ESCA GRADUATE ATTRIBUTES PROJECT

Quality Assurance 344

D J Dillon
27093131

Contents

Introduction

This report analyses the data of a business selling various hardware and software products and aims to identify any quality issue present. The data will be cleaned and interpreted using an array of statistical methods to gain a better understanding of the trends that occurred. These results will be reported on in the context of service and reliability and aim to help the business identify inefficiencies and maintain consistent quality.

Data wrangling will be performed to remove invalid data entries. Thereafter, descriptive statistical methods will be applied in order to better understand the patterns in the data. The process capabilities will be calculated, and a statistical process control analysis will be carried out. The probabilities of tpe 1 and type 2 errors within the delivery times will be evaluated and the delivery time will therefore be optimised. A 2-Way ANOVA analysis will be carried out, and the reliability of the service and products will be evaluated.

1. Data Analysis

1.1 Data Wrangling

The data wrangling process followed included cleaning, merging and finally validating the data from the separate provided data sets (customers, products, Head Office products, and sales). A single Analytics Base Table was thus formed with each instance representing a single product transaction. Data validity was check through logic rules, including detecting negative or zero values, invalid days and dates and finally duplicate entries. This validity check revealed 6 invalid instances, all of which were duplicate rows. These instances were removed, resulting in a final clean data set containing 99 994 valid instances. These 6 duplicate instances could be due to everyday human errors, such as data entry duplication as a result of miscommunication between staff processing the same order. This could also be attributed to a system synchronization issue, or simply a miscommunication between departments. However, these 6 duplicates out of 100 000 instances are a small minority, and therefore not a major system-wide issue that warrants major investigation by the business.

1.2 Descriptive Statistics and Analysis

Order Value Distribution

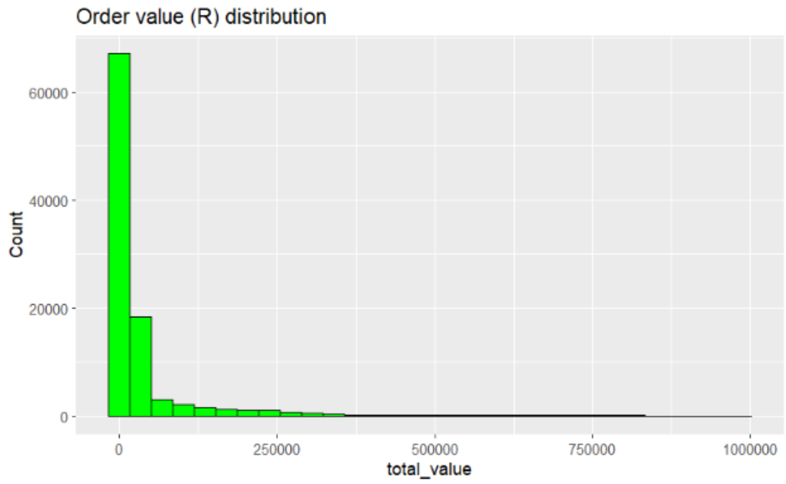


Figure 1: Histogram – Order Value

The price distribution follows an exponential distribution that is right skewed. A high number of the cheaper items were sold whereas a low number of the expensive items were sold. This long tail and high skewness suggest the presence of outliers much higher in value.

