ECSA FINAL REPORT

QA 344

Wilmien Van der Merwe

26912406

Van der Merwe, W, Miss [26912406@sun.ac.za]

Contents

Introduction:

This project applies statistical quality control and analytical optimisation techniques to evaluate and improve process performance across different datasets. The aim is to analyse variation and assess stability to be able to identify opportunities for improving both the quality and profitability of these processes.

Part 1: Descriptive Statistics

For part 1.2, three datasets were provided for analysis using descriptive statistics: customer_data, products_data and sales2022and2023. Each dataset represents a different aspect of the business and provides insights that can be used to strengthen the overall business strategy.

Customers – dataset:

The customer dataset provides demographic information about the company's customers, including gender, age, income, and city. It consists of 5000 rows (unique customers) and five columns, namely:

CustomerID: Unique identifier for each customer.

Gender: Male or Female

Age: Customer age in years

Income: Customer's annual income ($)

City: Customer's location

The average customer age is 51.5 years while the median age is 51 years. This indicates that the age distribution is nearly symmetrical with very few extreme outliers. Customer ages range from 16 to 105 years, which is a broad spread, but most customers fall around the early 50's range.

The gender distribution of customers is relatively balanced, with 2350 male customers and 2432 female customers. This results in a 0.97:1 male-to-female ratio.

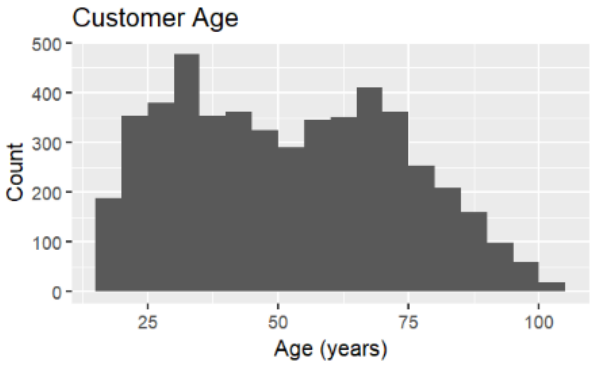
The mean income in the dataset is $80797 while the median income is $85000. Income data is typically right-skewed because there are always a few high earners; the median gives a better representation of the typical customer income. The difference in the median and mean indicates that some customers have much higher incomes, which pulls up the average. The standard deviation of customer income is $33150.11

Customers are located across seven major cities: New York, Houston, Chicago, San Fransico, Miami, Los Angeles and Seattle. This broad geographic distribution suggests a diverse customer base across key metropolitan areas.

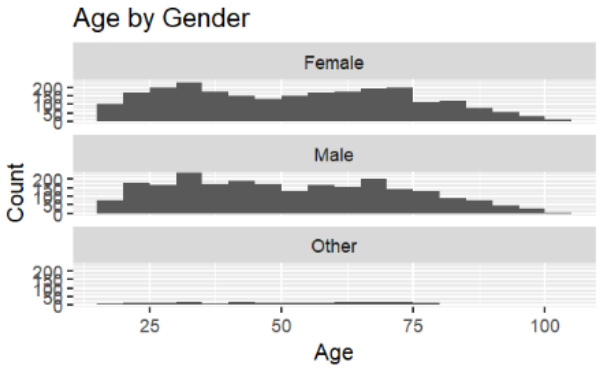
The number of customers per cities is given in the table below:

City	Number of customers
San Francisco	780
Los Angeles	726
New York	726
Chicago	724
Houston	724
Seattle	673
Miami	647

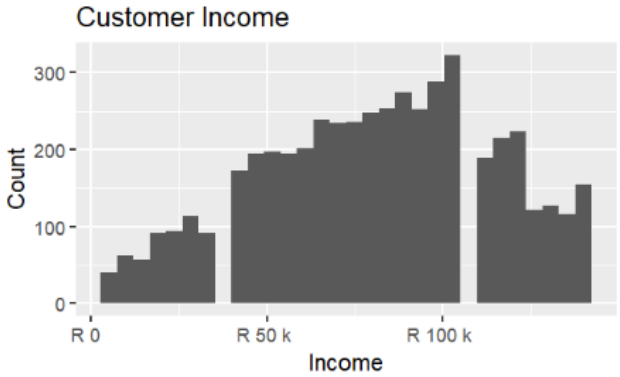
This table shows that the number of customers is evenly spread across Los Angeles, New York, Chicago and Houston. San Francisco has the most customers while Seattle and Miami are the only cities with less than 700 customers.



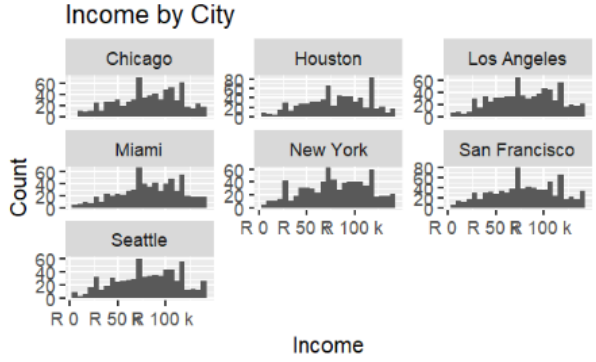
This histogram shows that most customers are between 25 and 75 years old with a concentration between 30 and 60 years old. The frequency gradually decreases for the older groups while there is a noticeable jump at around 18, which can be explained by younger people starting work / university and buying electronis they did not need in school.



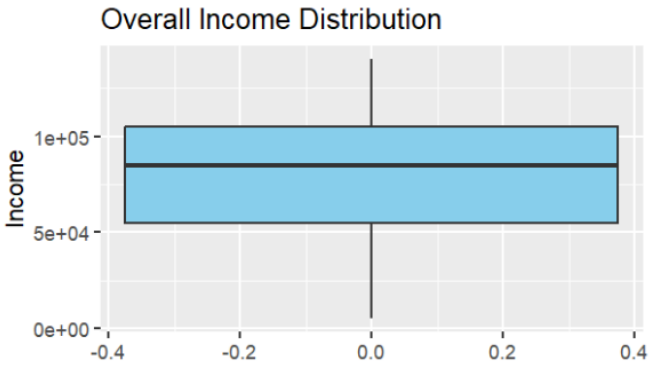
The customer age distributions split by customer age is almost identical. Both female and male customers are spread somewhat evenly across all age groups with a slight concentration at around 30 and 65 years old. Female customers have a higher representation across all age groups while the male customers have more noticable peaks. The customer base is well balanced by gender.



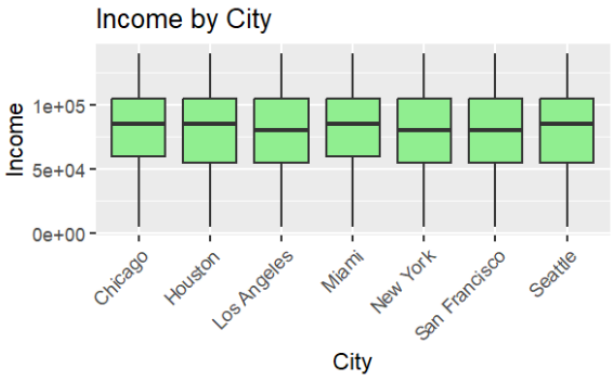
The customer income histogram shows that most customers earn between $50000 and $110000 annually. The shape of the graph is right-skewed, meaning there are a few customers with high annual incomes compared to the majority of customers. There is also a small percentage of customers that earns less than $25000. For the income distribution the median income gives is a more realistic picture of the typical customer than the mean because of the few very high earners.



