Quality Assurance 344
ESCA Report

By Adam Lewis
27078272

Table of Contents

Table of Figures

1. Introduction:

This report investigates and optimises a variety of service and delivery processes for a singular business offering a large and diverse amount of product types. Such products include Laptops (LAP), Monitors (MON), Keyboards (KEY), Software (SOF) , Mice, (MOU) and Cloud Subscriptions (CLO). The main objective is to ensure enhanced customer satisfaction, while maintaining process stability and maximising profitability. This is all done using statistical and analytical methods within our engineering framework. The report makes use of SPC (Statistical Process Control) techniques such as X-bar and S-charts to help evaluate variation in delivery performance and expose instances of poor process stability. Capability instances such as Cp, Cpk, Cpu and Cpl are then calculated in order to determine if each process meets customer needs and specification limits, while the risks associated with Type 1 and Type 2 errors are assessed to determine the reliability of the process monitoring decisions. The report further combines optimisation and modelling processes that create a staffing plan for 2 coffee shops that balances efficiency with service quality and profitability. In addition to this, the report includes DOE (Design of Experiments) and ANOVA (Analysis of Variance) to investigate the statistical differences between time periods and product categories, giving us insight into the process variability and overall performance trends. Finally, we look at the reliability of service for a car rental agency through a binomial modelling approach that delivers the likelihood of achieving reliable service based on the number of employees on duty. By combining process control, optimisation and reliability analysis, the report shows a comprehensive capability to investigate complex engineering problems using data-driven reasoning and statistical evaluations.

2. Descriptive Statistics:

2.1 Customer Data:

The table below shows the average income of each city where our customers are based. This indicates which cities have higher average income levels. Data like such can be very useful and allow the business to make use of strategies like targeted pricing and marketing. Miami and Chicago have the highest average incomes, showing us, where targeted marketing and pricing strategies would be most effective.

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

Figure 1 - Average Income vs City

Next, we look at the corelation between age and income. It is denoted by a Pearson coefficient of 0.158, which is a weak but positive linear relationship. This means that as age increases, income increases slightly. This can be further seen on the scatter plot below that compares age vs income. The blue line, which is not drastically steep, indicates this weak but positive relationship. From this, we can conclude that a few other factors, apart from age also, have an influence in determining income levels.

```
Pearson Correlation Coefficient (Income vs Age): 0.158
Interpretation: Weak positive linear relationship.
```

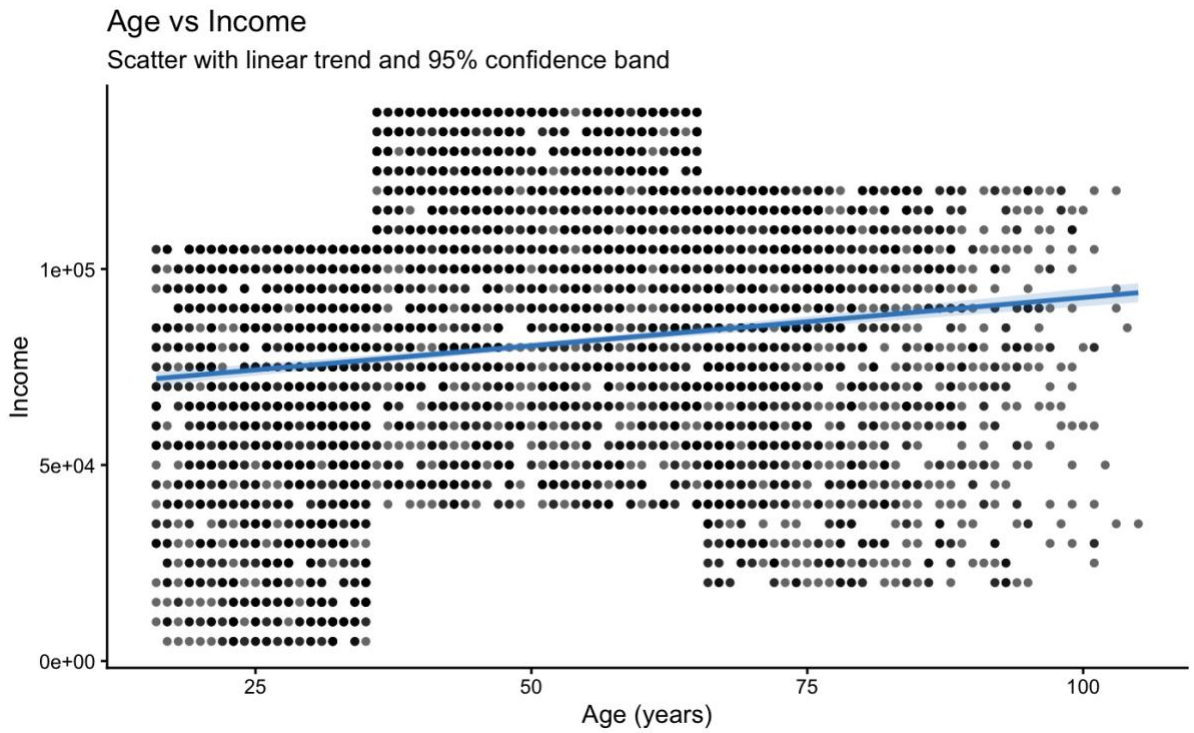


Figure 2 - Age vs Income

When determining if there is a noticeable difference between the average incomes (purchasing power) between males, females and other genders, there are a few components that need to be considered. The one-way ANOVA table below shows us a p-value of 0.998, which means that there is no statistically significant difference between the mean income levels of the genders. This tells us that any small defences seen are a result of random variation rather than a true underlying pattern in purchasing power between genders.

The graphs below further confirm this statement as true, as the genders all have the same range of values in the boxplot. Their median incomes are all quite similar and they have interquartile ranges that overlap, showing us no clear difference in central tendencies or spread between groups. The density plot can be seen showing the same outcome, as the curves for all the genders generally overlap, meaning that the probability distributions of income are nearly identical. All of this indicates no significant differences between the purchasing power of males, females and other genders.

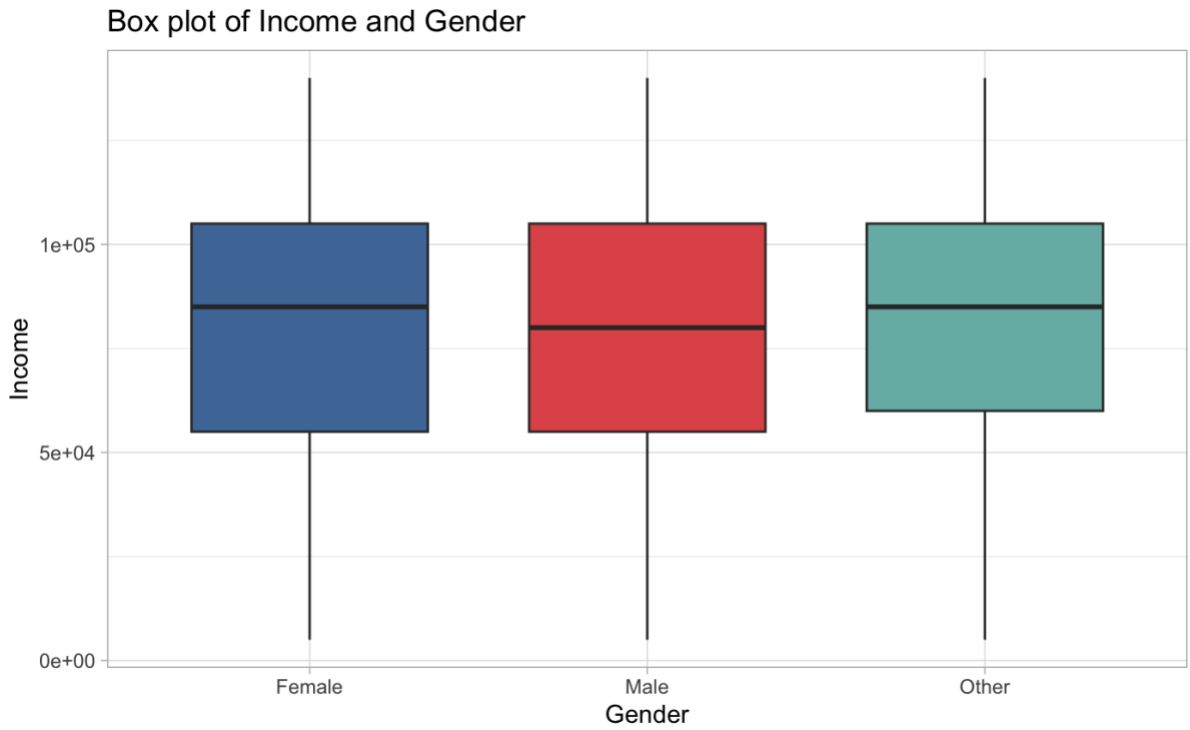


Figure 3 - Income vs Gender

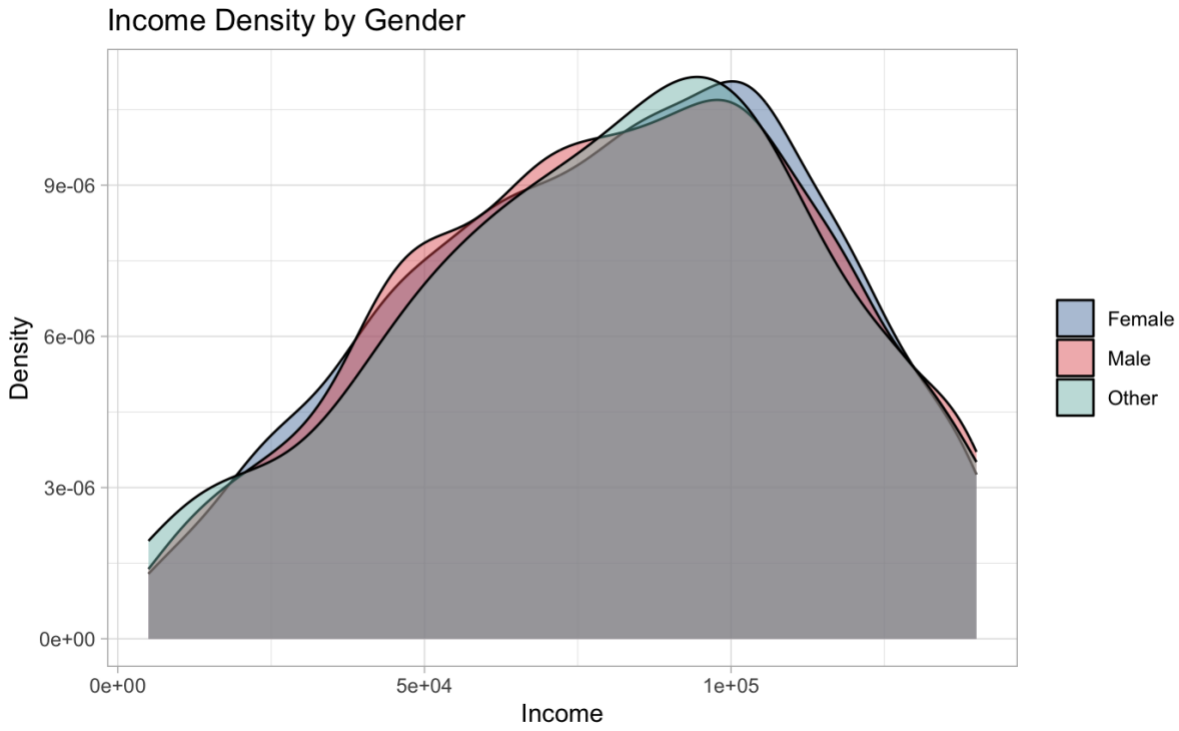


Figure 4 - Income Density by Gender

Age distribution is also a very useful set of data to understand. This is because it helps a business tailor age-specific product offerings and market strategies. As seen in the table below, the age bracket of 65+ has the highest count of customers, being 1484 customers. This means that most products must be designed for this age group, as they are the largest target market. The age distribution graphs can also be seen below.

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

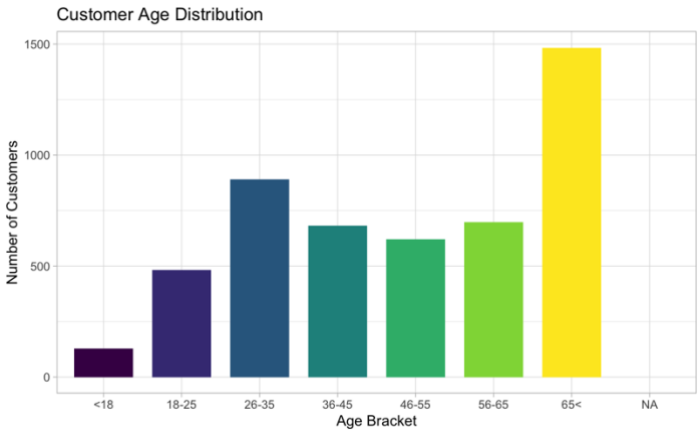


Figure 6 - Customer Age Distribution

Figure 5 - Age Distribution per City

The two visualisations above both show the age distribution of the customer data, one of all the customers and one of each city's customers more specifically. They both indicate the same results as the table above. That being that the majority of customers are of the 65+ age bracket.

