – old) ($)
Cloud Subscription	3 691.86	1 019.06	-2 672.80

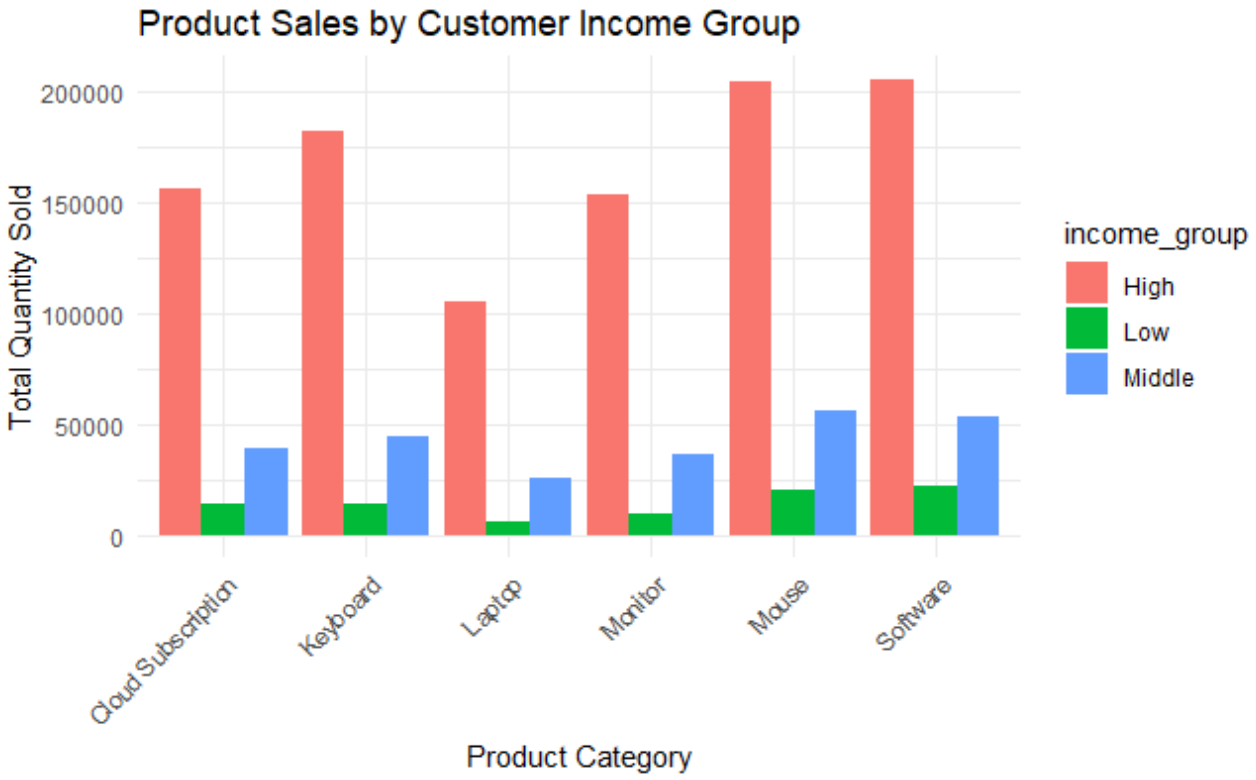
Keyboard	4 638.17	644.66	-3 993.51
Laptop	5 217.55	18 086.43	12 868.88
Monitor	5 014.17	6 310.53	1 296.36
Mouse	4 585.47	394.70	-4 190.77
Software	3 814.34	506.18	-3 308.16

Category	Old Average Markup (%)	New Average Markup (%)	Change (new – old) (%)
Cloud Subscription	20.553	19.956	-0.597
Keyboard	20.161	23.981	3.820
Laptop	20.623	18.403	-2.193
Monitor	20.727	23.868	3.141
Mouse	20.668	20.495	-0.173
Software	20.038	16.040	-3.998