Order Value by Age Band

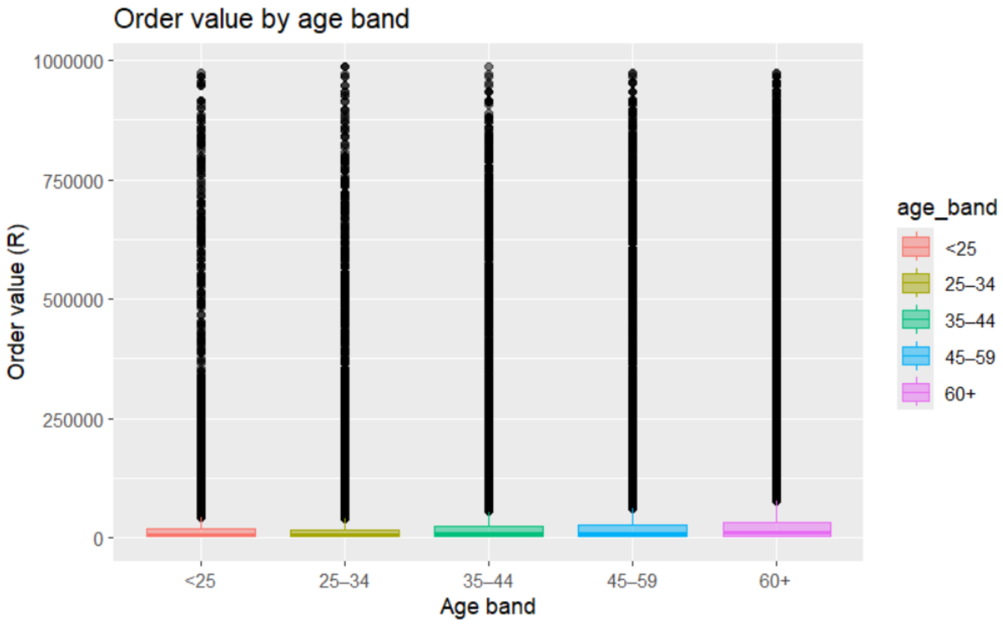


Figure 2: Box Plot - Order Value by Age Band

Order values can be seen to be highly right-skewed across all age bands, in correlation with the order value distribution. Typical spend differs very little by age, and each band shows less frequent abnormally large purchases that skew the data.

Revenue and Sales by Age Band

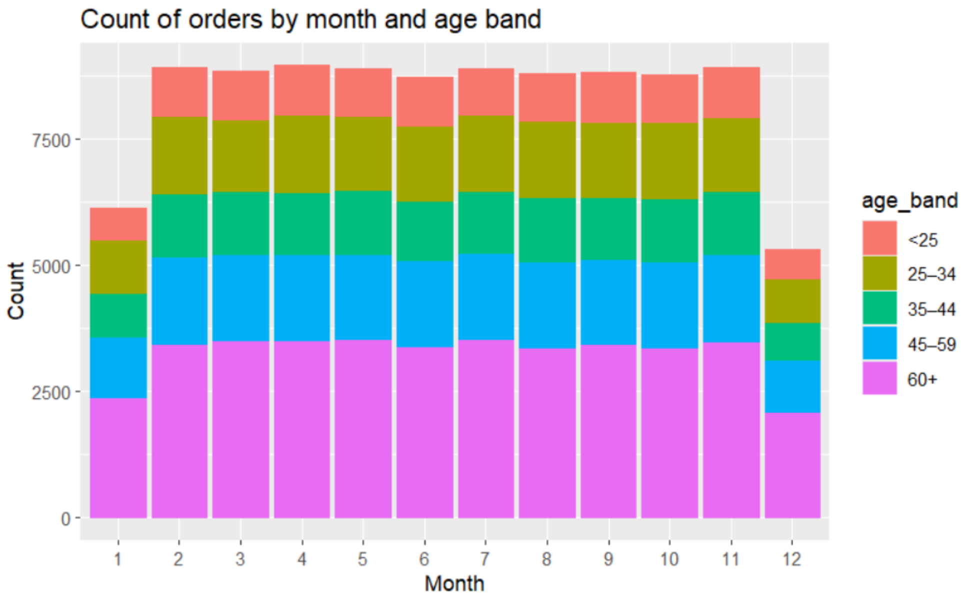


Figure 3: Stacked Bar Plot - Unit Selling Price per Month

It can be seen that the lowest sales months are January and December, with a higher consistent plateau from February to November. This suggests that the sales follow a typical dip at the start and end of the year, following the traditional December holidays. The 60+ band is responsible for the majority of sales throughout the year, while the <25 band is consistently the minority. The proportions of the age bands stay relatively stale throughout the year, meaning that the seasonality of sales is the same for all age groups roughly. A similar seasonality is revealed when monthly revenue is plotted per month by age band, instead of count of sales.