2.2 Product Data:

The product data was first fixed, as it was noticed that the categories weren't consistent with the product ID's. It was then ensured that each product ID was set with the correct corresponding product category and a new csv file called products_corrected. This data was then used for the data analysis. The cost price of each individual product was also calculated and stored in this file for further analysis.

The table and boxplot below show the key metrics about pricing, markups and the cost structure of all the product types. average markup, average selling price, average cost price and average markup as a percentage of cost. From this we can see that the business's most expensive products have the lowest percentage margins, suggesting price sensitivity and competitive market pressure.

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

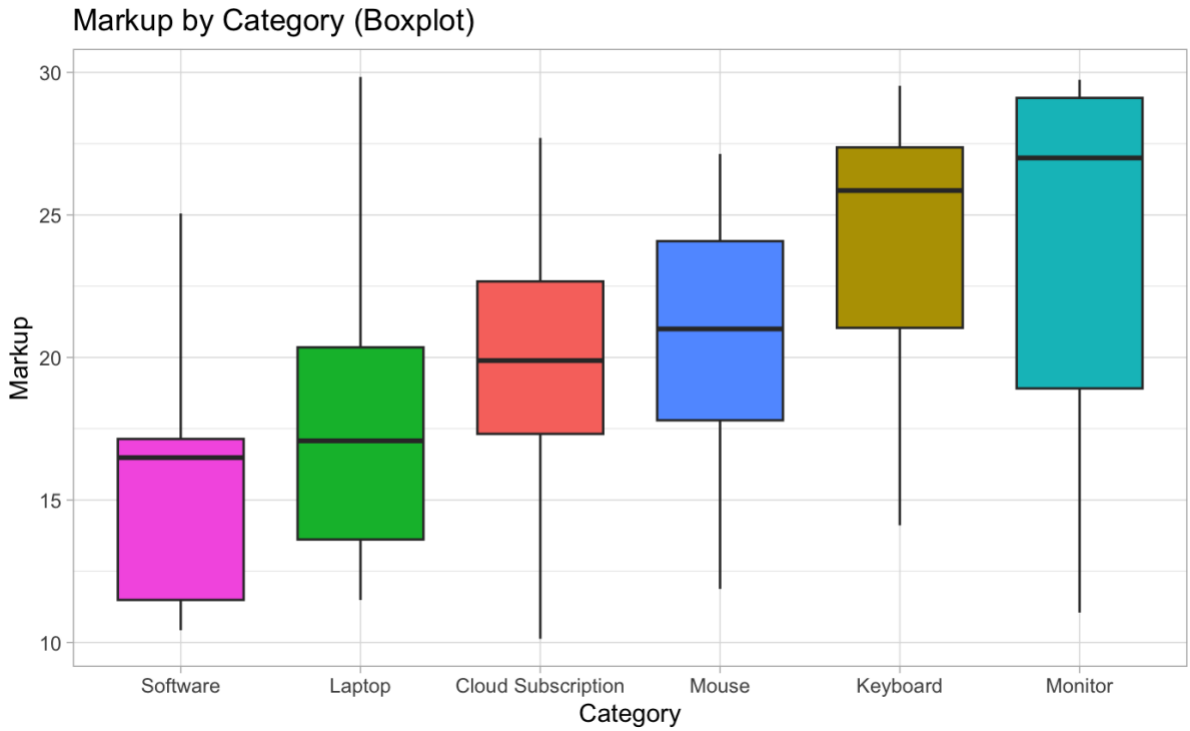


Figure 7 - Markup by Category

2.3 Sales and Merged Data:

The sales data first was altered to make sure that it was in the correct order by date. This then created a new csv file called order_corrected_sales_data. The data analysis was done using this data.

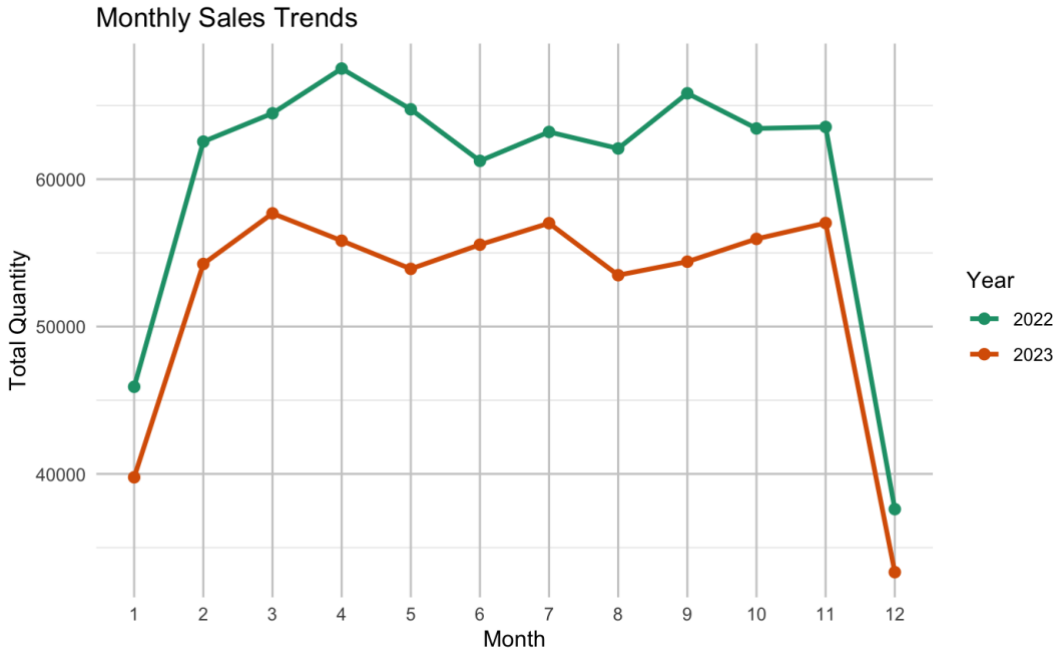


Figure 8 - Monthly Sales Trends

The first visualisation plots revenue by month for both 2022 and 2023. This graphs shows the distinct seasonal fluctuations in the revenue data for the business, with more revenue being brought in during year 2022 vs 2023. In both year's, one can see that revenue rises sharply during the middle of the year (months April to June), before tapering off at the end of the year (months of November and December)

This observed pattern suggests two possible business dynamics. The first being that demand may naturally peak in the mid-year as a result of promotional campaigns or industry cycles, with end-of-year periods associated with lower activity. The second being that the dip in late 2022 and 2023 indicates possible operational bottlenecks, such as supply chain shutdowns that impact order completion during these critical months.

This highlights the need for targeting promotional or marketing campaigns by the business in quarter 4 of the year to try offset this drop, creating a more consistent revenue level over the entire year.

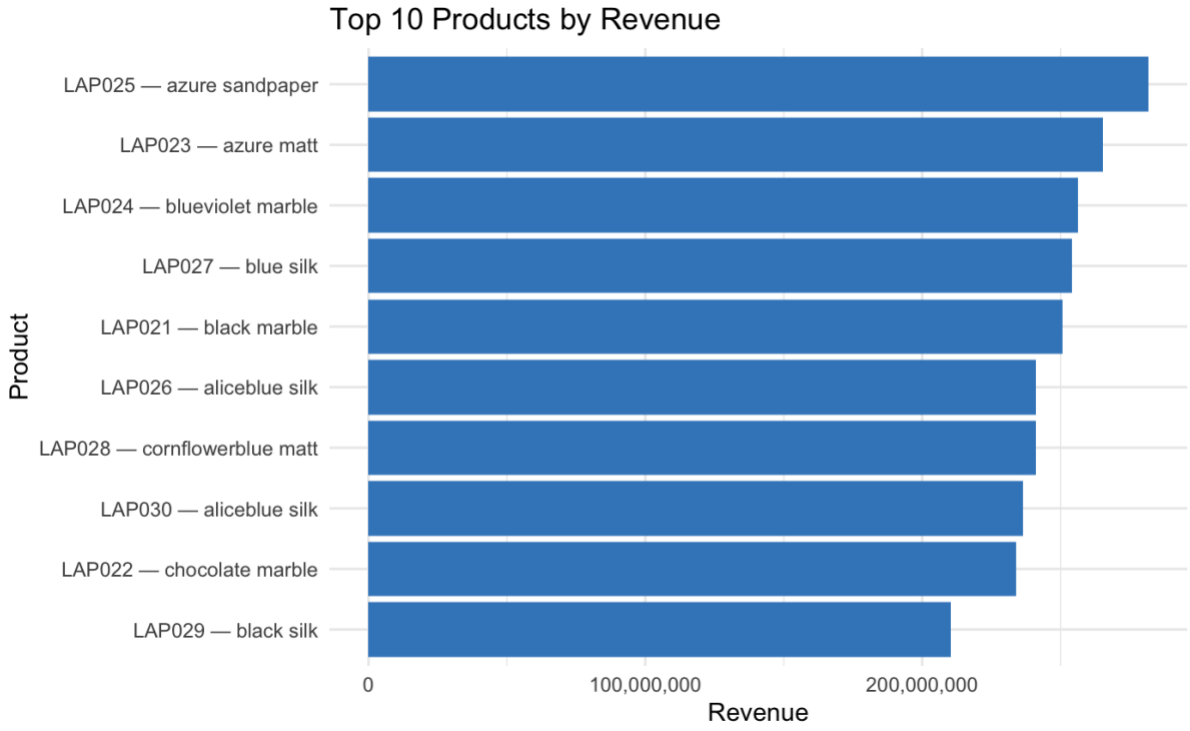


Figure 9 - Top 10 Products by Revenue

The top 10 products by revenue bar chart above shows which products create the biggest revenue stream for the business. One can see that sales are very heavily concentrated in laptop products, with the LAP025 (“Azure Sandpaper”), LAP023 (“Azure Matt”), and LAP024 (“Blueviolet Marble”) models dominating the category.

This product concentration risk of these products comes with both advantages and disadvantages. On the one hand, it shows the clear demand for premium laptop models, backing up the product portfolio. On the other hand, it shows a heavy reliance on a handful of SKU’s (Stock Keeping Units), meaning that any disruptions in supply or changes in customer preferences and spending patterns could have a significant impact on revenue.

The take away from this is that the company should consider diversifying marketing across mid-range laptops, accessories and subscription services to try reduce their dependency on a small set pf high performing SKU's.

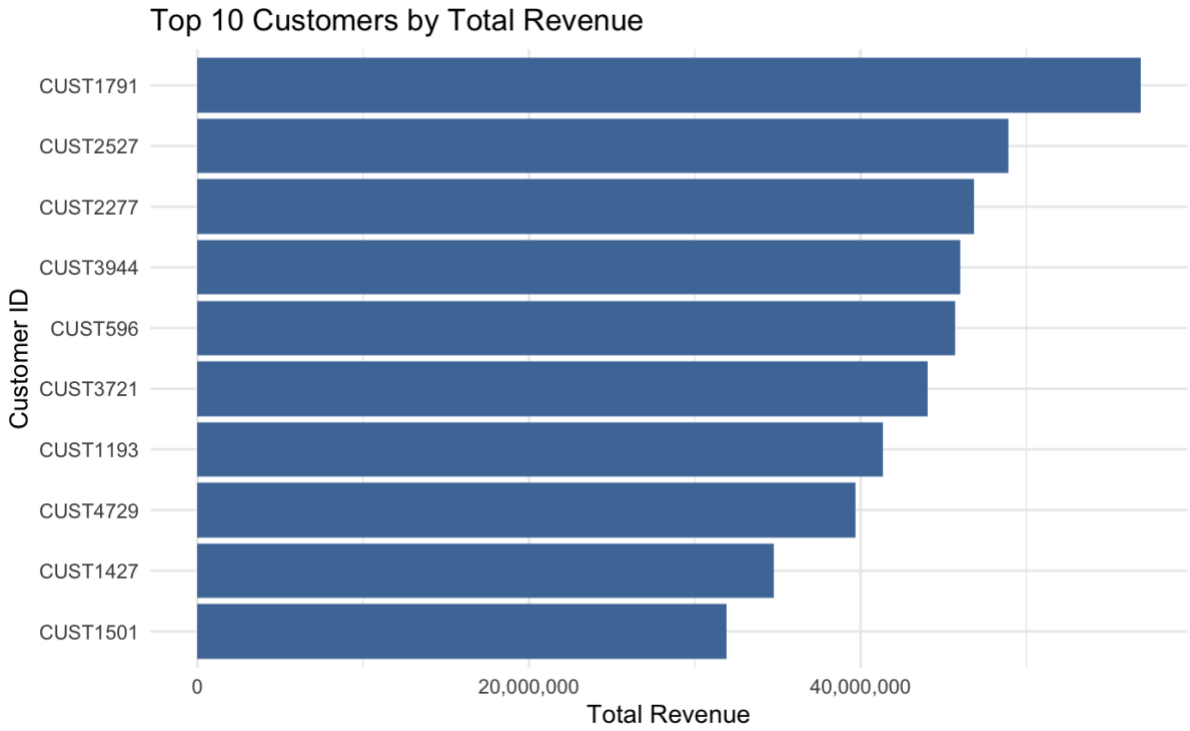


Figure 10 - Top 10 Customers by Total Revenue

The top 10 customers by revenue visualisation shows the customers that contribute the most towards the businesses revenue stream. One can see that a small number of customers contribute disproportionately to revenue, with several based in Los Angeles and San Francisco. For example, CUST1791 (Los Angeles) is the single largest customer, followed closely by others from the same region of Los Angeles.