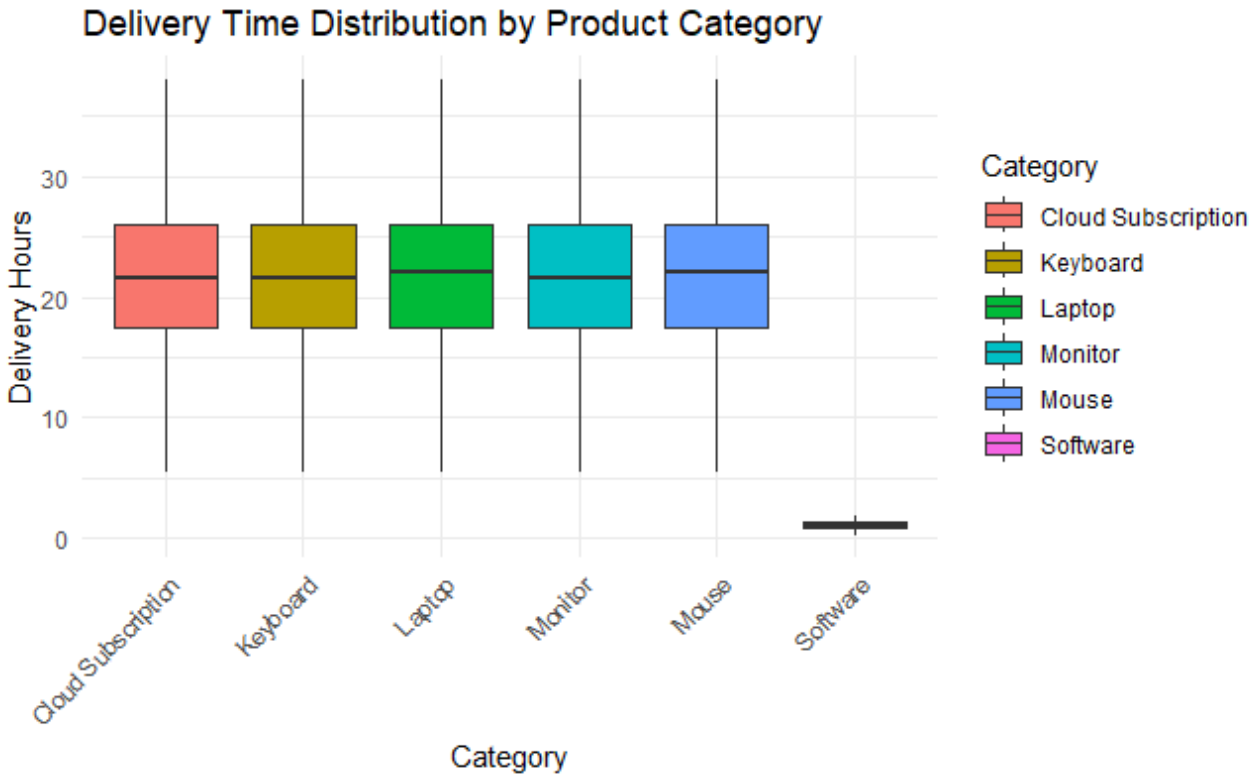
These changes highlight severe misrepresentations of both selling price and markups for each product type in the old data. This was due to mislabelling of product categories and product IDs, which led to incorrect values being recorded for each product type. High-value items were underreported while low-value items were overreported, seen by dramatic increases and decreases. The new values appear much more realistic for each of the product categories. The change in these values would also significantly impact the relationships between variables, and the corrections improves accuracy for revenue and profit calculations in later sections. Thus, data visulisations should reassessed.

3.4. Corrected Data Visualisations

As mentioned previously, only information pertaining to the products_data and products_Headoffice was updated. Thus, only visuals relaying information based on the values from these datasets were updated.

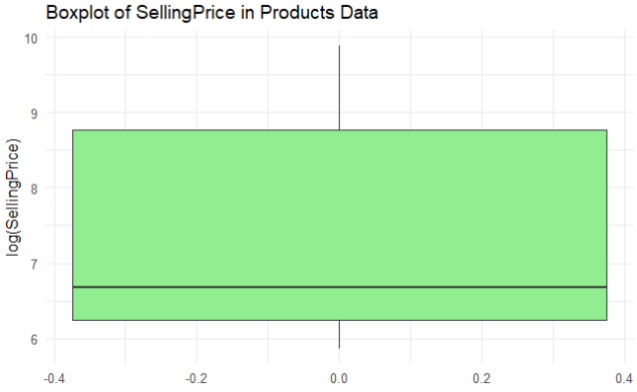


It is still clear that the majority of the company's customers come from the high-income group, while their smallest customer group is the low-income category. This maintains the fact that the company sells high-value products. However, this chart differs from the previous on as quantities sold to each income group are not consistent across categories. This demonstrates the company has products differing in popularity across income groups, highlighting products they should focus on. High-income customers purchase software the most, with mice being a close second, and laptops the least. Middle-income customers purchase mice the most and monitors the least. Low-income customers purchase mice the most and laptops the least. Since mice are popular across the income groups, it might be a sensible idea for the company to specialise in this product category if they do not already.