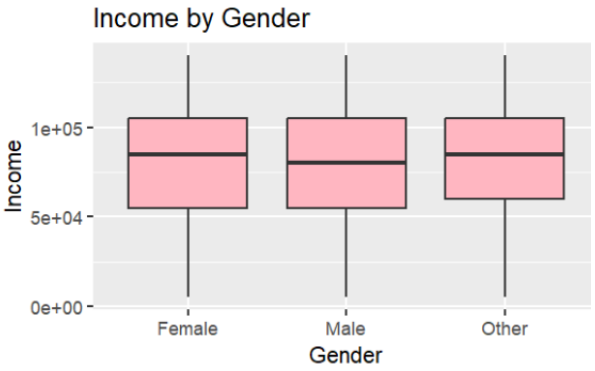
The histograms showing the income per city reveals that customers in San Francisco, Los Angeles and New York have higher incomes than the other cities. They are also the three cities with the most customers.



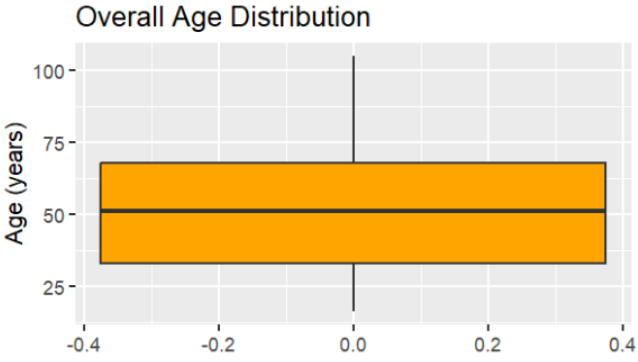
Overall income distribution shows the median is more towards the upper part of the box which means that most people have higher incomes but there are a few vir lower incomes that pulls the tail to the left. Therefore the distribution seems left skewed. Because the histogram and boxplot does not give the same distribution (right skewed vs slightly left skewed) it means there are gaps in the data and the data has multiple peaks, this can be seen on the histogram.



The incomes by city shows slight variation but all city incomes are between $50000 and $120000.



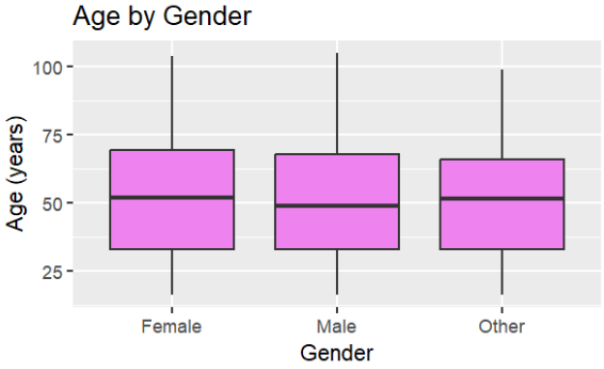
The median income is similar for male and female customers.



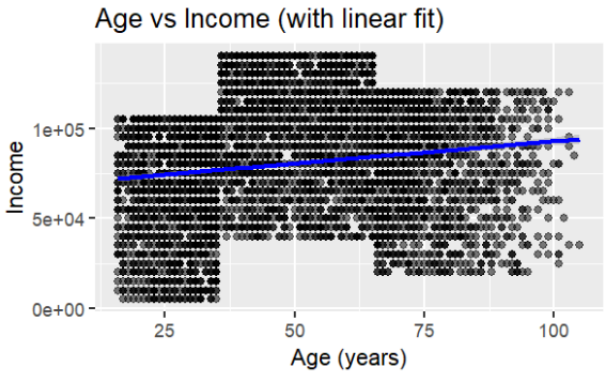
This boxplot shows the overall median which is around 51 years. The full range is between 16 and 105.

These boxplots based on income show no strong gender or city bias in either age or income.

The median age is similar across all cities.



The median age is similar for both males and females.



This scatterplot illustrates the relationship between customer age and customer income. The linear regression line indicates a positive correlation, as age increases, so does income. This pattern is logical as people who are older and have more work experience typically earn more money. Income among younger customers (16-35) varies the most, between $0 and $110000. The middle-aged customer group (36-68) also shows variation, but at a higher overall income level, typically between $40000 and $180000. In contrast, the older age group (69-105) contains visibly fewer observations, with incomes mostly between $35000 and $150000.

Products – dataset:

The products dataset gives more information about the products sold by the company. It describes products by their categories, gives a description of each product and gives its selling price and markup. The dataset has 60 entries (unique products).

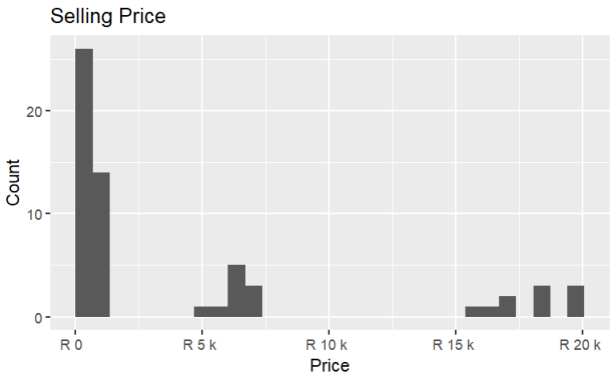
The products' SellingPrice's have an average of R4493.6 and a median of R794.2. The large difference in mean and median indicates that the dataset has many extreme values that influence the average, whereas the median is resistant to outliers. The mean is much larger than the median, which indicates that there are a few products with extremely high selling prices that make the average selling price higher, even though most products' selling prices are much less (closer to R794.2).

The average product's markup is 20.46%, and the median is 20.34% this is typical of a balanced, symmetrical distribution with no major outliers.

The table below gives a summary of the products' category, mean price, standard deviation of selling price and median price:

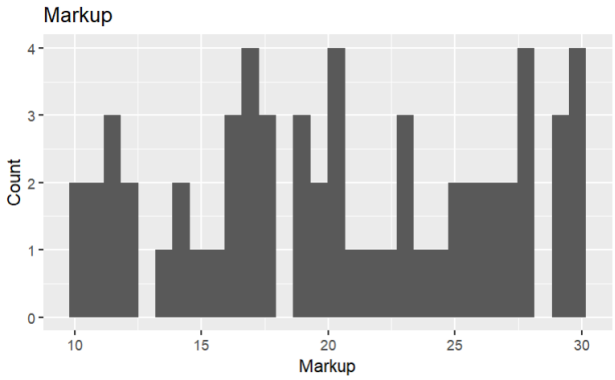
	Category	n_products	mean_price	sd_price	median_price	iqr_price	min_price	max_price	avg_markup	max_markup	cv_price
1	LAP	10	18086.429	1357.42780	18460.600	2321.6525	15851.74	19725.18	18.430	29.84	0.07505228
2	MON	10	6310.525	501.86771	6437.140	500.4475	5346.14	6806.08	23.868	29.74	0.07952868
3	CLO	10	1019.062	118.31961	1069.040	119.5225	728.26	1128.98	19.956	27.70	0.11610639
4	KEY	10	644.660	107.23236	645.040	154.7100	512.40	835.62	23.981	29.53	0.16633940
5	SOF	10	506.183	44.46783	513.840	40.6150	396.72	549.02	16.040	25.05	0.08784932
6	MOU	10	394.698	33.84428	384.945	54.4625	350.45	454.04	20.495	27.14	0.08574729

Histograms:

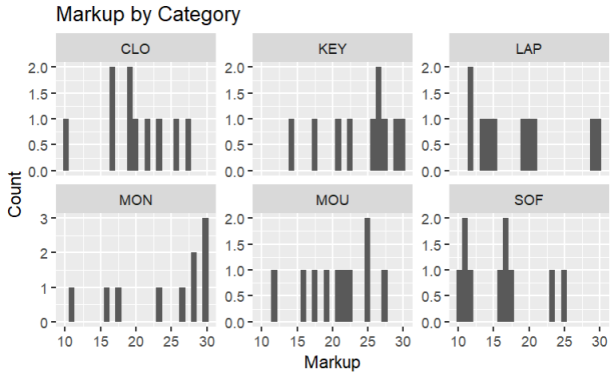


The selling price histogram indicates large price gaps in the product selling price. Most products have selling prices of less than R2000. There is a small peak between R4800 and R7500 and another between three small peaks between R15000 and R20000. The distribution is right skewed because most products have lower selling prices with only a few higher value products like laptops raising the average selling price.