The geographic concentration could introduce possible risks. If one or more of these top customers were to reduce their spending, the company could possible see a disproportionate decline in its total revenue, severely impacting the business. This also highlights the importance of strong and healthy customer relationships. Retaining and strengthening ties with top customers should be a priority, but equal focus should be put onto expanding the base of medium-sized revenue providing customers to try reduce over-reliance.

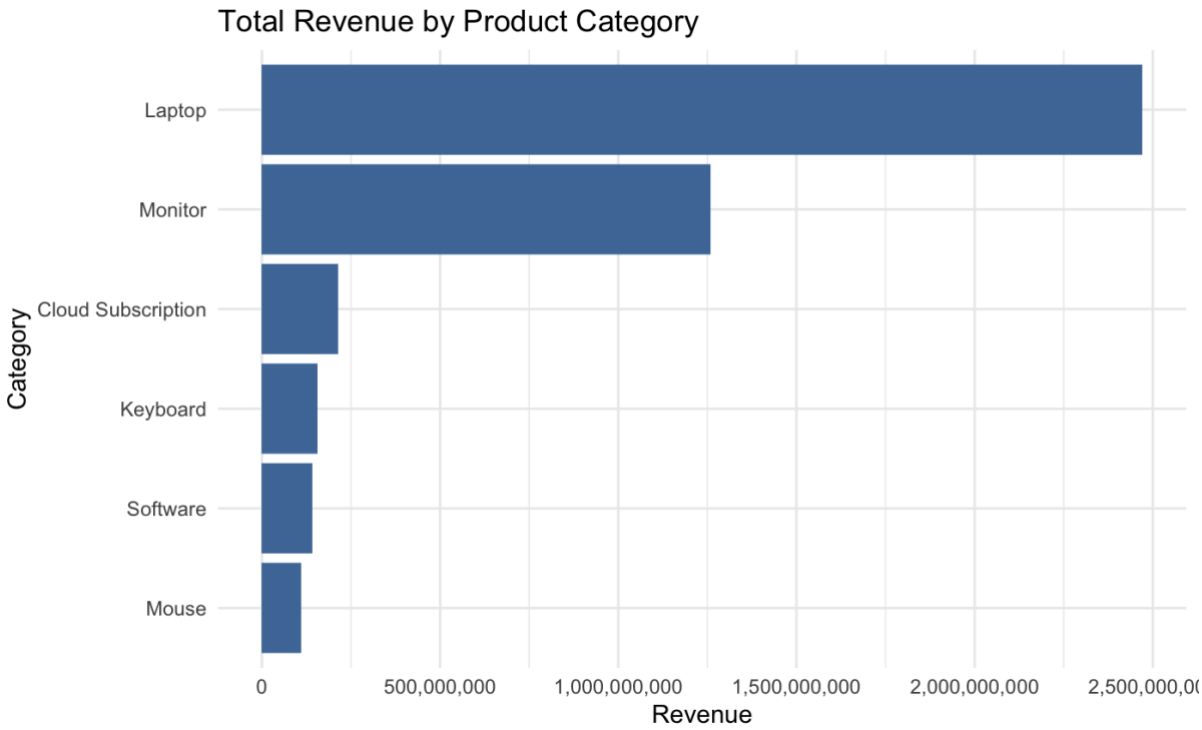


Figure 11 - Total Revenue by Product Category

The Revenue by Category chart seen above provides a broader view of how the different category of products influence the company's revenue stream. Laptops and Monitors dominate the total revenue contribution, with Software, Keyboards, Cloud Subscriptions and Mice making up a small amount of the revenue stream.

This imbalance shows a possible opportunity, while hardware seems to be the heart of the company's revenue stream, software and subscription services can generally offer higher margins and recurring revenue potential. Expanding into subscription offerings and actively marketing the, could improve both the profitability and revenue resilience of the business.

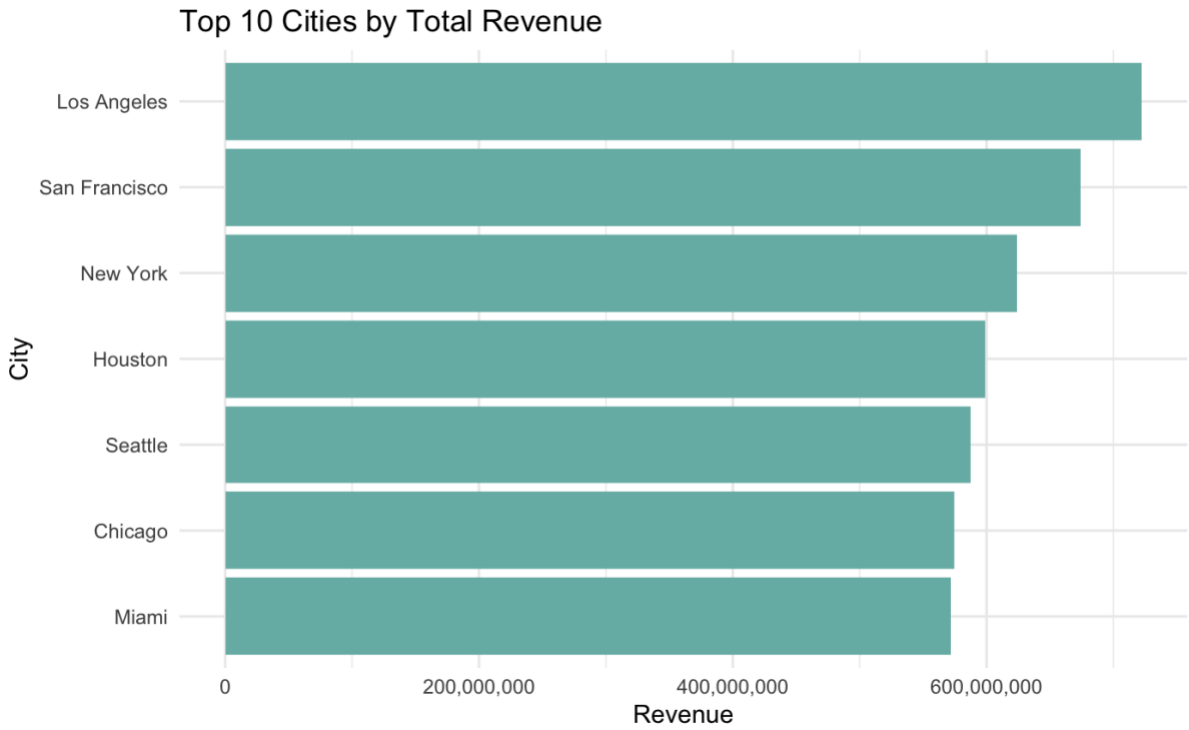


Figure 12 - Top 10 Cities by Total Revenue

The bar chart above shows Revenue by City, splitting the company's revenue into each city's contribution. The bar chart shows that Los Angeles, San Francisco and New York are the largest revenue centres. Houston, Seattle, Chicago and Miami also feature strongly, but at lower levels than the rest.

The implication of this is that sales are highly dependent and concentrated on a few large metropolitan areas. While these cities remain critical to the company's operations, significant growth opportunities may lie in expanding company presence in smaller or under-represented areas. The use of targeted marketing in mid-tier cities could diversify the revenue base and reduce the high dependency on a few geographical markets.

Looking at purely the sales data, the table below shows the mean is 13.50347, while the median is only 6. This means that the sales data distribution is right skewed and tells us that there are only a few high value transactions among a large amount of low value transactions. The standard variation is also very high when being compared to the mean, showing large variability in sales.

```
================ DESCRIPTIVE STATISTICS FOR SALES =================
Mean Sales:                    13.50347
Median Sales:                  6
Standard Deviation of Sales:   13.76013
Minimum Sales:                 1
Maximum Sales:                 50

====================================================================
```

This right skewness can be further verified by the histogram below. We can see that the majority of transactions are in the lower range, with higher sales being less frequent.

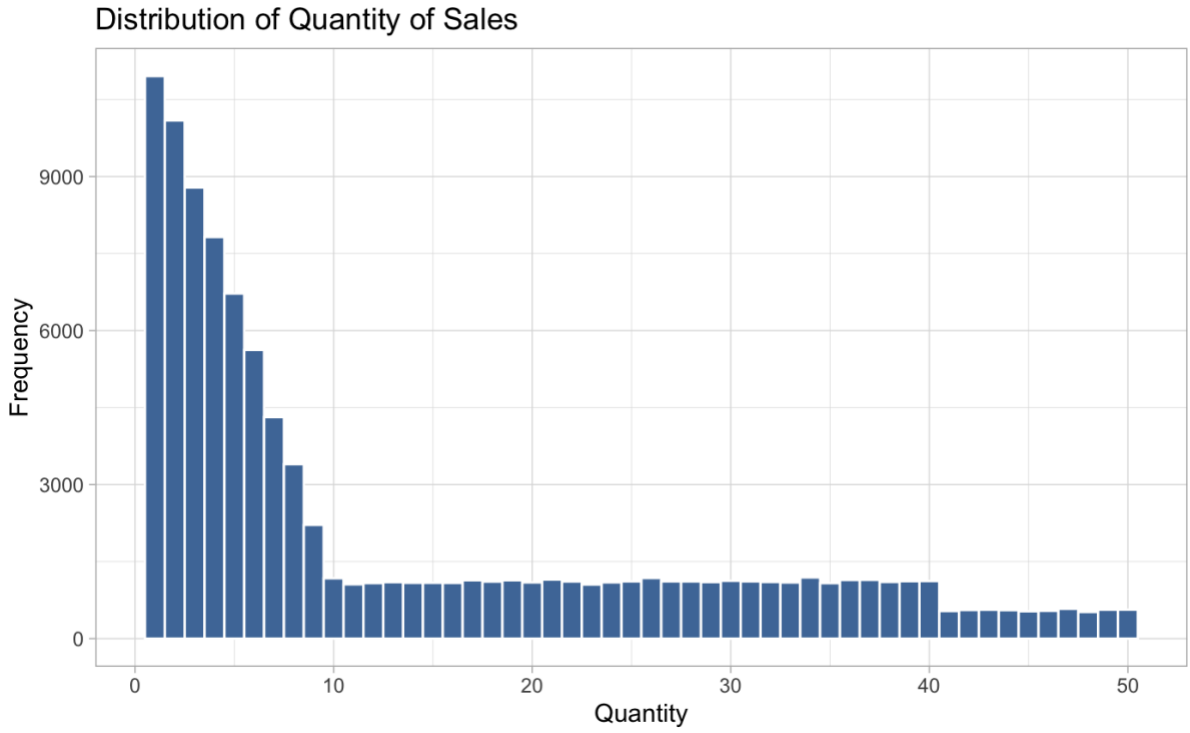


Figure 13 - Distribution of Quality of Sales

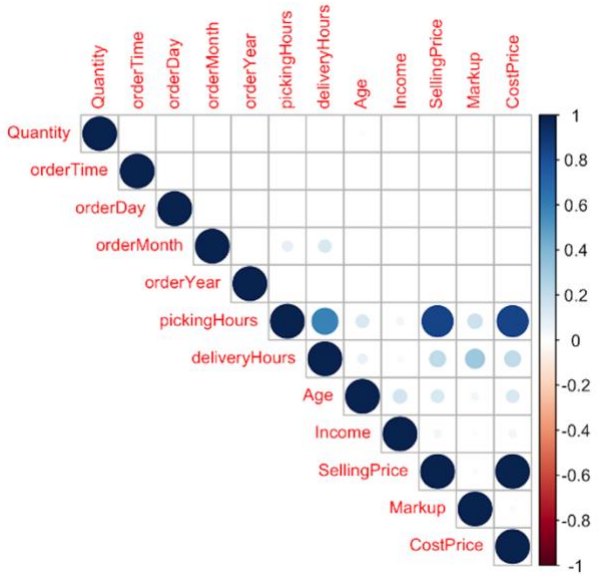


Figure 14 - Correlation Matrix

The correlation matrix on the left highlights a few clear and logical relationships in the data. The strongest correlation is between selling price and cost price, which is expected since cost price is derived from selling price and markup, confirming that the pricing data is consistent. There is also a strong link between picking hours and delivery hours, indicating that longer picking times generally lead to longer delivery times.

Moderate relationships between selling price, cost price, and markup suggest that higher-priced items tend to have higher markups and costs, which aligns with normal pricing behaviour. In contrast, variables such as quantity and order timing show little correlation with other factors, meaning they do not strongly influence operational or pricing patterns. Overall, the relationships shown are logical and confirm that the dataset is structured and functioning as expected.

The density plots below show the quantity by day, month and year. It is clear to see that the order quantities remain constant over year's, months and days, which is extremely valuable as it means that accurate forecasting can be done, leading to savings and lower costs on inventory, handling, stockouts and ensuring that sales are maximised. A consistent demand is very valuable and desirable, and it is shown that this company has such.