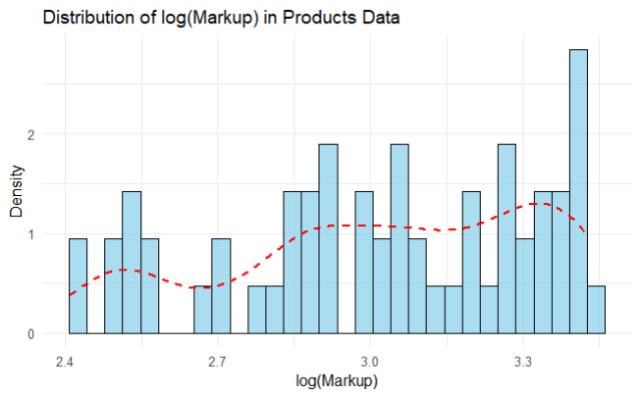
There is a very evident change in the distribution of delivery times by product category. The median remains approximately the same across product types (excluding software), with a value of roughly 22 hours for each product time. This is similar to the old median of approximately 20 hours. The variation of delivery times across the product types appears highly consistent, suggesting that the same process is applied for each product is the same, which contradicts the previous assumption. Furthermore, the software product type has gone from having the highest variation in delivery time to the lowest variation. It also has much shorter delivery times than the other product types, with a smaller range overall. This is expected, and is in line with the previous idea that the delivery process for software is likely automated and faster as this product does not have to be physically distributed.

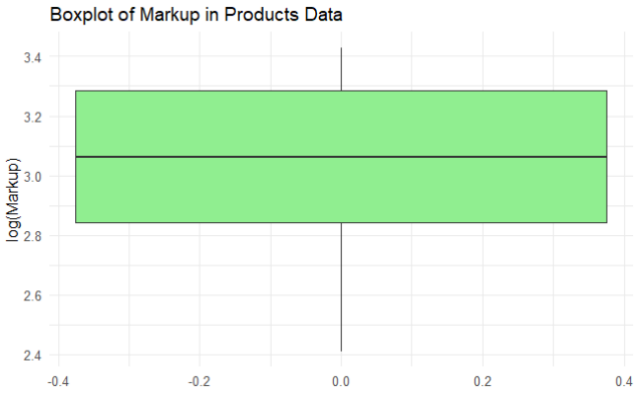
Selling price:



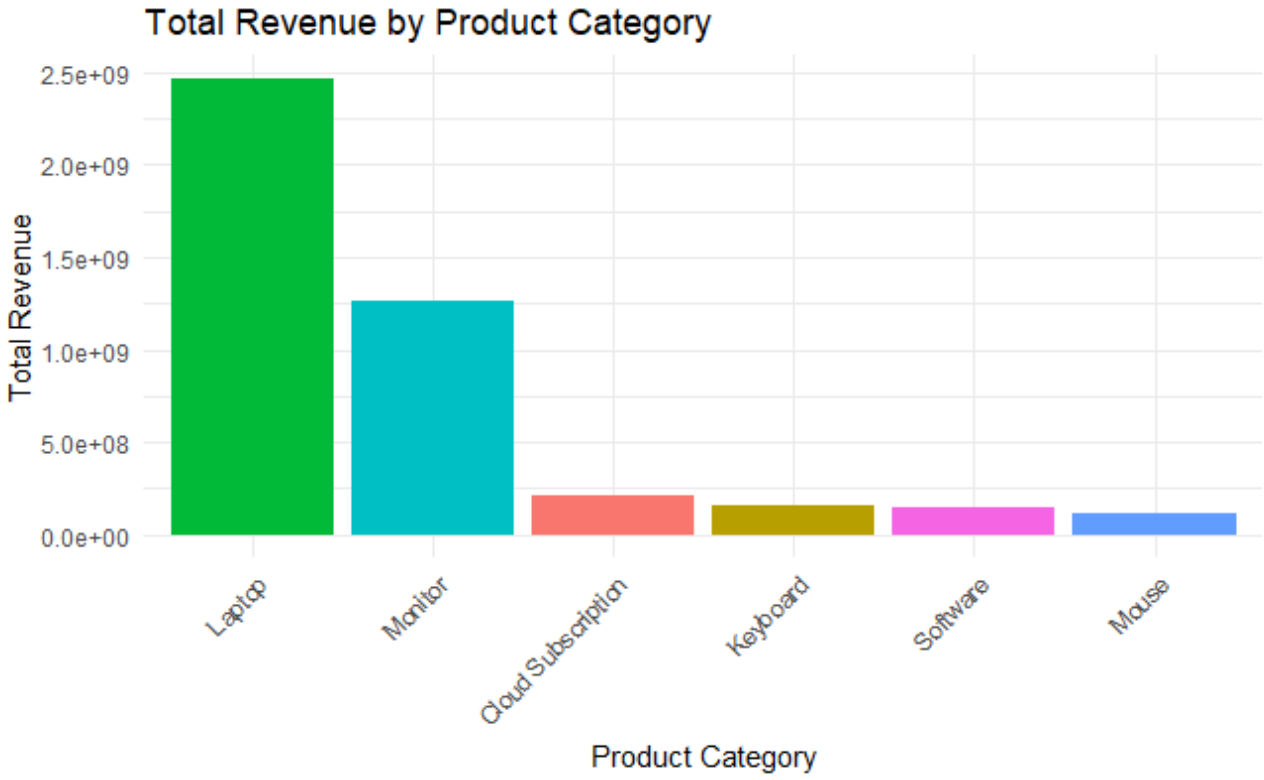
While the overall skew of the selling price didn't change, there appears to be a change in frequency of each selling price.