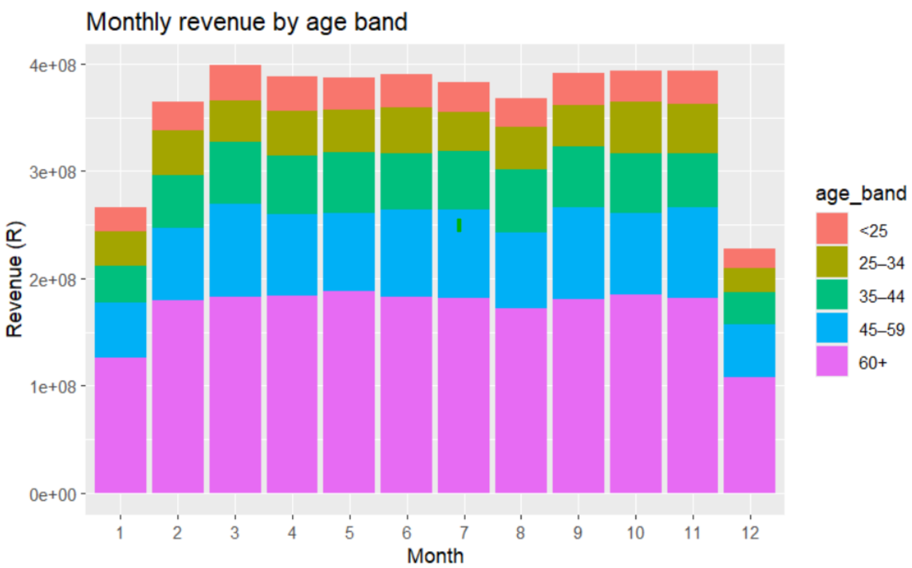


Figure 4: Stacked Bar Plot of Monthly Revenue per Month by Age

Delivery Hours by Month

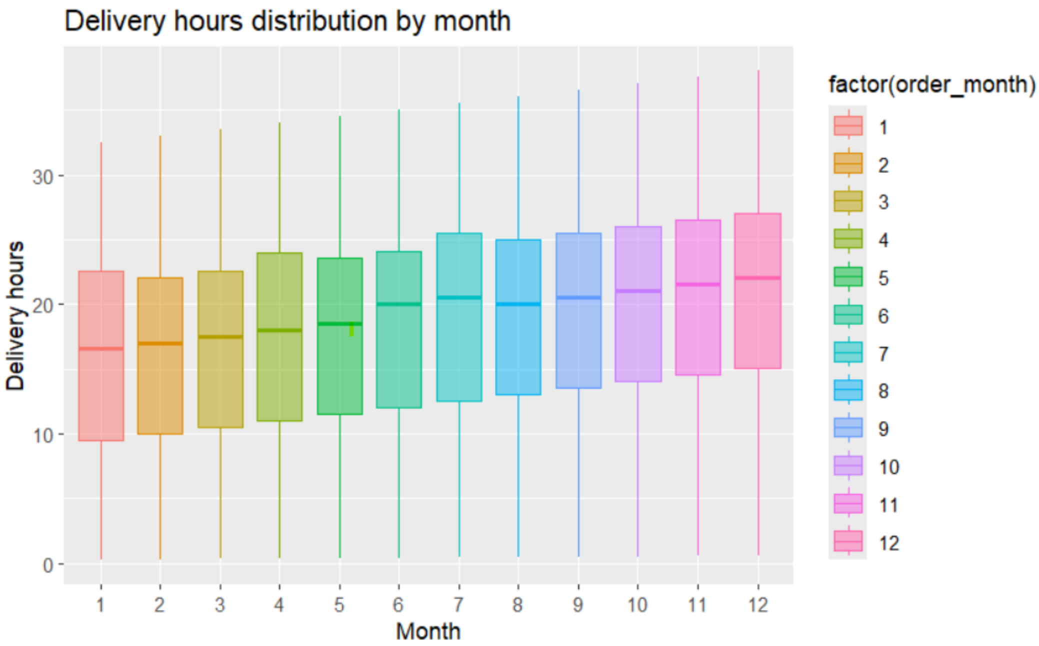
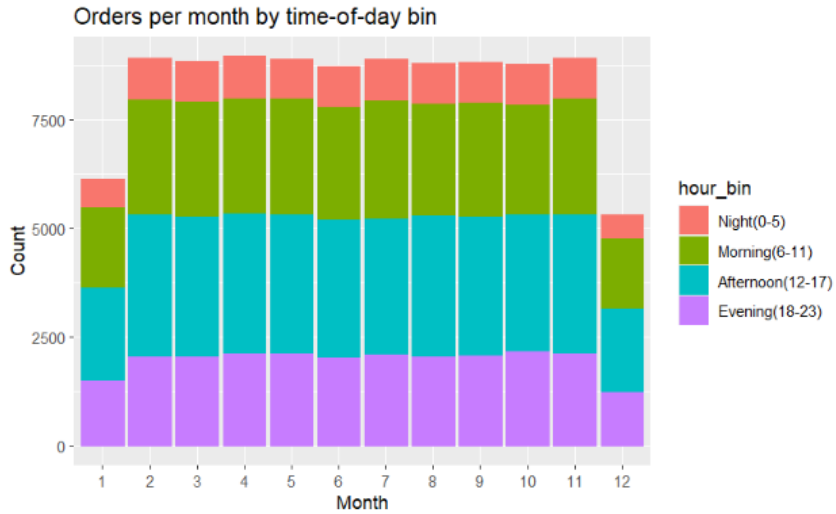