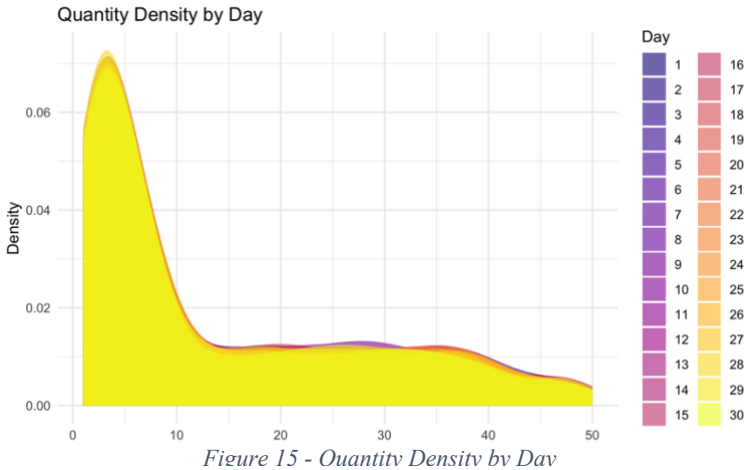


Figure 15 - Quantity Density by Day

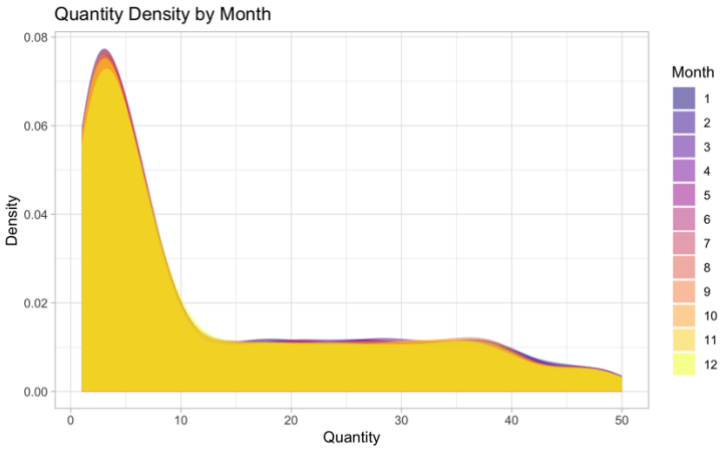


Figure 16 - Quantity Density by Month

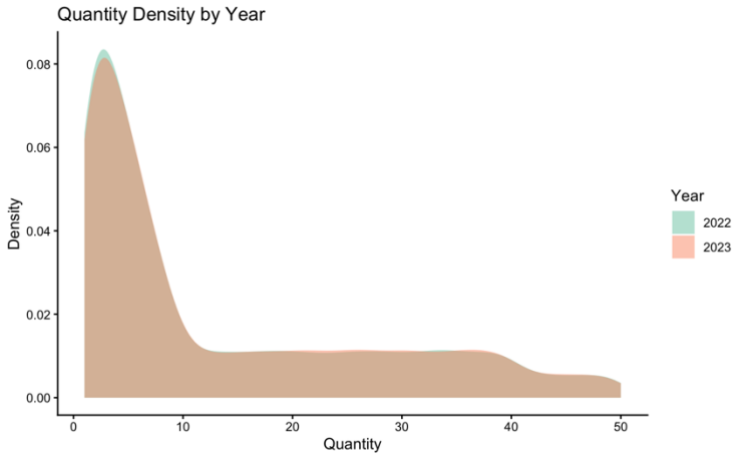


Figure 17 - Quantity Density by Year

This scatter plot below shows a clear positive relationship between picking and delivery hours, meaning that orders taking longer to pick also take longer to deliver. Laptops and monitors generally have the longest times, likely due to their higher value and handling needs, while software orders are completed almost instantly since they require no physical processing. The patterns are similar across 2022 and 2023, suggesting consistent operations, though some variation indicates differences in order complexity or workload.

Figure 18 - Picking Hours vs Delivery Hours by Category

3. Statistical Process Control (SPC)

3.1 Initialisation of Delivery-Time Processes (X-bar & s Charts)

During this section, we prepared the delivery-time data so it behaves like real time: for each product we sorted by year, month, day, and time, then split the timeline into back-to-back subgroups of 24 deliveries. Using only the first 30 subgroups, we calculated each subgroup's mean (X-bar) and its standard deviation (s). From those 30 values we computed the overall average of the subgroup means (X-double-bar) and the average subgroup standard deviation (s-bar). These graphs can be seen in Appendix A.

With standard SPC factors for n = 24 (A3, B3, B4) we set the fixed centre lines and three-sigma control limits: for the X-bar chart the centre line is X-double-bar and the limits are X-double-bar ± A3·s-bar; for the s chart the centre line is s-bar and the limits are B3·s-bar and B4·s-bar. We also added one-sigma and two-sigma guide bands by splitting the gap between the centre line and the three-sigma limits into thirds. This initial 30×24 window is Phase-1 (blue line in the graphs in Appendix A) and is used only to establish stable, fixed limits; it is plotted in blue.

3.2 Ongoing Control of Delivery-Time Processes (X-bar & s Charts)

Next, we continued down the timeline, still grouping data into subgroups of 24 (samples 31, 32, and onward) and, for each, calculating X-bar and s and plotting them against the fixed Phase-1 limits. This is Phase-2 monitoring and is shown by the green in the graphs in Appendix A . We always look at the s chart first; if a subgroup's spread is out of control, the X-bar result for that subgroup is not trusted. Any subgroup that exceeds its control limits is highlighted in red, and we also apply sensitising checks such as runs above the +2σ band to spot shifts and drifts. In practice this tells each product manager when the process appears stable and when it likely needs investigation or adjustment.

In graph A1: The X-bar chart for laptops shows a mean delivery time of roughly 19.5 hours, with points generally clustering around the centre line. Most samples remain within the 3-sigma limits, although several red points appear in the later phase, indicating out-of-control signals. This suggests that while the overall average performance is acceptable, the process mean tends to shift upward at certain times, possibly due to inconsistent workload or variations in operator efficiency. These shifts show that the process is not fully stable and may need closer monitoring to maintain consistent performance.

The S-chart shows that process variation is centred near 5.9 hours. During the early setup phase, variation is slightly higher, but in the control phase it stays mostly stable and within limits. A few spikes occur where the spread briefly increases, which could be linked to special causes such as resource shortages or temporary system delays. Overall, the laptop process appears mostly in control, but occasionally shows mean drift and small periods of increased variability.

In graph A2: The X-bar chart for monitors shows a mean delivery time of about 19.4 hours, with most data points clustering around the centre line. While many samples stay within the 3-sigma limits, there are several out-of-control points spread across the control phase, especially after sample 100 and again after sample 350. This pattern suggests that although the average delivery performance is generally acceptable, the process mean shifts upward at times, possibly because of uneven workload, order surges, or operational inconsistencies.

The S-chart shows that process variation is centred near 5.9 hours. The control phase remains mostly within the control limits, showing good consistency in spread. A few spikes appear throughout the chart, indicating brief increases in variability that may result from special causes such as delivery delays, equipment issues, or temporary resource shortages. Overall, the monitor process is largely in control, but some late-stage mean shifts and occasional increases in variation suggest the process could benefit from workload balancing and resource management.

In graph A3: The X-bar chart for mice shows a mean delivery time of about 19.2 hours, with most data points lying close to the centre line. While a large portion of the samples stay within the 3-sigma limits, several red points appear during the control phase, especially in the middle and towards the end of the series. This indicates that the process mean experiences occasional upward shifts, likely linked to fluctuations in workload or small delays in handling. Despite these out-of-control signals, the general process trend remains steady, suggesting that the average performance is mostly acceptable but could benefit from tighter monitoring to stop further drift.

The S-chart shows the process variation centred near 5.7 hours. Variation remains relatively consistent throughout both setup and control phases, with most samples well within the control limits. A few spikes in the spread occur at random points, which could be caused by temporary disruptions such as uneven staffing, equipment downtime, or short-term demand surges. Overall, the mice delivery process is fairly stable, but the presence of scattered out-of-control means and occasional variation spikes suggests that further attention to scheduling and process flow could improve long-term consistency.

In graph A4: The X-bar chart for keyboards shows a mean delivery time of about 19.2 hours, with most data points close to the centre line. Although many samples fall within the 3-sigma limits, several out-of-control points appear during the control phase, especially from around sample 200 – 400 and 600 - 750. This pattern tells us that the process mean shifts upward sometimes, due to changes in workload, scheduling, or operator performance. The clusters of red points indicate recurring deviations from normal operation, showing that the process occasionally becomes unstable even though the average delivery time remains acceptable overall.

The S-chart shows the process variation centred near 5.9 hours. Variation remains fairly consistent across both phases, with the majority of points within the control limits. A few short spikes occur, but there is no clear pattern of sustained instability. This means that the overall process variation is well controlled, and the fluctuations that do occur may result from temporary influences such as higher order volume or short resource interference. Overall, the keyboard process performs steadily, though attention should be given to the recurring upward mean shifts.

In graph A5: The X-bar chart for cloud subscriptions shows a mean delivery time of about 19.1 hours, with most points clustering near the centre line. While the majority of samples fall within the 3-sigma control limits, there are several out-of-control points during the control phase, particularly from around sample 150 – 350 and 500 - 650. These red points indicate that the process mean occasionally shifts upward, suggesting that certain periods experience slower delivery performance. This could be linked to increased system load, staff changes, or inconsistent scheduling.

The S-chart shows that the process variation is centred near 5.9 hours and remains fairly steady throughout both setup and control phases. Most samples are within limits, and although some spikes appear, they do not follow a strong pattern. The stable spread suggests that variation in delivery times is generally under control, with only short-term fluctuations likely caused by temporary delays or resource constraints.

In graph A6: The X-bar chart shows a mean delivery time of about 0.96 hours, with points mostly near the centre line. While most samples are within the 3-sigma limits, several out-of-control points appear after sample 200, showing slight upward shifts in the mean. These may be caused by workload peaks or short delays, suggesting mild instability despite generally good performance.

The S-chart shows variation centred around 0.30 hours, remaining steady across both phases. Most points stay within control limits, with only small, random spikes. This indicates that variation is well controlled and predictable. Overall, the software process is stable, with minor mean shifts that should be monitored to maintain consistent delivery.

3.3 Capacity Analysis

Process capability indices are important in assessing whether a delivery process can meet the customer's needs, expectations and performance requirements for a service or product. The Cp value represents the potential capability of a process, assuming it is perfectly centred between the Lower and Upper Specification Limits. It reflects how wide or narrow the process spread is compared to these limits. The Cpk value, on the other hand, measures the actual capability of the process by considering how well the mean is centred within the specification limits. A Cpk value of 1.30 or higher suggests that the process has the potential to meet customer requirements. Cpu and Cpl form key components of Cpk, indicating performance relative to the upper and lower specification boundaries. Cpu evaluates how close the process comes to the upper limit, where a lower value signals a risk of exceeding this limit, while Cpl indicates how well the process stays above the lower limit, with lower values showing potential underperformance.

For cloud services (CLO), the Cp value of 0.90 shows that the process spread is wider than the specification limits, suggesting inconsistent delivery performance. The Cpk value of 0.72 indicates that the process mean is not well centred, resulting in deliveries that may exceed the upper limit. The Cpu value of 0.72 confirms this, showing a higher chance of delays beyond the acceptable range, while the Cpl value of 1.08 shows relatively stable control below the lower limit. CLO (Cloud Services) is not capable of satisfying the VOC and requires process adjustments to reduce variation and improve centring.

For keyboards (KEY), a Cp value of 0.92 suggests a process with medium potential to meet the specification limits. However, a Cpk value of 0.73 reveals that the mean delivery time is not centred, causing inconsistent performance. The Cpu value of 0.73 indicates that deliveries occasionally exceed the upper limit, while a Cpl value of 1.10 shows stable performance below the lower limit. The process is close to acceptable capability but needs better control of average delivery times to consistently meet customer expectations.

For laptops (LAP), the Cp value of 0.90 shows that the process has a wider spread than the set limits, which means that deliveries vary more than desired. The Cpk value of 0.70 indicates that the process mean is off-centre, contributing to delivery delays. The Cpu value of 0.70 highlights a recurring tendency to exceed the upper limit, while the Cpl value of 1.10 shows good control below the lower limit. LAP (Laptops) therefore not capable of meeting the VOC and would benefit from process optimisation to achieve more consistent results.

For monitors (MON), the Cp value of 0.89 shows that the process spread is just below the acceptable benchmark, while the Cpk value of 0.70 confirms that the process mean is poorly centred. The Cpu value of 0.70 suggests frequent risks of exceeding the upper limit, and the Cpl value of 1.08 shows that the process remains steady below the lower limit. MON (Monitors) is not capable of meeting the VOC, as both process centring and spread need improvement to attain stable and predictable delivery times.

For mice (MOU), the Cp value of 0.92 indicates that the process has a moderate spread relative to the limits, but the Cpk value of 0.73 shows that it is not centred effectively. The Cpu value of 0.73 suggests that the process occasionally exceeds the upper specification limit, while the Cpl value of 1.10 shows minimal variation below the lower limit. MOU (Mice) is not capable of satisfying the VOC, and small improvements in centring and reducing variation could help the process perform within the desired range.