Markup:





Similar to the updated selling price values, the overall skew of markup values remained the same as it was with the previous data, however the frequency of each markup changed.



Unlike before, this visual description of revenue generated per product category varies greatly for different product types.

3.5. Discussion

The type I errors demonstrate that it is unlikely for a s chart to portray a process as out-of-control when it is in fact within the control limits, that a pattern of seven consecutive subgroups found to be within one sigma of the centre line is likely a true occurrence, and that four points above sigma are not likely to be a false alarm. Furthermore, the type II error demonstrated that the charts were not sensitive to slight shifts in the process mean. Additionally, the correction of the discrepancies in the products datasets lead to relationships between variables changing, which altered some of the conclusions drawn in the original analysis of the data.

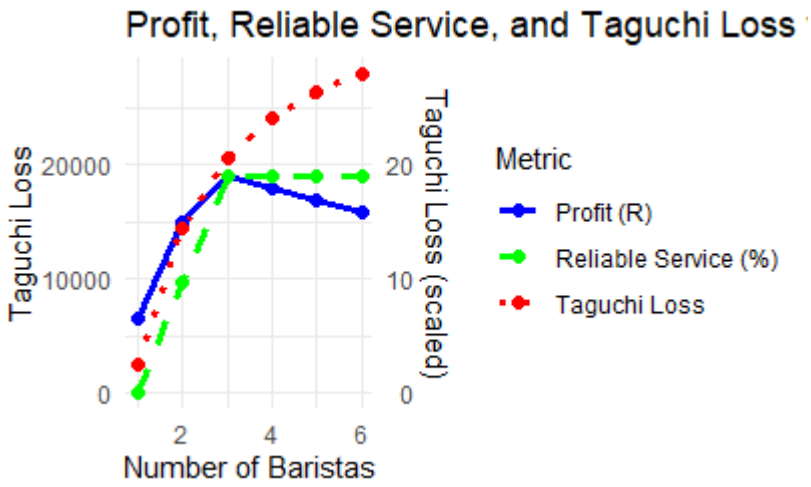
4. Operational Profit Optimisation

The following section uses a practical example to demonstrate how increasing the workforce can increase the profit of a business. Typically, the speed of service along with the number of customers serves increases with an increase in workforce. However, simply selecting the maximum number of employees is not necessarily the best solution for optimising company profit, since the costs of employees increases too.

The example analyses two coffee shops that are part of the same chain. The two shops differ in the times it takes for their customers to be served. It is assumed that both coffee shops are situated in high-traffic areas, such as an airport. According to the Dojo Business Team (2025), a high-volume coffee shop has an average of 750 customers per day. This value is used as the customer cap. In an article by Prateek Vasisht (2023), it is noted that the average time Starbucks customers spend waiting for their order to be served to them is 4.5 minutes, or 270 seconds. This is the value used as the target time. The reliable service threshold is set as an interval of 240 seconds and 300 seconds. Service times outside of this interval are considered to be unreliable and thus negatively affect customer satisfaction.