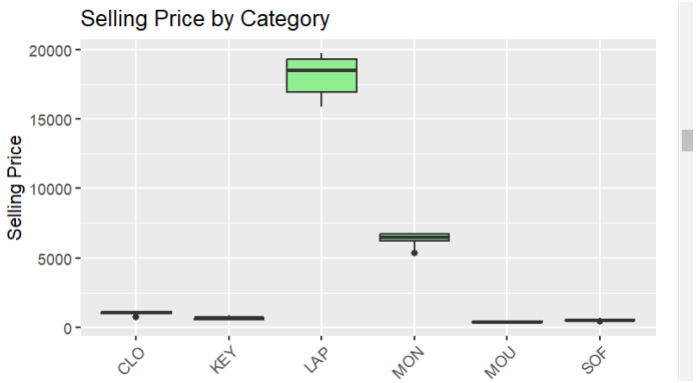
The selling price per category shows which product types are of higher value. Laptops have the highest selling prices, and Monitors have medium range selling prices. Cloud subscriptions, Keyboards, Mouses and Software have low selling prices. This supports the previous graph's interpretation that the overall selling price graph is right-skewed, because Laptops are the only high value product compared to the five other less expensive product types.



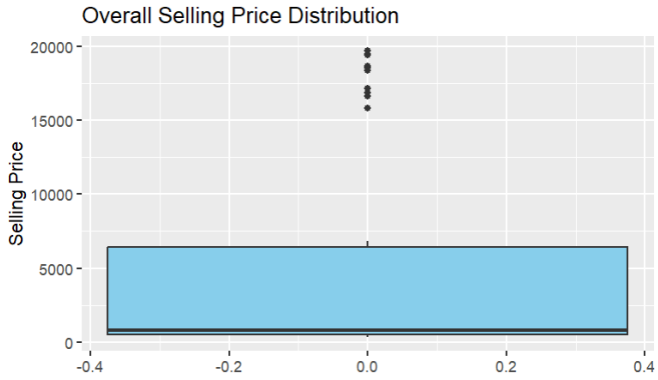
The product markups fall between 10% and 30% with some peaks at 12%, 17%, 21%, 23%, 28% and 30%.



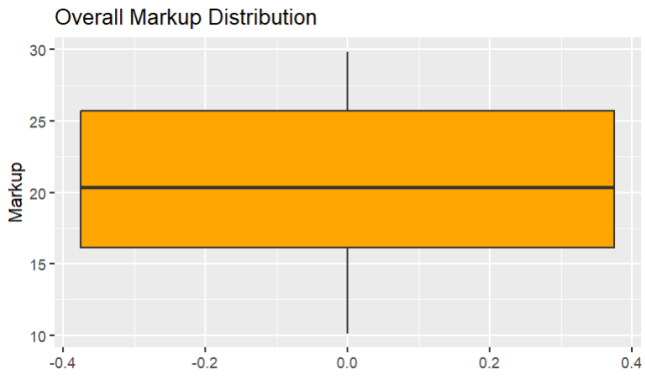
The markup per category histograms indicates that most monitors and keyboards have quite high markups, while the computer mice and cloud storage has mid-range markups and Software and Laptops have the lowest markups of the product categories.



Laptops are the higest priced category, with prices ranging from R17000-R20000. Monitors are the second highest category with prices between R5000 and R7000. Cloud subscriptions, Mice and Software all have low selling prices, generally below R1000.



The median selling price is low, most products are priced under R6000. The majority of selling prices are clustered at the lowered end with only some products close to the R20000 range.



This boxplot shows that most products have markups between

Sales – dataset:

The sales dataset contains information about which customers bought what products, the quantity bought, the order time, order day, order month, picking hours and delivery hours of sales during 2022 and 2023. These variables help evaluate customer demand patterns and process performance.

The average quantity bought by customers per order is approximately 13.5 units, with a relatively high standard deviation of 13.8, indicating that the order sizes vary widely. Some customers place single order items while others purchase in bulk.

Orders are placed throughout the day with an average order time around 13:00 and the standard deviation of order time is 5.5 hours, suggesting that most sales take place in the middle of the day. The mean order day is day 15.5 and mean order month is 6.4 – this indicates that the sales are fairly distributed across the months and years and does not have spikes in sales during specific times.

Picking hours, the hours needed to prepare products in the warehouse for dispatch after an order is placed, took on average 14.7 hours with relative variability of 10.4 hours. Furthermore the average delivery time is 17.5 hours with a standard deviation of 10 hours. This shows that even though the company's sales stay relatively stable each month of 2022 and 2023, the efficiency in the company's warehouse and delivery operations does vary.

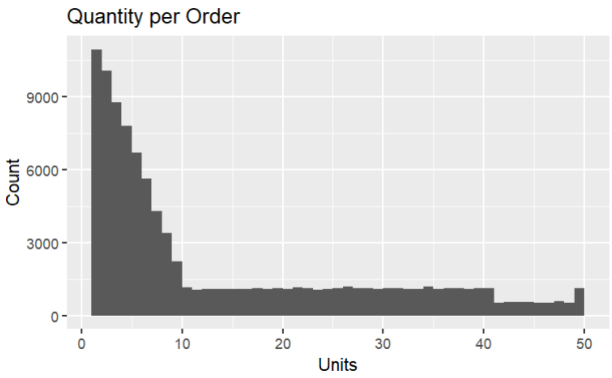
	avg_quantity	sd_quantity	avg_orderTime	sd_orderTime	mean_orderDay	mean_orderMon	avg_picking	sd_picking	avg_delivery	sd_delivery
1	13.50347	13.76013	12.9323	5.495127	15.49683	6.44813	14.69547	10.38733	17.47646	9.999944

The following table separates the 10 categorical variables over every product type:

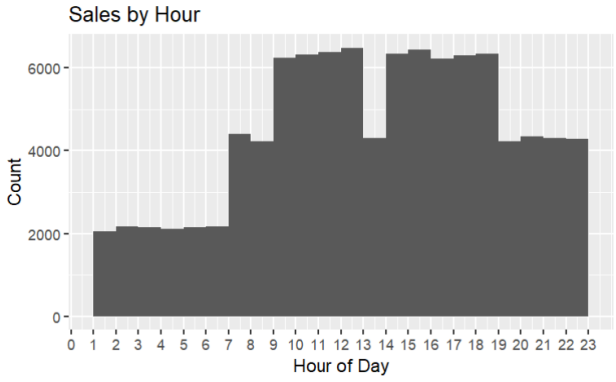
	orderYear	orderMonth	n_orders	avg_quantity	avg_picking	avg_delivery	YearMonth
1	2022	1	3305	13.89168	13.25196	15.43555	2022-01-01
2	2022	2	4816	12.98816	13.20893	15.45866	2022-02-01
3	2022	3	4744	13.59001	13.87758	16.06160	2022-03-01
4	2022	4	4831	13.97371	13.87869	16.57090	2022-04-01
5	2022	5	4819	13.43495	13.96651	16.87520	2022-05-01
6	2022	6	4665	13.12819	15.01092	17.52932	2022-06-01
7	2022	7	4762	13.27299	14.90669	17.90924	2022-07-01
8	2022	8	4755	13.05657	14.85524	18.03578	2022-08-01
9	2022	9	4738	13.89363	15.50752	18.51669	2022-09-01
10	2022	10	4710	13.46985	15.78980	18.93434	2022-10-01
11	2022	11	4734	13.42290	15.89860	19.33072	2022-11-01
12	2022	12	2848	13.20541	16.30882	19.84898	2022-12-01
13	2023	1	2829	14.05832	13.15555	15.18284	2023-01-01
14	2023	2	4096	13.24414	13.58427	15.75661	2023-02-01
15	2023	3	4107	14.04431	13.77374	16.02045	2023-03-01
16	2023	4	4128	13.52374	13.86460	16.53859	2023-04-01
17	2023	5	4077	13.22418	14.43478	16.91913	2023-05-01
18	2023	6	4059	13.68736	14.64929	17.31565	2023-06-01
19	2023	7	4137	13.78148	14.74864	17.64338	2023-07-01
20	2023	8	4045	13.22274	15.14628	18.13330	2023-08-01
21	2023	9	4083	13.32427	15.39753	18.44895	2023-09-01