For software (SOF), the Cp value of 18.14 shows that the process variation is extremely small compared to the specification limits, indicating a very tight and efficient process. The Cpk value of 1.08 suggests that the process is almost centred within the limits, although slightly below the typical capability benchmark of 1.30. The Cpu value of 35.19 shows that the process rarely exceeds the upper limit, and the Cpl value of 1.08 confirms steady control above the lower limit. Although SOF (Software) narrowly misses the capability threshold, it remains a highly stable and efficient process with strong overall performance.

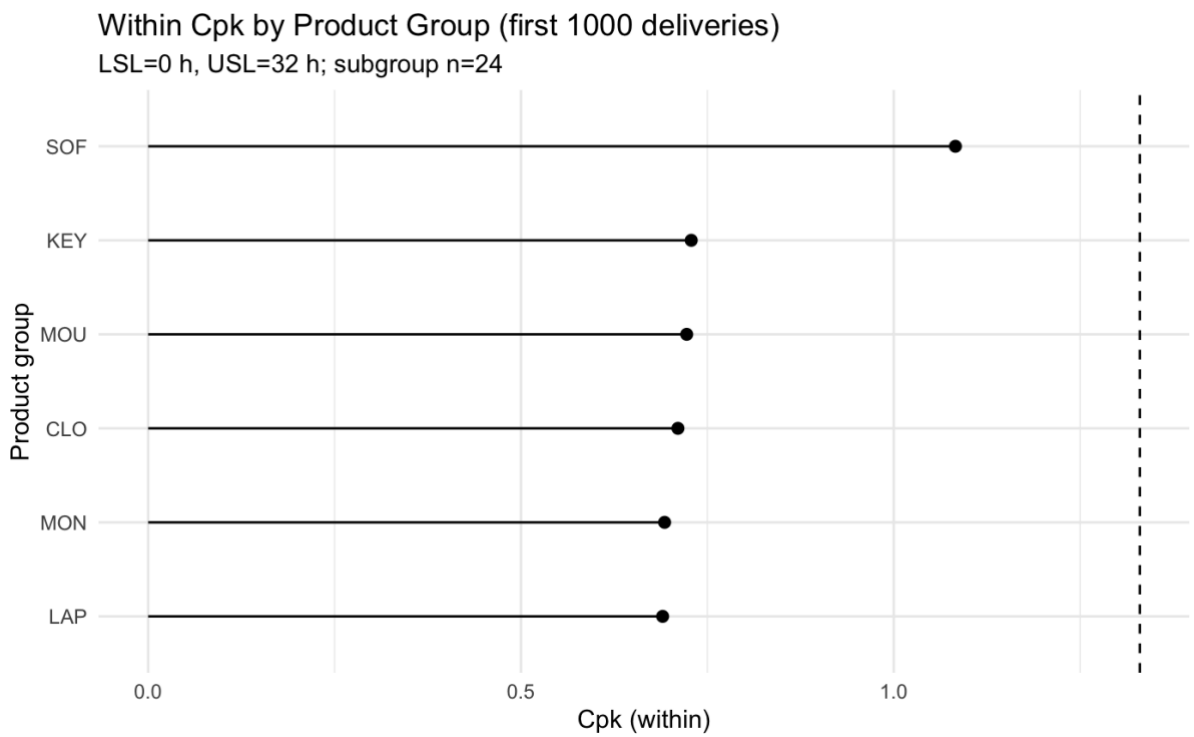


Figure 19 - CPK per Product Group

The images below show a summary of the process capability indices for each product type.

```
=== Capability (first 1000 deliveries) — compact table (rounded) ===

Product Category: CLO
Mean Delivery Hours: 19.23 hours
Standard Deviation: 5.94 hours

Cp: 0.90
Cpu: 0.72
Cpl: 1.08
Cpk: 0.72

Product Category CLO is NOT capable of satisfying the VOC.

Product Category: KEY
Mean Delivery Hours: 19.28 hours
Standard Deviation: 5.82 hours

Cp: 0.92
Cpu: 0.73
Cpl: 1.10
Cpk: 0.73

Product Category KEY is NOT capable of satisfying the VOC.

Product Category: LAP
Mean Delivery Hours: 19.61 hours
Standard Deviation: 5.93 hours

Cp: 0.90
Cpu: 0.70
Cpl: 1.10
Cpk: 0.70

Product Category LAP is NOT capable of satisfying the VOC.

Product Category: MON
Mean Delivery Hours: 19.41 hours
Standard Deviation: 6.00 hours

Cp: 0.89
Cpu: 0.70
Cpl: 1.08
Cpk: 0.70

Product Category MON is NOT capable of satisfying the VOC.

Product Category: MOU
Mean Delivery Hours: 19.30 hours
Standard Deviation: 5.83 hours

Cp: 0.92
Cpu: 0.73
Cpl: 1.10
Cpk: 0.73

Product Category MOU is NOT capable of satisfying the VOC.

Product Category: SOF
Mean Delivery Hours: 0.96 hours
Standard Deviation: 0.29 hours

Cp: 18.14
Cpu: 35.19
Cpl: 1.08
Cpk: 1.08

Product Category SOF is NOT capable of satisfying the VOC.
```

3.4 Identification of process control issues (Rules A–C)

In this question, it was asked that the data should be analysed for any possible process control issues according to the following rules.

Rule A (1 s sample outside +3σ limits): If a sample's standard deviation (s) exceeds the upper 3-sigma control limit, it signals excessive variation. This indicates that the process spread has increased unexpectedly and may be out of control.

Rule B (Longest run of s within ±1σ limits): A long sequence of consecutive s values within one sigma of the centre line shows consistent variability. This is a sign of good control and process stability.

Rule C (Four or more consecutive X-bar samples above +2σ) : When four or more sample means fall above the upper 2-sigma line, it suggests a shift or upward trend in the process mean. This may indicate that the process average has moved and requires investigation.

These were the results:

```
===== Q3.4 – Control Rule Summary by Category =====

Category: Cloud Subscription
A) s above +3σ (UCL): total=0; first3={}; last3={}
B) longest run s within ±1σ: length=35, start=474, end=508
C) x-bar runs over upper 2σ (>=4 in a row): runs=20, points=263; first3={122-125; 179-183; 192-200}; last3={567-602; 604-626; 628-649}

Category: Keyboard
A) s above +3σ (UCL): total=0; first3={}; last3={}
B) longest run s within ±1σ: length=15, start=730, end=744
C) x-bar runs over upper 2σ (>=4 in a row): runs=25, points=294; first3={112-117; 172-175; 187-191}; last3={698-719; 721-724; 726-746}

Category: Laptop
A) s above +3σ (UCL): total=0; first3={}; last3={}
B) longest run s within ±1σ: length=19, start=116, end=134
C) x-bar runs over upper 2σ (>=4 in a row): runs=12, points=159; first3={119-122; 130-140; 154-167}; last3={361-369; 374-391; 393-425}

Category: Monitor
A) s above +3σ (UCL): total=0; first3={}; last3={}
B) longest run s within ±1σ: length=34, start=238, end=271
C) x-bar runs over upper 2σ (>=4 in a row): runs=23, points=226; first3={134-137; 179-182; 190-194}; last3={580-608; 610-613; 615-618}

Category: Mouse
A) s above +3σ (UCL): total=1; first3={592}; last3={592}
B) longest run s within ±1σ: length=16, start=672, end=687
C) x-bar runs over upper 2σ (>=4 in a row): runs=23, points=324; first3={194-197; 235-239; 280-286}; last3={777-805; 811-842; 844-860}

Category: Software
A) s above +3σ (UCL): total=0; first3={}; last3={}
B) longest run s within ±1σ: length=21, start=659, end=679
C) x-bar runs over upper 2σ (>=4 in a row): runs=25, points=334; first3={202-205; 237-240; 244-247}; last3={774-801; 803-840; 842-864}
```

4.1 Type 1 Error for Rules A–C

For rule A, the probability of a type 1 error is calculated using the formula:

$$a = 2 \times [1 - \Phi(3)]$$

For one point outside the ±3 sigma limits the Type I error is based on the two-tailed probability of the normal distribution beyond ±3 standard deviations. The value of Φ(3) = 0.9987, where Φ(3) is the cumulative probability of the standard normal up to z = 3. This equations gets you a probability of 0.0027. This means that even if the process is stable, there will be at least 3 out of 1000 points that will fall outside the limits by chance.

For rule B, the probability of a type 1 error is calculated using the formulas:
The probability that one point lies ± 1 sigma: $P = P(-1 < Z < 1) = \Phi(1) - \Phi(-1)$ = 0.6827 and the chance that n consecutive s values lie within that range is: $P = (0.6827)^n$. But since this pattern is a sign of control rather than an alarm, the type 1 error $a \approx 0$.

For rule C, the probability of a type 1 error is calculated using the formulas:

$$a = [1 - \Phi(2)]^n$$

For four consecutive X-bar points beyond the +2 sigma on the same side the type 1 error is the one sided probability raised to the 4th power, $a = [1 - \Phi(2)]^4 = [0.0228]^4 = 2.7 \times 10^{-7}$. This means that the chance of four such points occurring in a stable process is about 3 in 10 million

Rule	Description	Type 1 Error (a)	Meaning
A	One sample outside ±3σ limits	0.0027 ≈ 0.27 %	3 in 1000 points trigger false alarm
B	Run within ±1σ (likely good control)	≈ 0	none (rule identifies stability)
C	Four X-bars beyond ±2σ same side	0.000000268 ≈ 0.0000268 %	≈ 3 in 10 million runs false alarm

4.2 Type 2 Error for Bottle Filling Process

For the bottle filling process, the probability of a type II error is calculated using the formula:

$$\beta = \Phi\left(\frac{UCL - \mu_1}{\sigma^-}\right) - \Phi\left(\frac{LCL - \mu_1}{\sigma^-}\right)$$

This formula above provides the probability that the new mean and variation still fall within the old control limits, even though the process has shifted. The control limits are UCL = 25.089 and LCL = 25.011 litres, the new process mean is μ_1 = 25.028 litres and the new standard deviation of the X – bar chart is σ^-= 0.017 litres. Substituting these values gives us Z_l = (25.011−25.028)/0.017 = −1.00 and Z_u = (25.089−25.028)/0.017 = 3.588.

Therefore, the probability is then $\beta = \Phi(3.588) - \Phi(-1.00)$ = 0.99983 – 0.15865 = 0.8412. This means that there is an 84.12% chance that the process will not be identified as out of control even though it has shifted. The power of the chart (the ability to detect the shift) is then 1 – β = 0.1588 = 15.88%. This shows that the chart is not very sensitive to small mean shifts and that the process could continue producing slightly incorrect fills without being flagged.

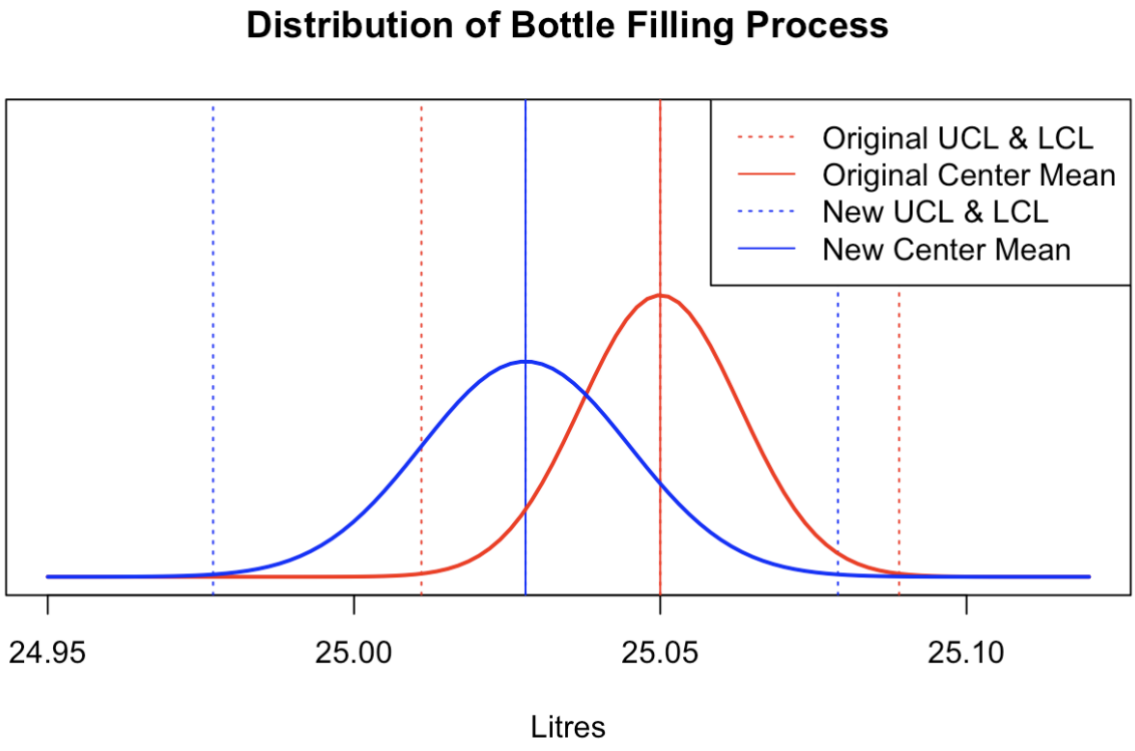


Figure 20 - Distribution of Bottle Filling Process

4.3 Fixing Head Office Data and Reapplying Data Analysis

In this step of the project, the product and head office data files were corrected and the discrepancies were handled following the instructions of the head office. The incorrect “NA” product codes were all given the correct prefixes relative to their product categories, such as “SOF”. The selling prices and markup percentages were also standardised using the first 10 reference values from the local products_data.csv file. These values were then repeated every 10 records so that product types 11 – 60 matched the correct models. Both corrected datasets (product_data2025 and products_Headoffice2025) were then saved and the main data analysis was ready to be redone with the now correct data.