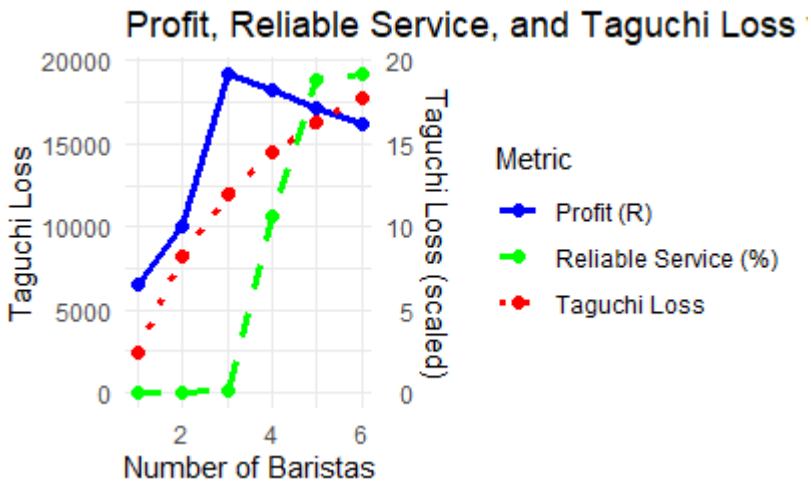
It is also important to define Taguchi's Loss before discussing the results. This is the loss a company experiences when a process deviates from its target (OpEx Learning Team, 2017). In this case, it represents customer dissatisfaction if the service time strays from 270 seconds, even if it falls within the limits of the reliability threshold.

4.1. Coffee Shop 1



The first coffee shop analysed has an optimal number of 3 baristas. This results in the shop obtaining a profit of $19 086.33 and a reliable service percentage of 100%. However, employing this many baristas results in a Taguchi loss of 413.67.

4.2. Coffee Shop 2



The second coffee shop analysed also has an optimal number of 3 baristas. This results in the shop obtaining a profit of $19 201.11 but a reliable service percentage of 0.7%. Further investigation should be done to determine whether the service is too fast or too slow, and improvements can be made from there. The Taguchi loss for this dataset is 238.89, which suggests that the service time deviates less from the target value than the service time for coffee shop 1 does.

4.3. Discussion

By comparing the two coffee shops, it is evident that the optimal solution to increase profit for one does not necessarily yield the same results for the other. Thus, when a business is

trying to optimise its profit, should analyse specific processes or individual components separately.

It is also important to note that the Taguchi loss appears to increase with an increase in the number of baristas employed. This is likely due to the fact that increasing the number of baristas decreases the service time, causing it to deviate largely from the target service time. It is important to balance optimising the profit with minimising the Taguchi loss, as a large loss implies a great amount of customer dissatisfaction.

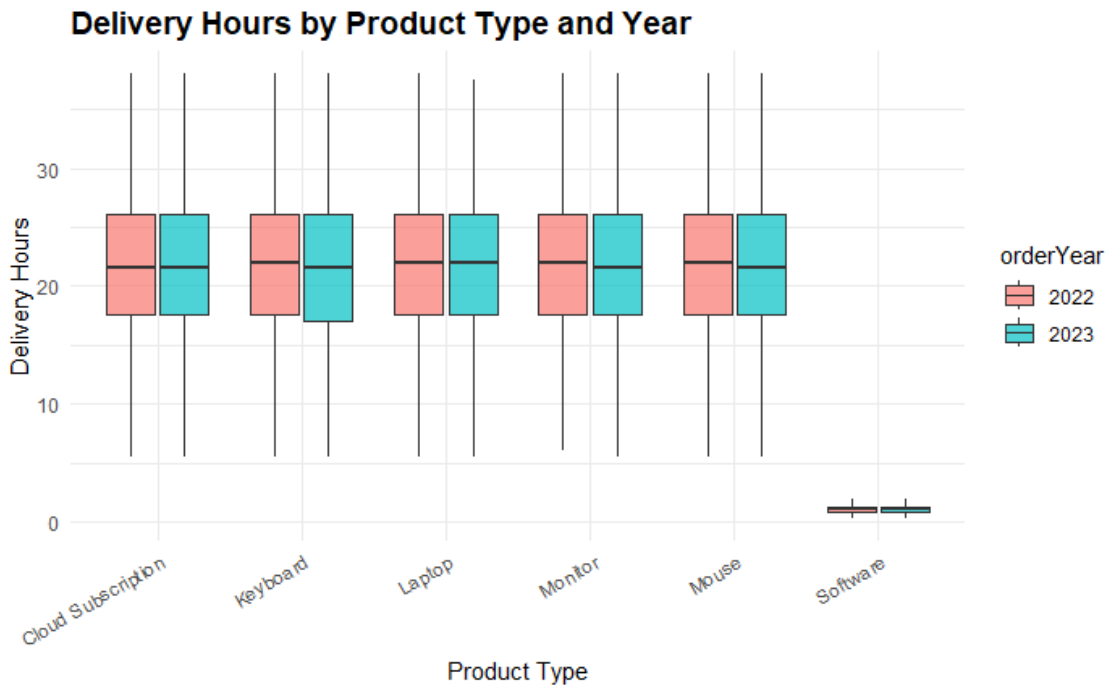
5. Design of Experiments and Variance Analysis

Design of experiments is important to recognise the relationship between inputs that affect both a process itself and the outputs of the said process. Variance analysis determines whether differences between group means are statistically significant. Together, these can be used to conduct experiments on the data and interpret the results of the experiment, allowing for data-driven process improvement.