22	2023	10	4078	13.71972	15.59522	18.66972	2023-10-01
23	2023	11	4178	13.64888	16.02486	19.21401	2023-11-01
24	2023	12	2456	13.57166	16.31599	19.51881	2023-12-01

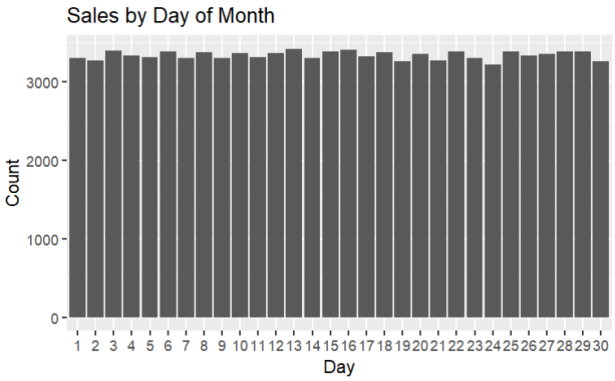
This table shows that the number of orders each month stays relatively constant except for January and December of each year. This could be because of holidays and people having more expenses than in normal months. The sales could be increased by having promotions in January and December. Although there are less orders in January and December the quantity of the orders that happens in these months stays the same as months in the middle of the year. The average picking hours varies between 13 hours and 17 hours for the two years. The average delivery times also varies consistently between 15 hours and 20 hours, which means the company usually delivers in less than a day after dispatch and less than 2 days from the time the order was placed.



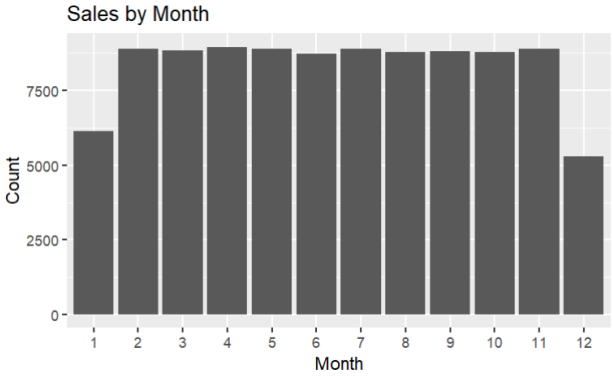
This histogram shows that there are many orders containing only 1 or 2 units. The number of orders drastically decreases from 1 unit per order to 10 units per order. Thereafter, there is almost the same number of orders that have between 10 and 42 units per order. Then it decreases again with fewer orders that have between 42 and 50 units per order. At 50 units per order, there is a small peak.



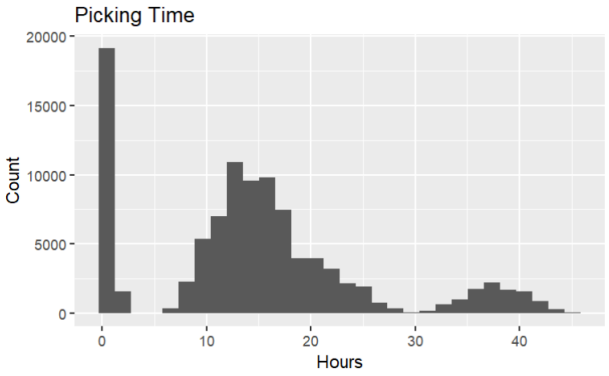
The sales by hour is the lowest between 12 AM and 7 AM, but by 8 AM it increases which reflects typical business hours. At around 9:30 AM, there is another increase, where peak sales are reached between 9 AM and 5 PM with a quick decrease between 1 PM and 2 PM, typically during lunch hours. After 6 PM sales start to decline with a sharp drop at 7 PM, but there are sales happening even after 10PM. The company experiences sales activity throughout the day but mostly during normal working hours.



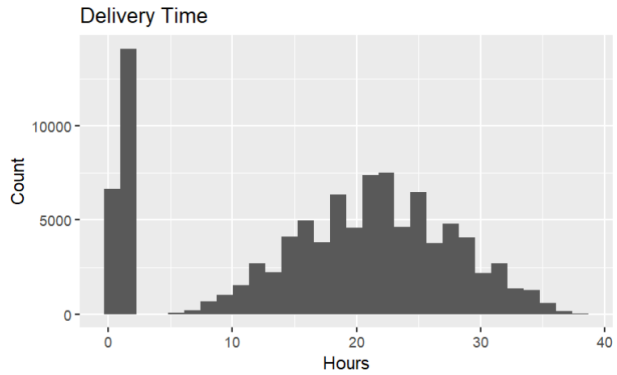
The sales stay consistent for each day of the month with very little variation.



The sales per month are consistent from February through November, with noticeably less sales during January and December.



Picking time has a significant peak at around 0 hours, which means the company's warehouse operations are very efficient, because the time it takes when an order is placed before it is dispatched is minimal. There is another peak between 10 and 20 hours, this is applicable to products that needs to be assembled in the warehouse before it goes out for delivery. Another small concentration forms around 35 hours to 40 hours for products where components might need to be made/ordered before it can be assembled.



Most deliveries happen within 5 hours, proving that the company's delivery system works efficiently and is reliable. Between 10 hours and 40 hours, a normal distribution is formed.

Part 3: Statistical Process Control

3.1 Initialising X-charts and s-charts:

The sales data for 2026 and 2027 were used to set up X-charts and s-charts for every product type. Firstly, the data was ordered as it would've arrived in real time. The data was ordered by Year, Month, Day and Ordertime. Samples of 24 were then created for each process. The first 30 samples of 24 each are used to determine the centre lines, outer control limits, the 2-sigma control limits and the 1-sigma control limits for the charts. For product types with fewer than 30 samples, provisional control limits were computed using all available subgroups and updated when more data became available. From these samples, the following control values were calculated for each product type:

Type: **CLO** | subgroups of 24: 645
Phase-1 limits (30 subgroups):
X-bar CL:19.1259
X-bar ±1σ :18.0036, 20.2483
X-bar ±2σ :16.8812, 21.3707
X-bar 3σ :15.7588, 22.4931
S-CL:5.9108
S-3σ:3.2824, 8.5391

Type: **KEY** | subgroups of 24: 740
Phase-1 limits (30 subgroups):
X-bar CL:19.194
X-bar ±1σ :18.0533, 20.3347
X-bar ±2σ :16.9126, 21.4754
X-bar 3σ :15.7718, 22.6162
S CL:5.858
S 3σ :3.2531, 8.4629

Type: **LAP** | subgroups of 24: 420
Phase-1 limits (30 subgroups):
X-bar CL: 19.5239
X-bar ±1σ :18.3788, 20.6689
X-bar ±2σ :17.2337, 21.814
X-bar 3σ :16.0886, 22.9591
S CL:5.906
S 3σ :3.2798, 8.5322

Type: **MON** | subgroups of 24: 615
Phase-1 limits (30 subgroups):
X-bar CL:19.4259
X-bar ±1σ :18.3228, 20.5291
X-bar ±2σ :17.2196, 21.6323
X-bar 3σ :16.1165, 22.7354
S CL:5.9183
S 3σ :3.2866, 8.55

Type: **MOU** | subgroups of 24: 857
Phase-1 limits (30 subgroups):
X-bar CL:19.2391
X-bar ±1σ :18.1491, 20.3292
X-bar ±2σ :17.059, 21.4193
X-bar 3σ :15.969, 22.5093
S CL:5.6889
S 3σ :3.1592, 8.2187

Type: SOF | subgroups of 24: 860
Phase-1 limits (30 subgroups):
X-bar CL :0.9556
X-bar ±1σ :0.8987 , 1.0126
X-bar ±2σ :0.8417 , 1.0695
X-bar 3σ :0.7848 , 1.1265
S CL :0.2971
S 3σ :0.165 , 0.4292

3.2 Drawing samples of 24 for each product type

After establishing these limits using the first 30 samples, additional subgroups were analysed to check whether delivery times remained within the established control boundaries. For each product type, both the X-bar and S charts were monitored, and all points beyond ±3σ were recorded.