In step 1.3, I already noticed the problem in the product data and corrected it as I thought that it was a mistake. I then did some logical tests for the changes that I made in the head office data. It seemed that all of the product ID’s in the sales data were in the product data. This means that none of the changes made to the head office data will have influenced revenue, or other sales metrics. This was further confirmed by the visualisations in the code.

```
================ PRODUCT ID CHECK ================
Total Product IDs in products_data2025:  60
Total Product IDs in sales2022and2023:   60
Matching Product IDs:                    60
Unmatched Product IDs (in sales only):   0

---- Matching Product IDs ----
 [1] "CLO011" "CLO012" "CLO013" "CLO014" "CLO015" "CLO016" "CLO017" "CLO018" "CLO019" "CLO020" "KEY041" "KEY042" "KEY043" "KEY044" "KEY045"
[16] "KEY046" "KEY047" "KEY048" "KEY049" "KEY050" "LAP021" "LAP022" "LAP023" "LAP024" "LAP025" "LAP026" "LAP027" "LAP028" "LAP029" "LAP030"
[31] "MON031" "MON032" "MON033" "MON034" "MON035" "MON036" "MON037" "MON038" "MON039" "MON040" "MOU051" "MOU052" "MOU053" "MOU054" "MOU055"
[46] "MOU056" "MOU057" "MOU058" "MOU059" "MOU060" "SOF001" "SOF002" "SOF003" "SOF004" "SOF005" "SOF006" "SOF007" "SOF008" "SOF009" "SOF010"

All ProductIDs in sales are present in products_data2025.

==================================================
```

5 Optimisation:

5.1 Shop 1:

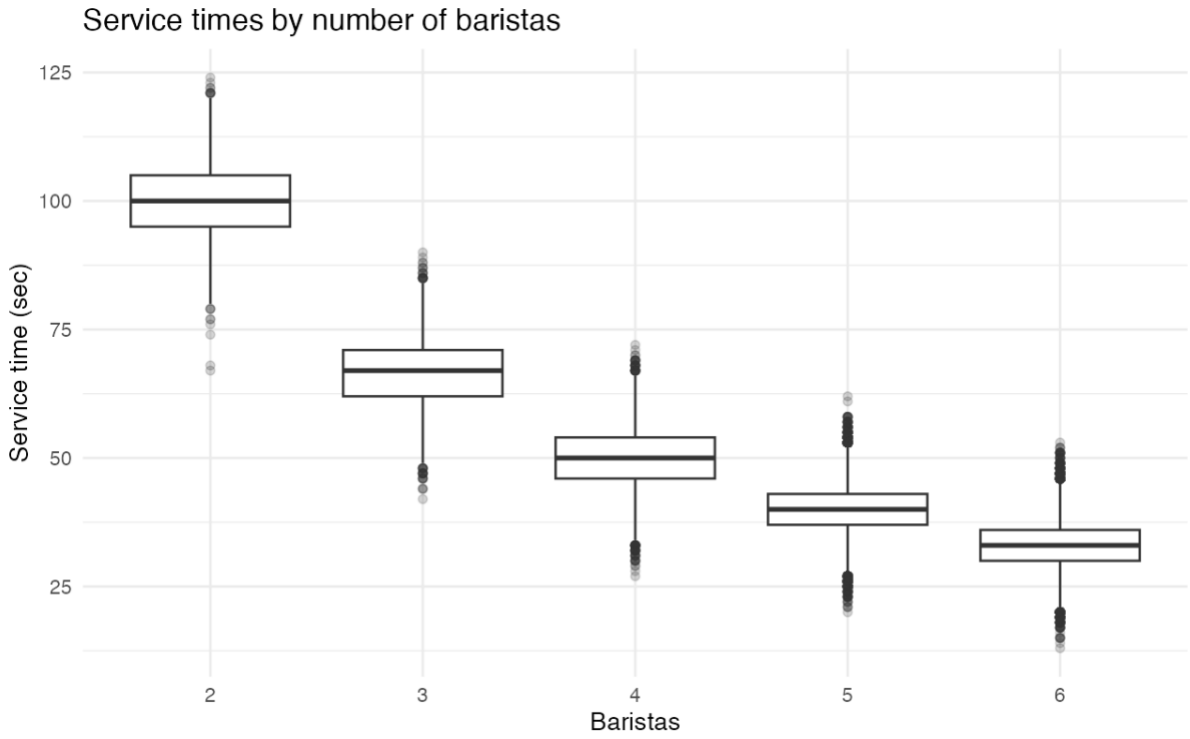


Figure 22 - Service Times vs Barristers (Shop 1)

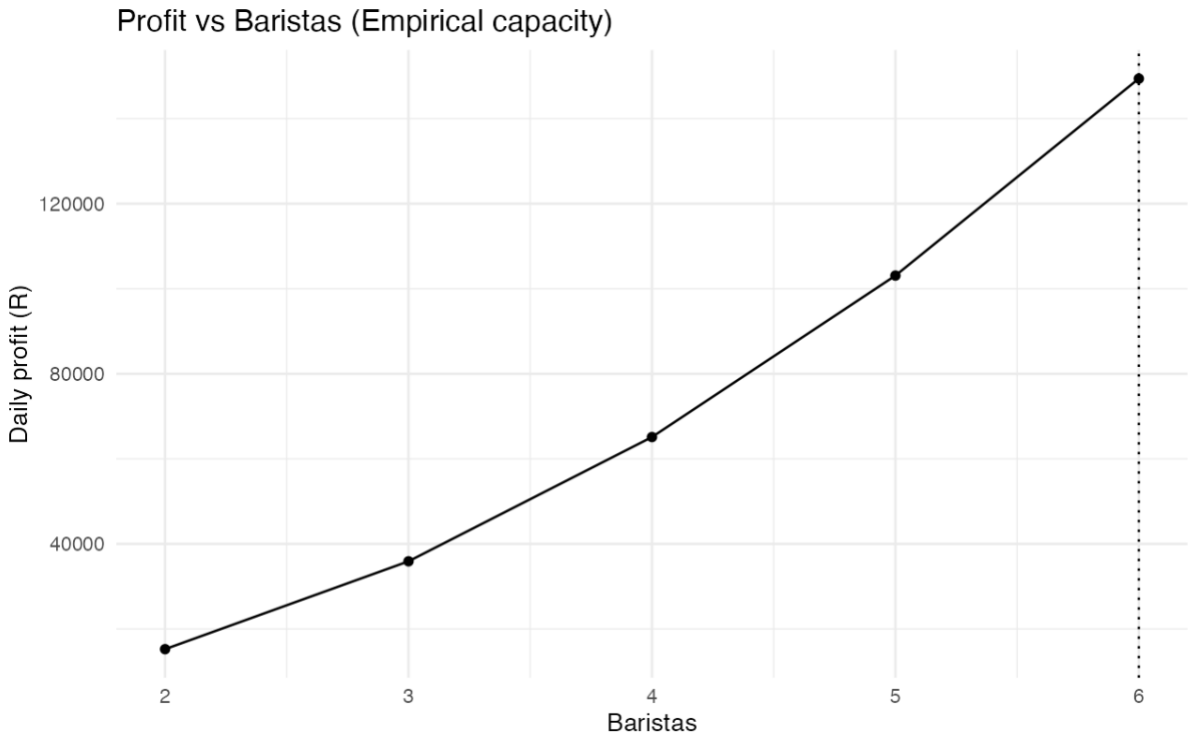


Figure 21 - Profit vs Barristers (Shop 1)

Baristas <int>	DailyRevenue <dbl>	DailyCost <dbl>	DailyProfit <dbl>
2	17250	2000	15250
3	38910	3000	35910
4	69120	4000	65120
5	108090	5000	103090
6	155400	6000	149400

Using the service-time data for shop 1, the analysis shows a strong negative relationship between the number of baristas and the average time to make a coffee, dropping from about 100 seconds with two baristas to around 35 seconds with six. As speed improves, so does reliability: under a 60-second Service Level Agreement (SLA), the model predicts that roughly 95 percent of customers will receive reliable service once four or more baristas are working, with reliability effectively reaching 100 percent at six baristas.

When profits were modelled by combining customer throughput, reliability, and staffing costs, daily profit continued to rise with each added barista. The maximum profit occurs at six baristas, where efficiency gains and high customer turnover outweigh the extra labour cost. Therefore, scheduling six baristas per weekday provides the best financial return while ensuring nearly all clients experience timely service.

5.2 Shop 2:

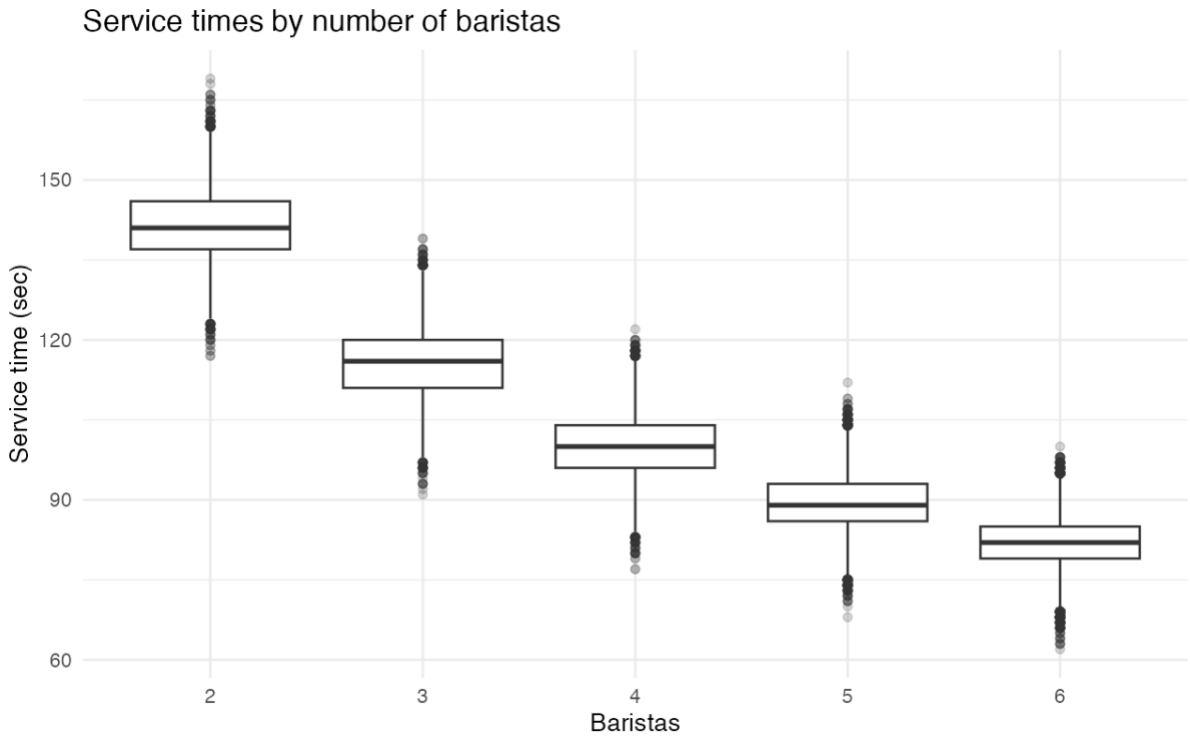


Figure 24 - Service Times vs Barristers (Shop 2)