This section of the report tests three hypotheses with different methods. A two-way ANOVA is used to determine whether mean delivery times differed across product types between the two years given in the data (2022 and 2023). The assumption that variability f delivery times is constant is tested using Levene's test. Lastly, a MANOVA is used to determine if delivery time is associated with order volume.

5.1. ANOVA

The hypothesis test by this ANOVA was that the mean delivery times were the same for all product types. The results indicated a highly significant effect of product type (p < 2e-16, which is less than 0.05), suggesting that average delivery times vary substantially between product categories. In contrast, the effect of order year and the interaction between product type and year were not statistically significant (p = 0.0976, which is greater than 0.05), implying that the differences in delivery performance across products remained relatively consistent between 2022 and 2023.



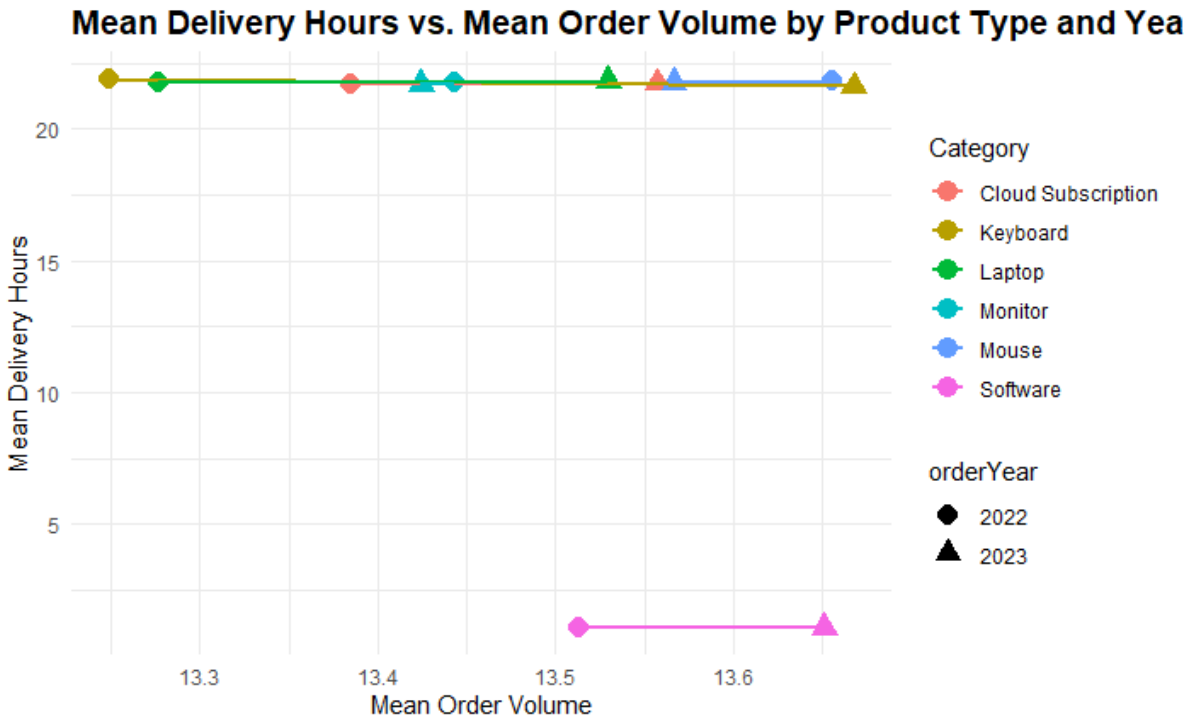
The above chart confirms the conclusions drawn. Each product type does not appear to have a different spread of values over the two years (seen by the minimum change in the blue and pink boxes), indicating the mean delivery time for a certain product type is consistent with time. However, it is apparent that the mean delivery time is not constant across product types. More specifically, it can be seen that the mean delivery time for software products is vastly different to the mean delivery time for the other products as it has a much smaller range of values for delivery time. The delivery process should be further investigated to see if any of its aspects can be incorporated into the delivery of other product types. Speeding up the delivery times may improve the satisfaction of the company's customers since the quality of the service offered to them is seen to increase. This could even lead to greater profits for the company.