The control charts indicate when the variation exceeds the expected natural limits. This should prompt the product managers to investigate possible causes of instability.

If no points are outside the ±3σ limit, then the process is operating in statistical control.

3.3 Process Capability Indices for product delivery times

Process capability indices (Cp, Cpu, Cpl, Cpk) were calculated for the delivery time process of each product type by using the first 1000 entries. The lower specification limit (LSL) was set to 0 hours, while the upper specification limit was set to 32 hours (USL). The benchmark for a capable process was set to Cpk >= 1.33.

Product types with Cpk >= 1.33 perform within customer specifications and if Cpk <= 1.33 then the processes require improvements to the variability and mean.

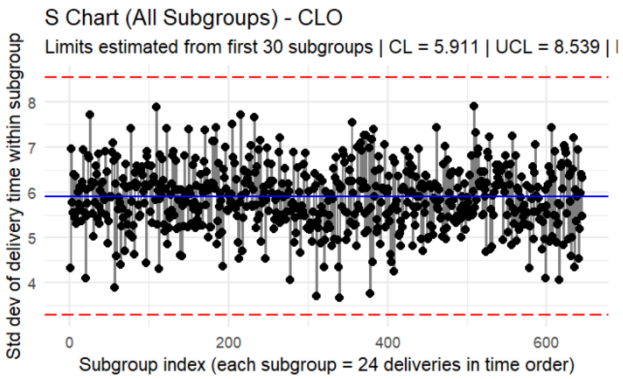
3.4 Process Control Issues

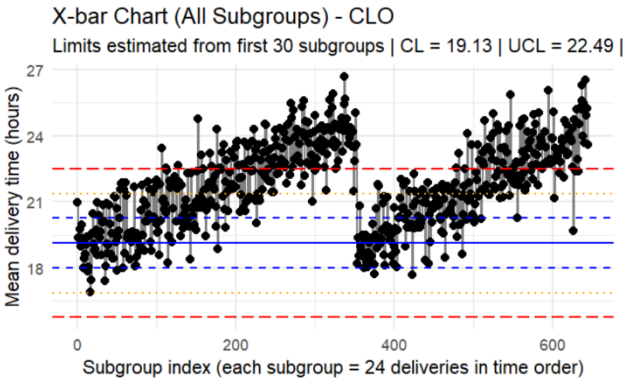
The following three rules were applied to monitor the delivery time performance of each product type. Each subgroup represented 24 delivery time observations, and the X-bar chart and S charts tracked the average delivery time and its variability over time.

Rule A:

Any single S-chart sample beyond the +3σ limit indicates an excessive increase in process variation.

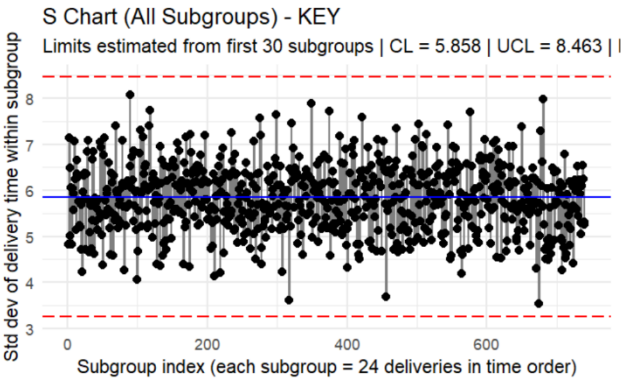
Cloud Subscriptions:

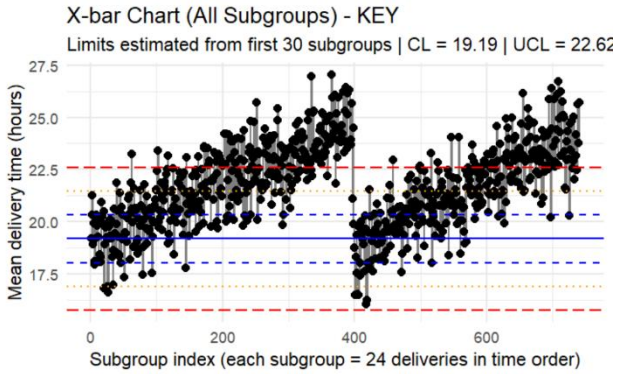




The S-chart for Cloud subscriptions remain steady within the 3σ limits, showing consistent short-term variability. The X-bar chart reveals a clear upward drift in the subgroup means over time. From around subgroup 100 and onward, the average delivery time increases steadily and eventually exceeds the +2σ and +3σ limits, meaning the process mean has shifted significantly.

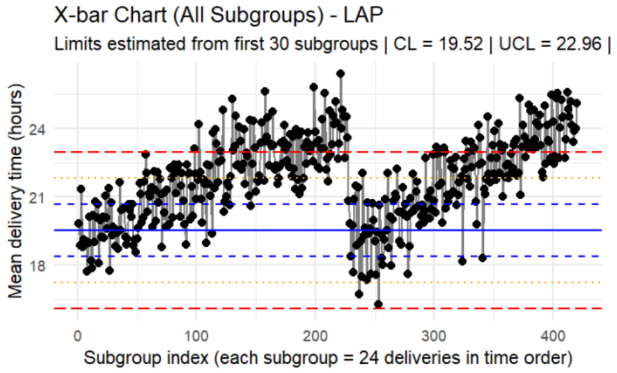
Keyboards:

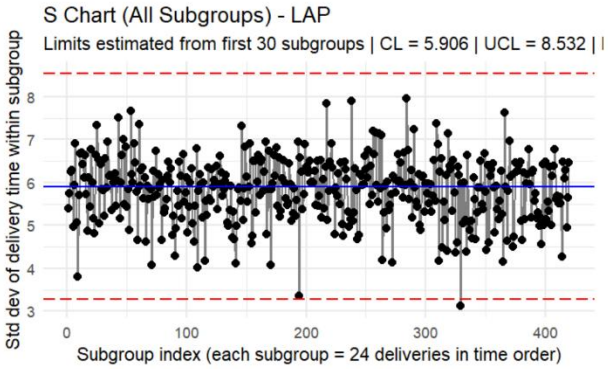




The S-chart for keyboards are stable, with the subgroup standard deviations fluctuating randomly around the centreline but staying within the control limits. In contrast, the X-bar chart shows a repeating pattern of upward drift. The mean delivery times gradually increase, then increase slightly and increase again. Several subgroups approached and exceeded the upper control limit. Day to day variation is consistent but the overall process average has shifted upward. This can be caused by delays in productio n or distribution scheduling.

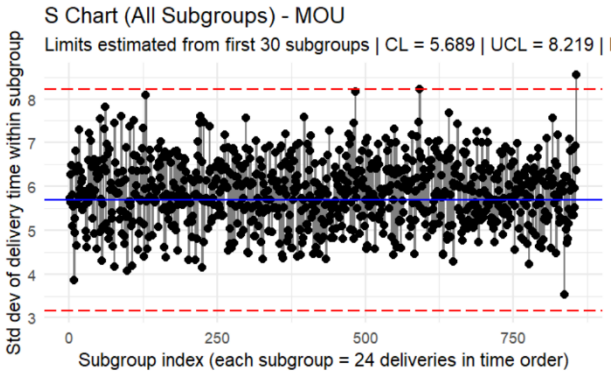
Laptops:

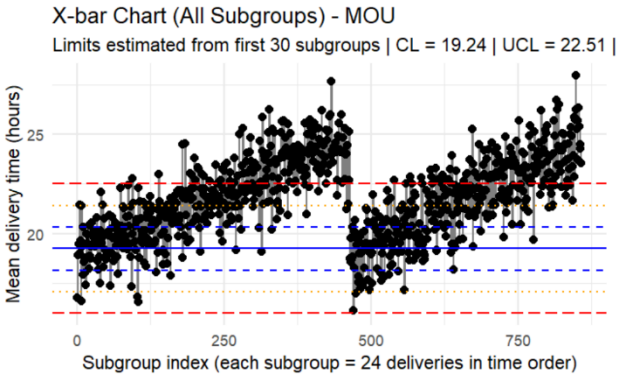




The S-chart shows stable variation with no significant points beyond the control limits. The -bar chart shows long term upward trends in the mean of delivery times of the subgroups. The process mean oscillates around the centre line at first but after about 100 subgroups the means rise and many samples above +2σ and near the +3σ line. Delivery times are gradually becoming longer on average even though internal variability remains within control.