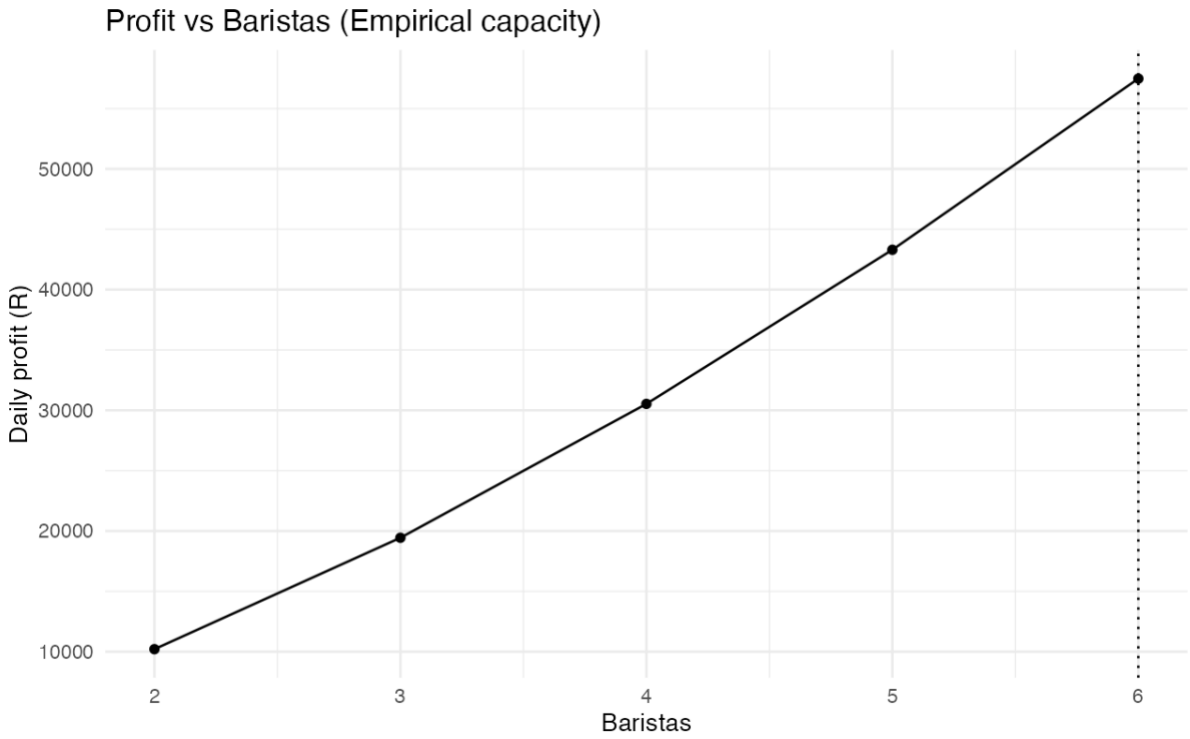


Figure 23 - Profit vs Barristers (Shop 2)

Baristas <int>	DailyRevenue <dbl>	DailyCost <dbl>	DailyProfit <dbl>
2	12210	2000	10210
3	22440	3000	19440
4	34530	4000	30530
5	48300	5000	43300
6	63480	6000	57480

The same result can be taken from shop 2's data. The service times have a negative relationship with number of barristers. When one barrister is used, the service time looks like it sits at around 140 second. This then drops to around 82 seconds when 6 barristers are employed. This means that more customers can be served per day, increasing the possible daily profit.

This can be confirmed by the profit table, where we can see the highest daily profit occurs when 6 barristers are employed. This profit sits at R57 480 per day and the graph above this table proves why. We can see that there is almost no diminishing returns, so the more barristers employed, the more profit gets made.

6 DOE and ANOVA:

6.1

I came up with the following question to be analysed by the ANOVA test:
Is there a significant difference in the mean delivery hours for the 4 physical products groups (Laptops, Monitors, Keyboards and Mice) that are offered by the business?

The Null Hypothesis (H0): There is no significant difference in the mean delivery hours among the physical product groups. (H0 : $\mu_{LAP}=\mu_{MON}=\mu_{KEY}=\mu_{MOU}$)

The Alternative Hypothesis (H1): At least one of the physical product groups has a significantly different mean delivery hour compared to the others.
H1: At least one μ_i differs

6.2

These are the results from the ANOVA test:

Source <chr>	Df <dbl>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
product_group	3	34.216	11.405	0.308	0.82
residuals	63649	2359638.566	37.073	*NA*	*NA*

These are the summary statistics:

product_group <fctr>	n <dbl>	mean <dbl>	sd <dbl>
Laptop	10207	21.782	6.048
Monitor	14864	21.739	6.047
Keyboard	17920	21.744	6.091
Mouse	20662	21.790	6.136

Looking at the summary statistics, we can see that all the products (LAP, MON, KEY and MOU) have very similar means and standard deviations. This tells us that the variation within each product group is relatively similar.

The ANOVA test gives a F-value of 0.308, which indicates the ratio of variance within each product type to the variance between the four product types. Since the F-value is small, we know that the difference in delivery time between product types is not substantial.
The ANOVA test also gives us another result, this being the p-value. The p value for this test is 0.82, which is much higher than the common significance level of 0.05. This tells us that we cannot reject the null hypothesis and that there is no statistically significant difference between the 4 product types when looking at delivery times.

These findings suggest that the delivery process operates consistently across all physical product groups. No product category takes significantly longer or shorter to deliver than the others, indicating that delivery times are stable and well-controlled across the different product types.

7 Reliability of Service:

7.1

Given the information, we know that problems occur when fewer than 15 staff are on duty. This means that when 15 or 16 staff are working, the service will be reliable. Out of the 397 observed days, there were 96 days with 15 staff and 270 days with 16 staff. This means that 366 out of the 397 days were reliable, giving a reliability of 366/397, which is equal to 0.922. That gives us roughly 337 reliable days in a year and 28 unreliable days in a year.

7.2

A binomial model was used to identify the number of workers that minimises total monthly costs while maintaining reliability. As more workers are scheduled, the chance of having too few people on duty decreases, but the wage cost increases. The results show that the total monthly cost is lowest when seventeen workers are scheduled, at about R430000 per month. Fewer than seventeen workers lead to losses from unreliable service, while more than seventeen increases costs without much improvement in reliability.

The graph below shows that the total cost curve has a minimum point at seventeen workers. This follows the Taguchi Loss Principle, which states that costs rise when conditions move away from the ideal balance point. Under staffing leads to lost sales, and over staffing leads to unnecessary labour costs. Having seventeen workers provides the best balance between reliability and cost efficiency.

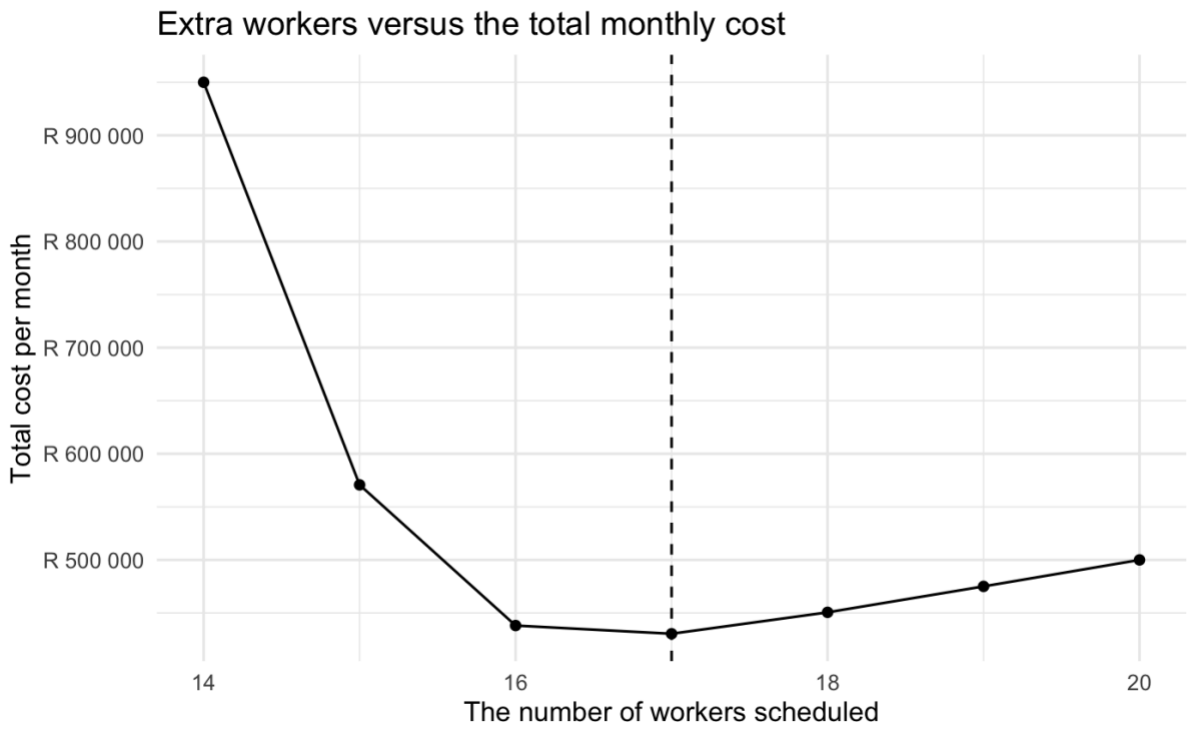


Figure 25 - Workers vs Total Monthly Cost

workers <dbl>	p_problem <dbl>	reliable_pct <dbl>	expected_loss_month <dbl>	staffing_cost_month <dbl>	total_monthly_cost <dbl>
14	1.0000	0.0000	600000.00	350000	950000.0
15	0.3262	0.6738	195707.53	375000	570707.5
16	0.0636	0.9364	38178.59	400000	438178.6
17	0.0091	0.9909	5442.72	425000	430442.7
18	0.0010	0.9990	624.08	450000	450624.1
19	0.0001	0.9999	60.82	475000	475060.8
20	0.0000	1.0000	5.22	500000	500005.2

7 Conclusion

This report investigated the company's delivery performance, process stability, and staffing efficiency using a range of statistical and analytical methods. The analysis of customer, product, and sales data gave insight into revenue concentration, pricing structure, and customer demographics. From the SPC results, it was seen that software delivery is highly stable and efficient, while physical product types such as monitors, laptops, keyboards, and mice showed signs of mean drift and small periods of increased variation. The process capability study confirmed that most product categories are not yet capable of consistently meeting customer requirements, suggesting that improvements in centring and variation control are needed.

The Type I and Type II error analysis demonstrated how sensitive the control charts are to small process shifts, highlighting the importance of continuous monitoring to detect performance changes early. The corrections made to the head office data ensured that all product information was standardised and accurate for further analysis. The optimisation analysis for the coffee shops showed that six baristas provide the best balance between efficiency, profit, and customer service reliability. Similarly, the staffing reliability model for the car rental agency indicated that scheduling seventeen workers achieved the lowest total monthly cost while maintaining high reliability.

The ANOVA test comparing mean delivery times across physical product types showed no significant differences, confirming that the company's delivery process is consistent across these categories. Overall, this report demonstrated that while most systems operate reliably, targeted improvements in process centring, resource allocation, and monitoring can further enhance performance, profitability, and customer satisfaction.

9. References

DataCamp. (2024). What is Data Analysis? Expert Guide.* [online] Available at: https://www.datacamp.com/blog/what-is-data-analysis-expert-guide [Accessed 1 October 2025].

Creighton, S. (2023). What are control limits in an SPC chart? [online] blog.lifeqisystem.com. Available at: https://blog.lifeqisystem.com/control-limits-in-spc-chart [Accessed 7 October 2025].

Quality-One International. (n.d.). Statistical Process Control (SPC). [online] Available at: https://quality-one.com/spc/ [Accessed 7 October 2025].

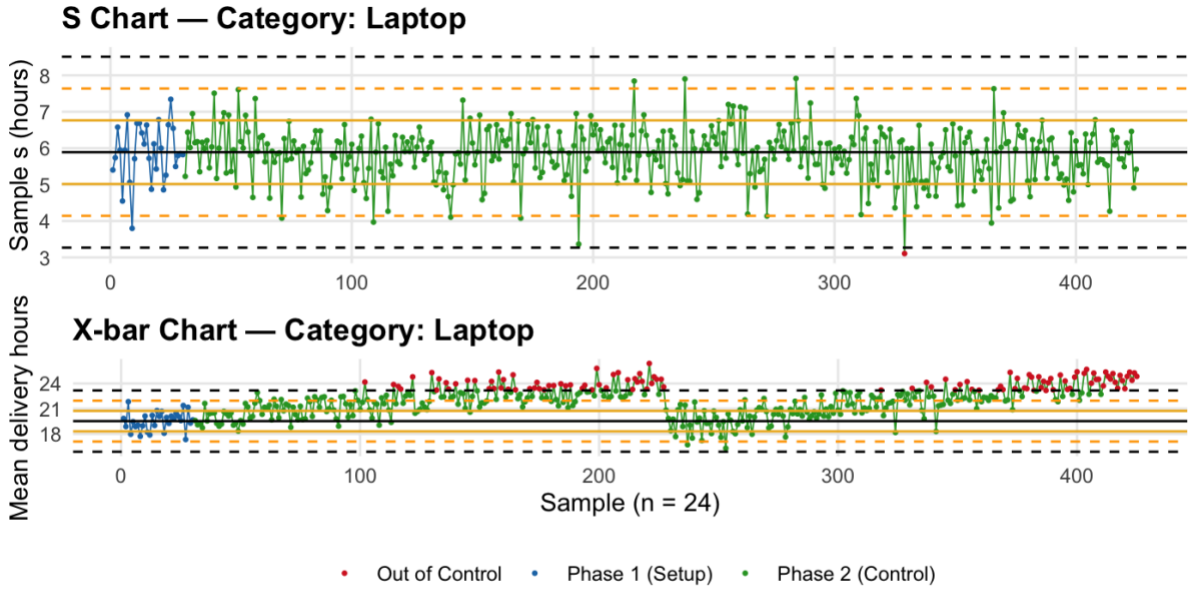
1Factory. (n.d.). A Guide to Process Capability (Cp, Cpk, Pp, Ppk). [online] Available at: https://www.1factory.com/quality-academy/guide-process-capability.html [Accessed 7 October 2025].