5.2. Levene's Test

The hypothesis tested by Levene's test was the variability of the delivery times is the same for all product types. Levene's test returned significant results (p < 2.2e-16), indicating that the variability of delivery times is not constant across product categories or years. This violates the homogeneity of variance assumption of ANOVA but is not uncommon with operational time data; it suggests that some product types experience greater inconsistency in delivery performance than others. As previously mentioned, the delivery processes should be investigated in greater detail to isolate where the issues may lie.

5.3. MANOVA

The hypothesis tested by the MANOVA was delivery time is not significantly associated with order volume. To test this, the joint relationship between delivery hours and order volume was explored.



The scatterplot of group means shows that mean delivery hours remain largely stable across categories despite little to no fluctuations in mean order volume. This supports the conclusion that order volume is not strongly associated with delivery performance, consistent with the non-significant MANOVA findings. However, since order volume has only changed minimally, investigating this relationship over a few years when large variations in order volume are apparent should be considered. This might illustrate the relationship better than the current data does.

5.4. Discussion

Overall, the analysis confirms that delivery performance differs by product type but is stable across years, and that delivery time and order volume behave largely independently. These findings highlight the importance of tailoring reliability improvements to specific product categories rather than to year-to-year operational changes.

6. Reliability of Service

The car rental agency's staffing data over 397 days were used to assess the reliability of daily service. Days were considered reliable when at least 15 employees were on duty. Based on the observed distribution, 366 days (92.2%) had 15 or more workers, suggesting that service was reliable on most days. This also suggests that service was unreliable 7.8% of the time. By converting this to a 365-day year, we expect approximately 336 reliable days and 29 problem days per year. This indicates that while the company performs reliably for the majority of the year, occasional under-staffing still leads to periods of reduced service reliability. This reduced reliability can lead to a loss of income, which in turn reduces business success by decreasing the agency's overall profit. To avoid this, the optimal number of workers must be found in order to ensure service is achieved with maximum reliability.

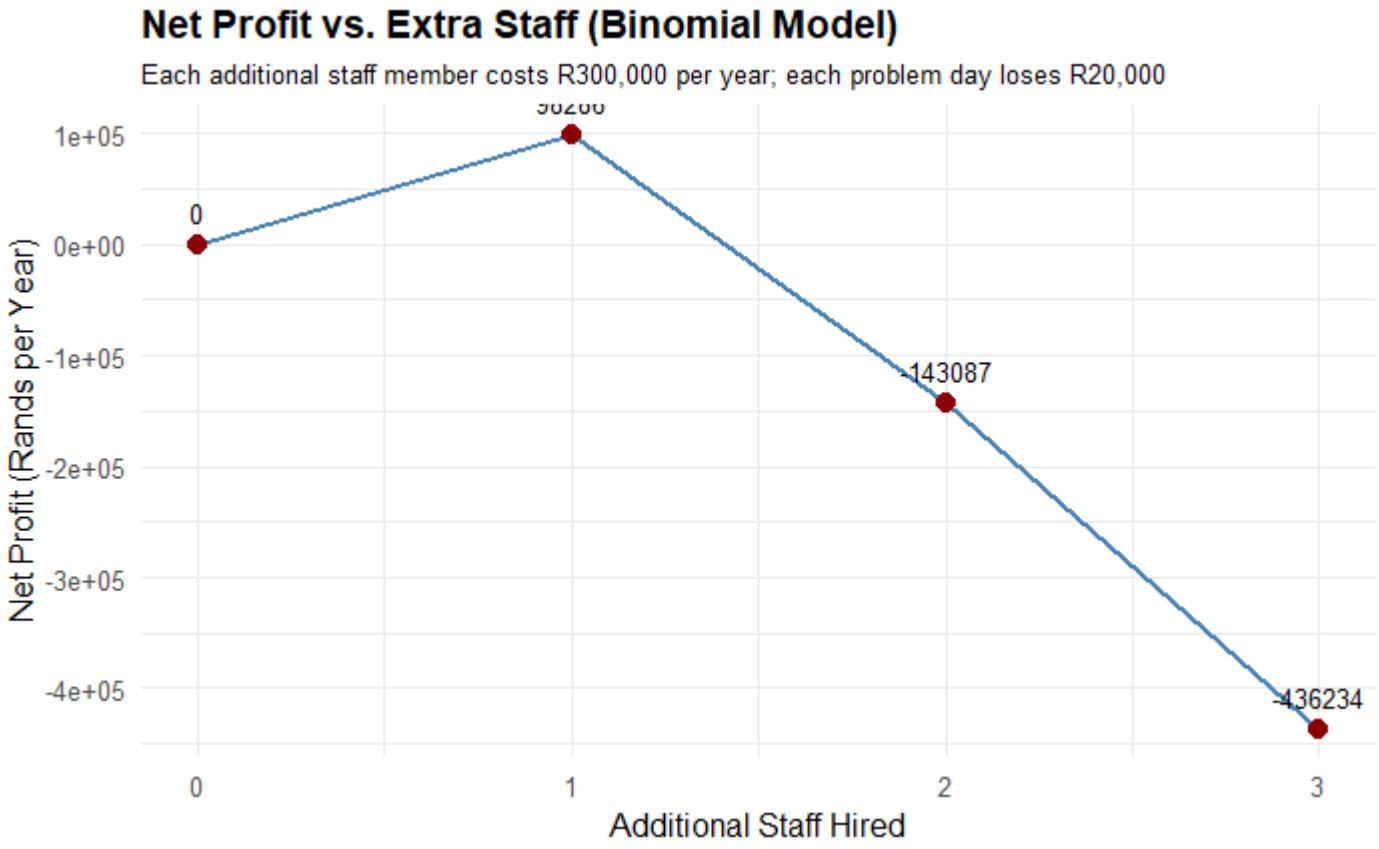
A binomial model was used to estimate the expected number of reliable service days per year. The model was appropriate because each day could be classified as either a 'success' (reliable service) or a 'failure' (unreliable service).