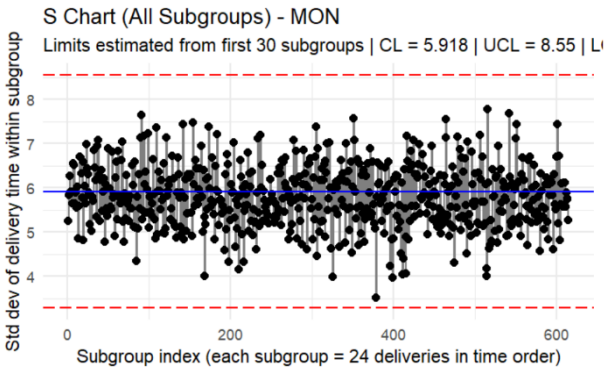
Mouse:

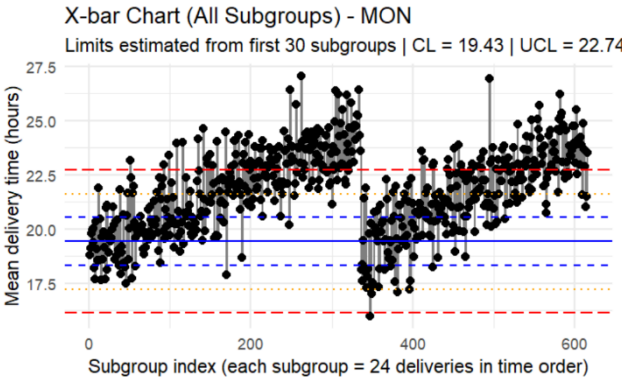




The S-charts indicates that the variation remains within the control limits and consistent. The X-bar chart shows a systematic trend in longer delivery times, with 2 samples above 27 hours. This can be caused by increased demand or the effects of backlogs.

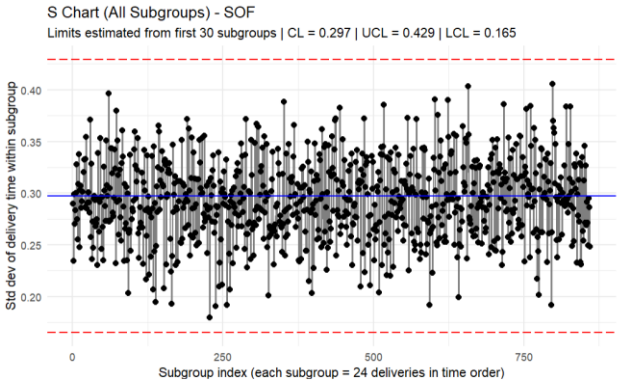
Monitors:

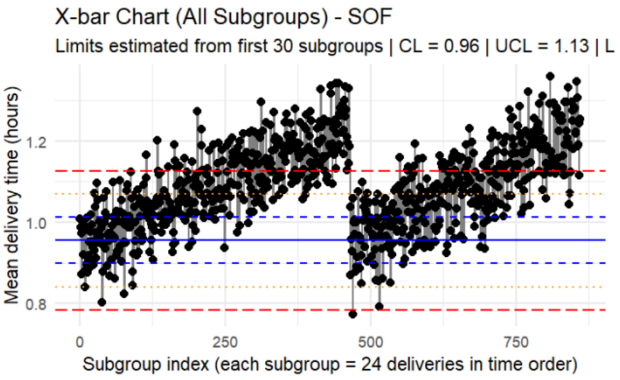




The S-chart remains tight and stable, with minimal fluctuation in the standard deviations of the subgroups. The X-bar chart shows that the mean delivery times increase gradually over time. It has more extreme delays than the other products, with some deliveries taking up to 27.5 hours. There is also a delivery that fell below the lower control limit, indicating the delivery was done much faster than expected. This upward trend can be because of monitors that are difficult to access in the warehouse or it needs to be assembled before dispatch.

Software:





The S-chart for software products are stable with no subgroup standard deviations reaching the control limits. The X-bar chart is less stable, with various subgroup means exceeding the upper control limits and 2 subgroups reaching the lower control limits. The mean of subgroup delivery times experiences a upward trend, but it is still much lower than the physical products. The increase in delivery time can suggest users struggle with the login process or download of the software products or that the user interface is difficult to use.

Rule B: Measures the longest run of consecutive S samples within +- 1σ, that is, how consistently stable the variation in delivery times is. Longer consecutive runs of S-values within ±1σ of the centre line indicate that there is minimal variation between the delivery times.

Cloud Subscription: The longest run of consecutive S samples within +- 1σ from the centre line is 19 samples. The variation in delivery times of cloud subscriptions is stable (usually downloaded online) because delivery happens digitally through activation or login.

Keyboards: The longest consecutive run of S samples within 1σ of the centre line for the keyboards is 16.

Laptops: The longest consecutive run of S samples within 1σ of the centre line for Laptops is 19 samples.

Monitors: The longest consecutive run of S samples within 1σ of the centre line for Monitors is 34 samples. This shows excellent statistical control.

Mouse: The longest consecutive run of S samples within 1σ of the centre line for computer mice is 14 samples – this is strong evidence of sustained stability in its delivery time variation.

Software: The longest consecutive run of S samples within 1σ of the centre line for ssoftware products is 19 samples.

Results of the application of Rule B to the S-charts of delivery times for all the product types shows that each product type has a long consecutive run of more than 14 samples that fall within the +- 1σ of the centre line. The short-term variability in delivery time is highly consistent, and the process operates under strong statistical control. The delivery process of all product types is stable and therefore predictable. This is a desirable outcome, however, it might also mean that it is being overcontrolled and unnecessary corrections are made to a process that is already stable, which could cost the company extra money.

Rule C: Checks to see if there are 4 consecutive subgroup means that all fall above the +2σ line, because this could indicate a gradual increase in the average delivery times.

Cloud subscription: 239 individual samples fall outside +2σ from the centre line, from which most are at least 4 consecutive samples.

Keyboards: 229 individual samples 2σ above the centre line and long consecutive runs.

Laptops: 120 individual samples 2σ above the centre line and long consecutive runs some of which are more than 18 consecutive samples.

Monitors: 192 individual samples 2σ above the centre line of which most are consecutive with short gaps.

Mouse: 273 individual samples 2σ above the centre line with a consecutive run of 57 samples. This is not normal and shows a gradual increase in the mean delivery time of computer mice.

Software: 270 samples more than 2σ from the centre line, with only a small number of individual samples that are not part of a consecutive run.

Across all product types, Rule C revealed consistent evidence of upward shifts in the average delivery times. The cloud subscription and software products, which are expected to have minimal to no delivery times, had 509 samples above the +2σ limit. This is likely data handling and measurement effects rather than real delivery delays. The physical product, Laptops, Mice, Keyboard and Monitors, displays long consecutive runs of subgroup means above +2σ. These

extended consecutive runs suggest a gradual increase in the average delivery time of products. The mouse deliveries are a particular concern, having a 57-sample consecutive run. This long consecutive run does not happen by chance and therefore confirms that the product delivery process is gradually starting to take longer than the usual average. This may require an investigation to identify the cause of the increase and prevent any further delays.