Amplitude. (2024). Type 1 and Type 2 Errors Explained. [online] Available at: https://amplitude.com/explore/experiment/type-1-and-type-2-errors-explained [Accessed 14 October 2025].

University of Virginia Library. (2024). Understanding t-tests, ANOVA, and MANOVA. [online] Available at: https://library.virginia.edu/data/articles/understanding-t-tests-anova-and-manova [Accessed 19 October 2025].

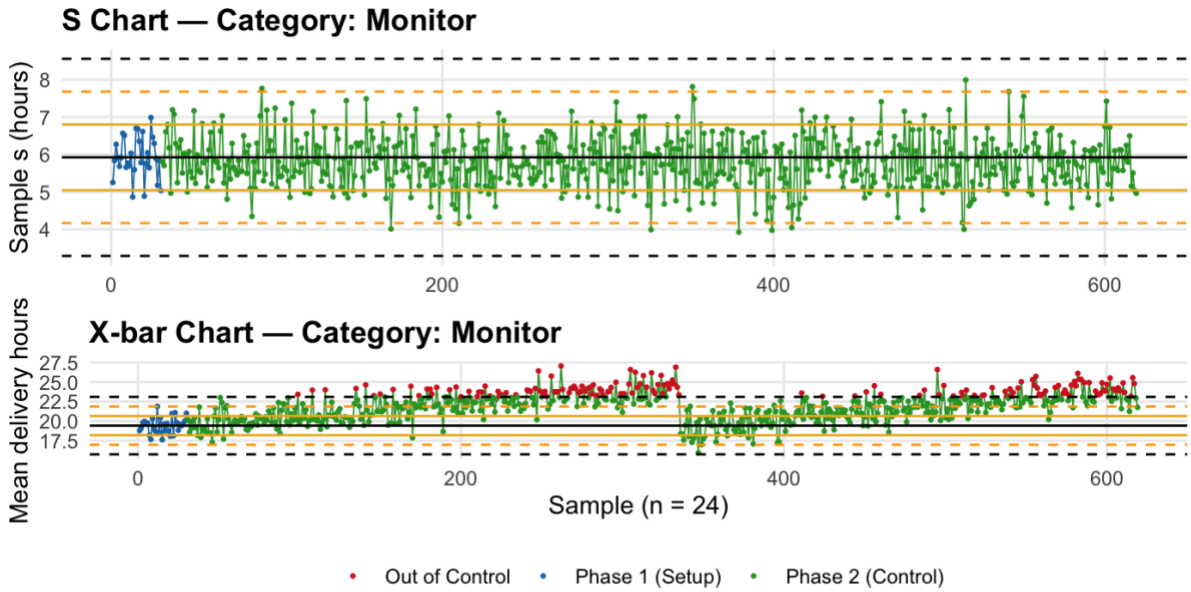
OpenAI. (2025). *ChatGPT. [online] Available at: https://chatgpt.com/ [Accessed 1 October 2025].

Appendix A



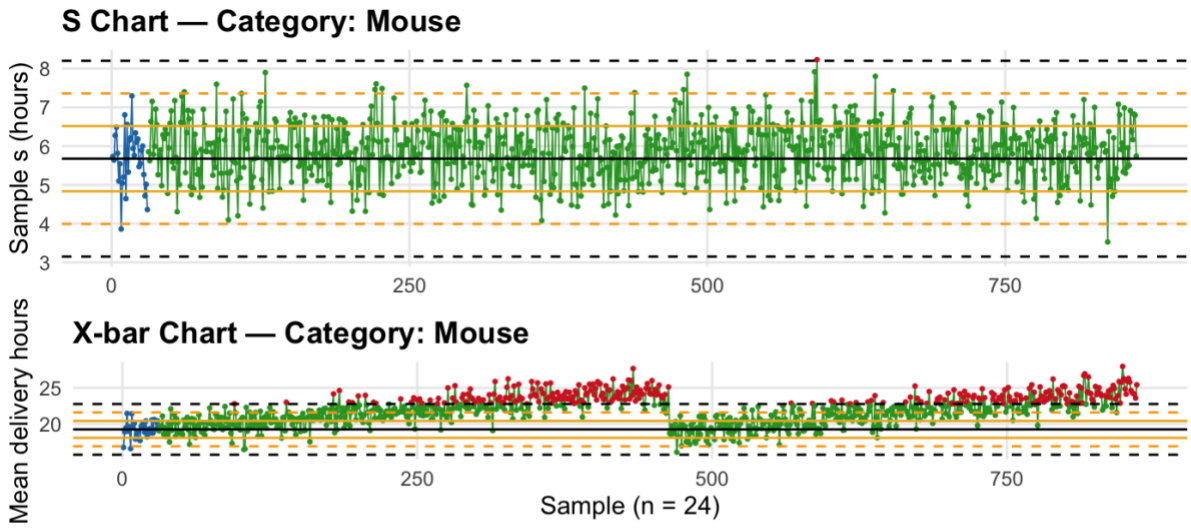
Measure	CL	L1	L2	LCL	U1	U2	UCL
$\bar{X}$	19.524	18.308	17.093	15.878	20.739	21.955	23.170
s	5.890	5.017	4.143	3.269	6.764	7.638	8.512

Figure 26 - Graph A1



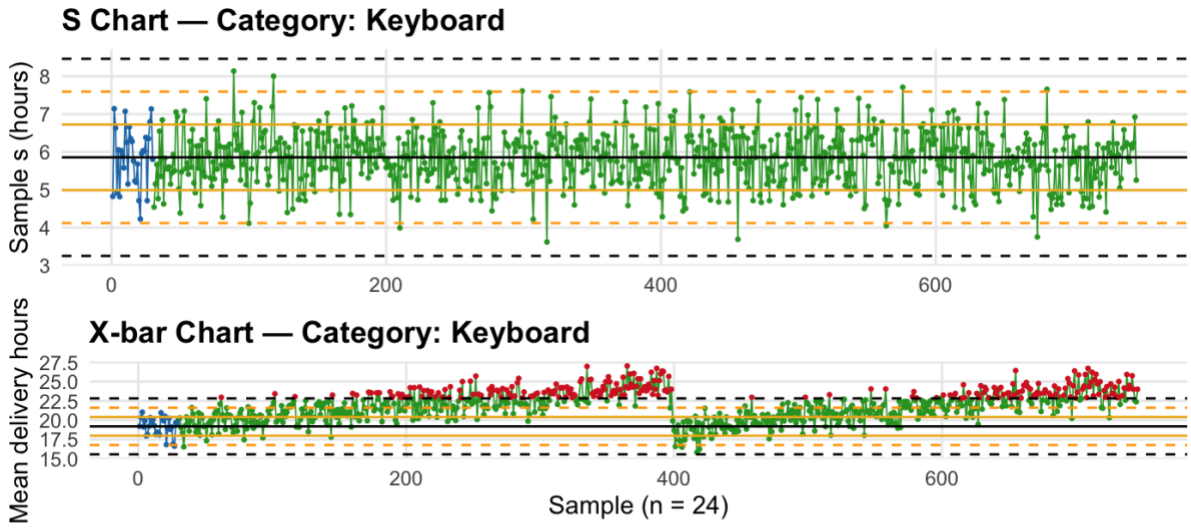
Measure	CL	L1	L2	LCL	U1	U2	UCL
$\bar{X}$	19.426	18.204	16.982	15.760	20.648	21.87	23.092
s	5.923	5.045	4.166	3.287	6.802	7.68	8.559

Figure 27 - Graph A2



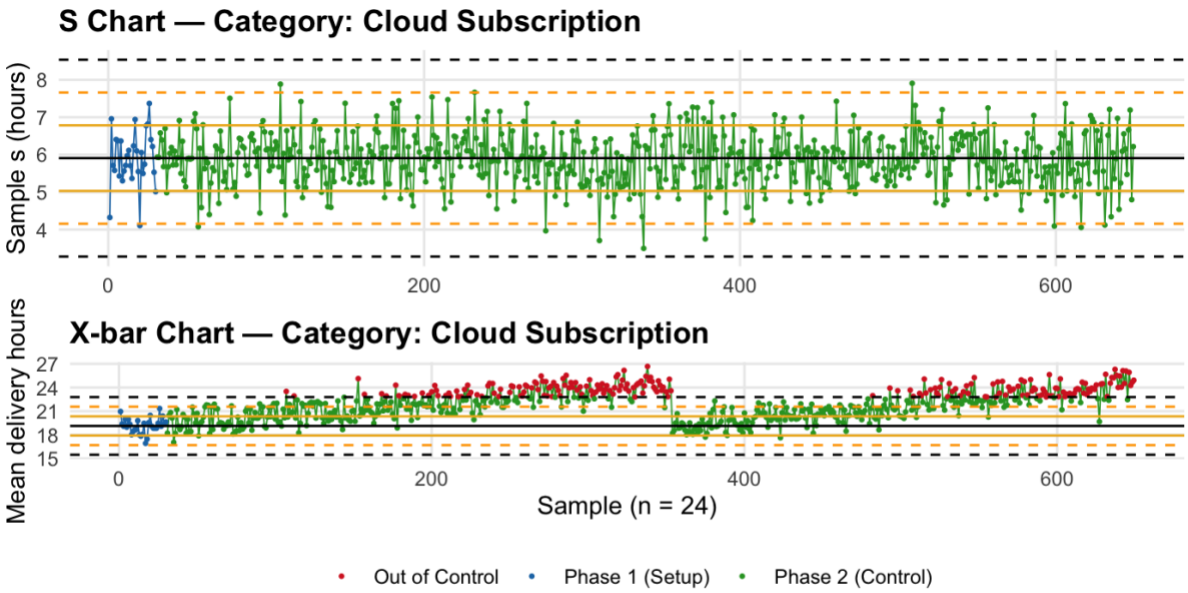
Measure	CL	L1	L2	LCL	U1	U2	UCL
$\bar{X}$	19.249	18.078	16.906	15.735	20.420	21.591	22.762
s	5.676	4.834	3.992	3.150	6.518	7.360	8.202

Figure 28 - Graph A3



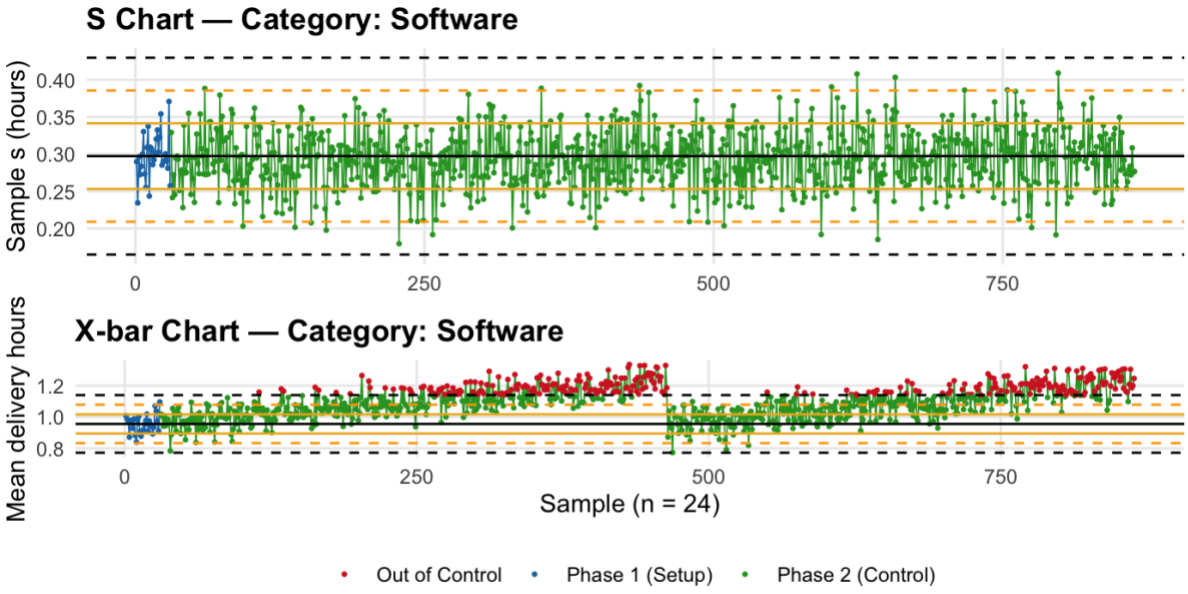
Measure	CL	L1	L2	LCL	U1	U2	UCL
$\bar{X}$	19.194	17.985	16.777	15.568	20.403	21.611	22.820
s	5.857	4.989	4.120	3.251	6.726	7.595	8.464

Figure 29 - Graph A4



Measure	CL	L1	L2	LCL	U1	U2	UCL
$\bar{X}$	19.126	17.907	16.688	15.469	20.345	21.564	22.783
s	5.908	5.031	4.155	3.279	6.784	7.660	8.537

Figure 30 - Graph A5



Measure	CL	L1	L2	LCL	U1	U2	UCL
$\bar{X}$	0.956	0.894	0.833	0.772	1.017	1.078	1.14
s	0.297	0.253	0.209	0.165	0.341	0.386	0.43

Figure 31 - Graph A6