6.1. Binomial Model for Profit Optimisation

As previously mentioned, staffing reliability with profitability, the number of workers on duty was modelled using a binomial distribution in order to link staffing with profit.

From the data, the average number of staff per day was estimated as 15.57, giving an attendance probability of 0.973. The Binomial model calculates the probability of understaffing as 0.0637 for the current roster. The probability of a problem day and corresponding financial outcomes were calculated for adding 0–3 extra staff members, and the results are displayed in the table below:

Extra staff	Problem days per year	Days saved	Benefit ($)	Cost ($)	Net profit ($)
0	23.23	0	0	0	0
1	3.31	19.91	398 286	300 000	98 286
2	0.38	22.85	456 913	600 000	-143 623
3	0.04	23.19	463 766	900 000	-436 234



As shown in the line graph above, hiring one additional staff member provides the highest expected annual profit of R98 286, while further hires result in diminishing returns due to increased labour cost. Adding one more worker effectively reduces the chance of having fewer than 15 workers, because now we can treat 14-worker days as reliable. After hiring one extra staff member, the Binomial model calculates the probability of understaffing as 0.0091.

6.2. Discussion

It is evident that service reliability is strongly dependent on staff availability. With the current staffing levels, the company is expected to operate reliably on around 336 days per year; however, low staffing levels on approximately 29 days lead to significant losses. By employing one additional worker, the probability of under-staffing drops from 6.4% to below 1%, improving service consistency and customer experience while still being financially beneficial. Thus, it can be concluded that reliability improvements contribute to

increases in profit, although over staffing must be avoided as it can diminish profitability as employee costs increase. This is why the option that provides the most benefit to the company may not lead to an increase in the company's profit.

7. Conclusion

This project utilised a variety of statistical techniques that analysed data regarding business processes. It highlighted the fact that a systematic method can be applied to data describing a business in order to gain greater understanding regarding its operations. In doing so, areas needing improvement can be identified, and possible solutions can be applied. The solutions are not simply applied based on trial-and-error, but rather on concrete data, ensuring companies experience the best results when implementing improvements.

References

Bobbitt, Z. (2021). *What is Considered a Low Standard Deviation?* [online] Statology. Available at: https://www.statology.org/what-is-a-low-standard-deviation/ [Accessed 29 Sep. 2025].

Menon, K. (2022). *The Complete Guide To Skewness And Kurtosis | Simplilearn*. [online] Simplilearn.com. Available at: https://www.simplilearn.com/tutorials/statistics-tutorial/skewness-and-kurtosis [Accessed 29 Sep. 2025].

OpEx Learning Team (2017). *What is the Taguchi Loss Function - 6sigma*. [online] 6sigma. Available at: https://6sigma.com/what-is-the-taguchi-loss-function/ [Accessed 23 Oct. 2025].

The Dojo Business Team (2025). *How many customers does a coffee shop have per day?* [online] BusinessDojo. Available at: https://dojobusiness.com/blogs/news/customers-per-day-coffee-shop [Accessed 6 Oct. 2025].

Vasisht, P. (2023). *The 9-minute takeaway coffee*. [online] Management Matters. Available at: https://medium.com/managementmatters/the-9-minute-takeaway-coffee-67045d359b57 [Accessed 6 Oct. 2025].