Conclusion:

The SPC results indicate that the delivery process is undergoing an upward trend, with all product types experiencing increased delivery times. According to Taguchi's principle, any deviation from the target is a loss to the company, even if the variation is within the control limits. According to this principal, the business is currently experiencing a loss because of this increase in delivery times and the variation experienced. The business must strive to minimize variation and maintaining the process mean near the centre to reduce any possible quality and profit losses. The increase in process mean could be a result of an increase in demand, understaffing, inefficient layouts in the distribution warehouse, or bottlenecks in the picking process.

Part 4: Control Chart Rules and Type II errors

4.1 Type I Error Probabilities

Rule A – One point beyond ±3σ:

We assume a standardized normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

$$P(\text{sample} > \mu + 3\sigma) = P(Z > 3)$$
$$P(Z > 3) = 1 - P(Z < 3) = 1 - 0.99865 = 0.00135$$

Therefore, there is only a 0.135% chance that any one in-control point will fall above the +3σ limit.

Rule B – Long consecutive run within ±1σ:

$$P(|\ Z\ | \leq 1) = 0.6827$$

The probability that *k* consecutive samples fall within ±1σ is:

$$(0.6827)^k$$

For $k = 15$:

$$(0.6827)^{15} = 0.003262 = 0.3262\%$$

Thus, there is only a 0.33% chance that 15 consecutive samples will all fall within ±1 σ of the centre line.

Rule C – Four consecutive points above +2σ:

$$P(Z > 2) = 1 - P(Z < 2) = 1 - 0.9772 = 0.0228$$

This means that any single in-control point has a 2.28% chance of being above +2σ.
The probability of four consecutive samples being above +2σ is:

$$(0.0228)^4 = 0.00000027 = 0.000027\%$$

Such an occurrence is *extremely rare* and would strongly indicate that the process is no longer in control.

4.2 Type II Error (β) – Missed Detection

Given:

$$\text{CL} = 25.050, \text{UCL} = 25.089, \text{LCL} = 25.011$$
$$\sigma_{\text{old}} = 0.013, \mu_{\text{new}} = 25.028, \sigma_{\text{new}} = 0.017$$

Compute the new z-scores:

$$Z_L = \frac{25.011 - 25.028}{0.017} = -1.00$$
$$Z_U = \frac{25.089 - 25.028}{0.017} = 3.5882$$

Now calculate β:

$$\beta = \Phi(Z_U) - \Phi(Z_L) = \Phi(3.5882) - \Phi(-1.00)$$
$$\beta = 0.9998 - 0.1587 = 0.8411 \approx 0.84$$

Interpretation:

With the mean shifted to 25.028 and the standard deviation increased to 0.017, about 84% of the sample means will still fall between the old control limits.
Therefore, the X̄-chart will likely fail to signal this change, meaning the process drift would go unnoticed. This weak detection occurs because the mean shift (from 25.011 to 25.028) is small, only one new standard deviation above the old LCL, and the UCL is still 3.59σ away from the new mean.

4.3 Data Cleaning

Before the analysis could be done, some issues in the products_Headoffice.csv and products_data.csv had to be fixed. The issues include:

- Fixing product ID's: Some products had incomplete or incorrect prefixes, so missing prefixes were filled in by using the last valid prefix above or below that entry.
- Product Type column: The first three letters of the ProductID column were used to create a new column that grouped the products by category.
- Repairing missing IDs: For rows that did not have a product code, a new productID was constructed based on the most recent valid entry to ensure that no rows started with "NA".
- Matching products_headoffice2025.csv and products_data2025.csv: Both datasets were checked to make sure the product types in both data sets corresponded.

Conclusion:

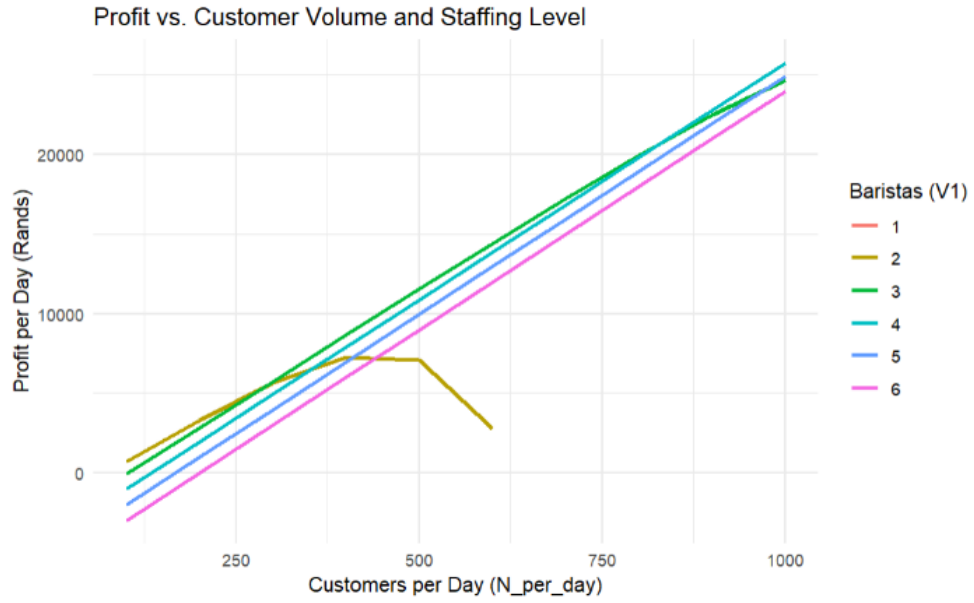
Even well controlled processes are subject to small probabilities of Type I and Type II errors. This part emphasized the importance of continual data monitoring. This section ensured that all datasets used were accurate and standardized to reinforce the validity of the data used.

Part 5: Optimizing the profits for 2 different coffee shops

Shop 1:

Assuming the coffee shop is open from 7 am to 5 pm daily and has a target service time of 4 minutes (240 seconds) per customer it means that the coffee shop can serve 150 customers per day. It is logical to think that as the number of baristas increases, reliability improves because customer wait less. However, adding more baristas also increase staffing costs which reduces profit beyond a certain point. Looking at the table, scheduling two baristas per day is optimal to maximise profits while keeping reliability at 89.82%. The maximum profit/day for an average of 150 customers is R2041.83. If the coffee shop experiences more customers, two baristas will no longer be enough to deliver reliable service. As seen in the figure below, at around 220 customers per day, the two scheduled baristas become overloaded, and the service reliability starts to decline. Beyond 500 customers per day, reliability drops drastically with only 2 baristas. At least 1 other barista should be scheduled for times when the average customers per day is expected to be more than 500 otherwise, profit will begin to decline due to slower service and long queues. At a low customer level, 2 baristas are optimal however, at higher customer volumes more baristas become necessary to maintain service reliability and keep customers happy. This highlights the importance of balancing cost efficiency with service quality. If the business just cared about profit, it would always only schedule 2 baristas, since it will produce the highest daily profit. However, this approach sacrifices service reliability and customer experience, in the long term this will harm the business' perception and profits. Adding a third barista during busy times will balance profitability with customer satisfaction.

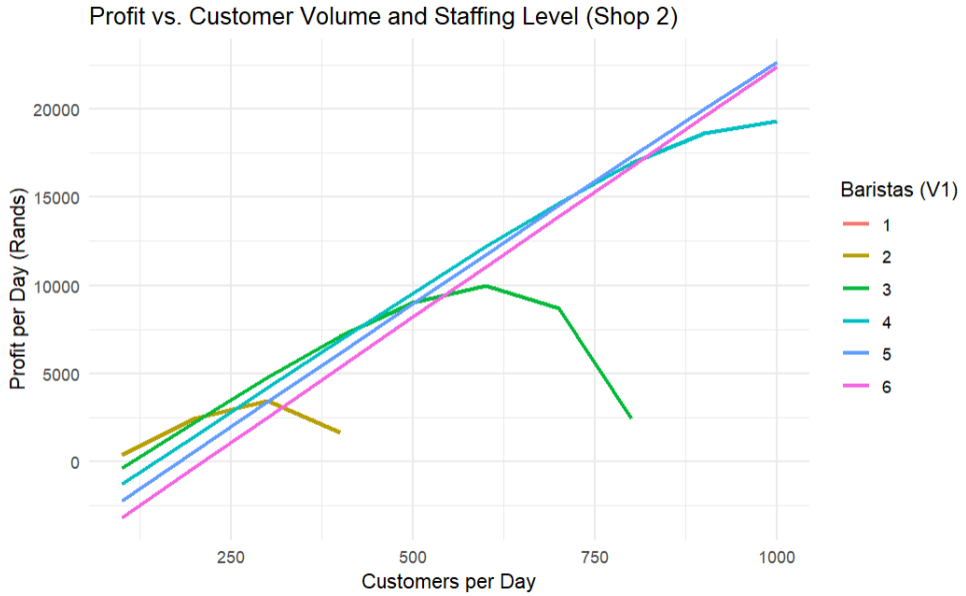
	V1 <int>	V2 <dbl>	mu <dbl>	stable <lgl>	R <dbl>	Profit <dbl>
2	2	100.17098	0.009982931	TRUE	0.8981846	2041.8305
3	3	66.61174	0.015012368	TRUE	0.9726930	1377.1183
4	4	49.98038	0.020007852	TRUE	0.9917854	463.0345
5	5	39.96183	0.025023876	TRUE	0.9975354	-511.0907
6	6	33.35565	0.029979932	TRUE	0.9992498	-1503.3759



Shop 2:

Under the same customer per day, operating hours and target service time assumptions, shop 2's optimal barista count per day is two. Two baristas will result in a 77.85% reliability and R1503.23 profit per day. Although shop 2 also needs two baristas for maximum profitability, it has a lower daily profit and service reliability. Shop 2's baristas start to become overloaded at around 190 customers per day, and service reliability drastically decreases at 310 customers per day. To maintain an acceptable service level, Shop 2 must schedule 3 baristas right before reaching 300 customers a day. If the shop gets busier and daily customers reach 620, a fourth barista should be scheduled. A fifth barista should be scheduled when there are around 875 daily customers. From the image below, shop 2 needs more baristas to keep the system stable, which increases its labour costs.

	V1 <dbl>	V2 <dbl>	mu <dbl>	stable <lgl>	R <dbl>	Profit <dbl>
2	2	141.51462	0.007066408	TRUE	0.7784965	1503.23417
3	3	115.44091	0.008662440	TRUE	0.8734700	930.61506
4	4	100.01527	0.009998473	TRUE	0.9092001	91.40062
5	5	89.43597	0.011181184	TRUE	0.9316739	-807.46737
6	6	81.64272	0.012248489	TRUE	0.9471151	-1737.98220



Conclusion

Both shops were analysed using the same operational assumptions: 9 working hours per day, 150 daily customers and a 4-minute target service level. Shop 1 had consistently faster service times across all staffing levels, which resulted in: a higher service rate per barista, a higher service reliability and a higher daily profit.

Shop 2 recorder longer average service times, which reduced its effective capacity. Even with the same number of customers per day, reliability fell below acceptable levels, arrival rate came close to or exceeded the maximum service capacity and overall profitability dropped as a result.

Shop 1 is performing efficiently with minimal staff, but management should focus on optimising processes in Shop 2; increasing the staff cannot be the optimal solution. If processes are optimized, revenue and customer service reliability will increase meaning increased long term profitability for the company.

Part 6:

6.1 DOE set up

In this experiment, we want to determine whether the mean delivery times differ significantly across different years and months. The factor under investigation is time (YEAR: 2026, 2027) and the variable is delivery time (in hours). The experiment was performed separately for each product type using 50 observations per year per product type. This represents a single-factor, multi-level DOE where each product type has repeated observations for each year. The analysis is performed using a one-way ANOVA.

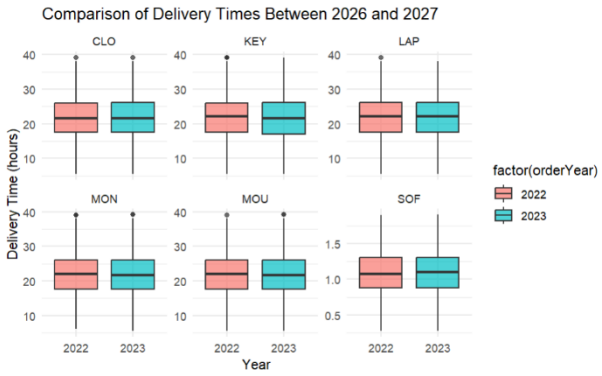
6.2 ANOVA results and analysis

A one-way ANOVA was applied to compare the average delivery times of each product type between 2026 and 2027. The null hypothesis stated that the mean delivery times were equal across years. The alternative hypothesis stated that the delivery times changed between 2026 and 2027 for at least one product type.

Product Type	F	p-value	Significant (α = 0.05)
CLO	1.55	0.2128798	No
KEY	6.01	0.0142655	No
LAP	0.55	0.4583654	No
MON	0.94	0.3316384	No
MOU	1.13	0.2878919	No
SOF	0	0.9794691	No

Conclusion

For all the product types, the p-values obtained from the ANOVA were greater than 0.05, indicating that there is no significant difference in mean delivery times between 2026 and 2027. This means that on average the delivery performance remained stable across years and small differences that occurred were likely random variation and not a real changing trend in the delivery process.



These boxplots compare the distribution of delivery times for each product type across the two years. In all 6 categories there was no significant change in mean delivery times from 2026 to 2027.

Part 7: Reliability and service

7.1 Number of days to expect reliable service:

Reliable service = at least 15 workers daily.

From the given graph:

- Days with 15 workers = 96
- Days with 16 workers = 270

That means the number of days with reliable service:

$$= 96 + 270 = 366.$$

Therefore, over the 397 days, 366 days had reliable service.

$$P(reliable) = \frac{366}{397} = 0.922$$

This means the shop is reliable about 92% of the time. The expected reliable days per year is then:

$$365\ x\ 0.922 = 336.53$$

So about 337 days per year of receives reliable service, and about 28 days per year will not have reliable service.

7.2 Optimising the profit for the company

Assumptions:

"Bad day": staff < 15

On a "bad day" the company will lose R20000

We can hire new people for R25000 per month per employee

7.2.1 Model reliability as binomial

Most days we have 16 staff members, some days drop to 15 and occasionally there are 14 or less staff members. So, if we plan for 16 staff members, most days we will have 16 staff members, but sickness / leave means there will only be 15 or 14 staff members. If we only plan for 15 we will have a lot more unreliable days because there will be days when only 14 or even 13 staff members will be at work. Higher planned staffing will reduce the probability of never having less than 15 staff members and reduce the R20 000 penalty days but it will also cost extra to hire more staff.

7.2.2 Cost of not being reliable

If 31 of 397 days had less than 15 staff members and therefore resulted in a penalty of R20000, that means: $31/397\ x\ 365\ =\ 28{,}5$ days will result in a penalty of R20000 per year.

If a R20000 penalty is paid on 29 days, then we spend $29\ x\ R20000\ =\ R58000$ per years on unreliability penalties. To decide if we hire more employees we need to calculate if the cost of staff will be less than the cost of not being reliable.

7.2.3 Staffing cost

Each additional employee costs R25000 per month, therefore if we hire 1 extra employee it will cost us $R25000\ x\ 12\ =\ R300000$ annually. This is less than what we are currently paying in penalties for being unreliable. Two additional employees will cost the us $R300000\ x\ 2\ =\ R600000$.

7.2.4 Comparisons

If we hire 1 additional employee, we will save $R580000 - R300000\ =\ R280000$ per year by eliminating most unreliable days.

If we hire 2 additional employees, the staffing cost will be more than the cost of being unreliable and we will lose $R600000 - R580000\ =\ R20000$.

Conclusion

To maximise the profit while maintaining an acceptable service reliability, the company should hire one additional staff member compared to the current staffing level. This will give the company a better financial outcome than not hiring another staff member because it avoids penalties and it is cheaper than aggressively overstaffing.