Figure 5: Box Plot – Delivery Hours by Month

The distribution of delivery hours drifts upwards throughout the year from January to December, from medians of roughly 17 hours in January to around 23 hours in December. Variability appears to increase from mid-year onward, as the whiskers of the box increase in length, meaning deliveries are less predictable in Q3-Q4. Months 10-12 (Q4) have the highest medians as well as the greatest variabilities. This therefore means that Q4 will have the highest risk of SLA misses and unhappy customers.

Orders per Month by Time-of-Day

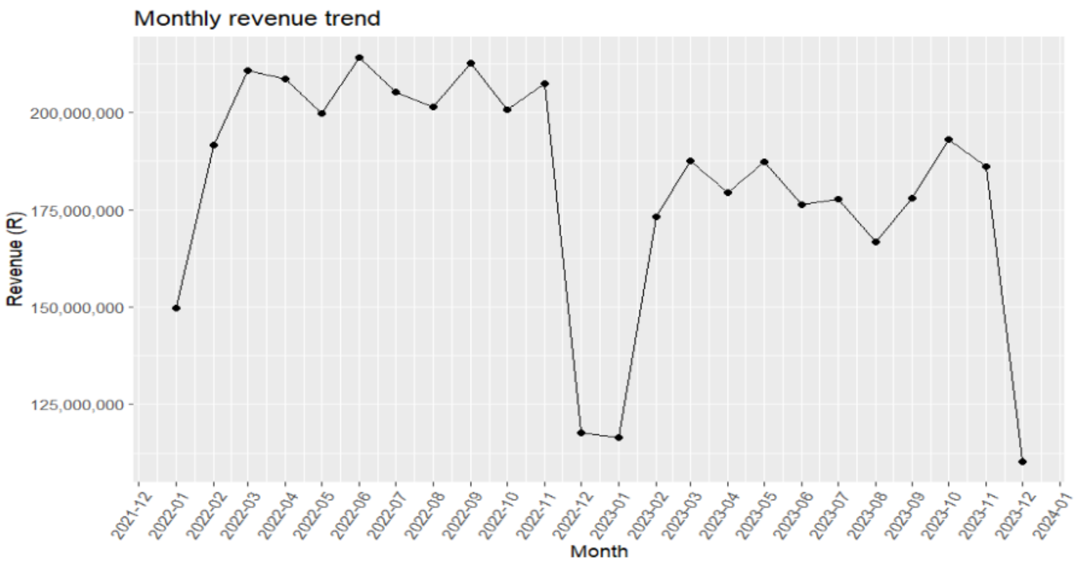
Figure 6: Stacked Bar Plot - Orders per month by time-of-day



The proportions of time-of-day count orders stays relatively stable across months, and staffing can therefore be planned accordingly with minimal risk. The mornings and afternoons dominate sales, and these times should therefore be staffed highest.

Monthly Revenue Trend

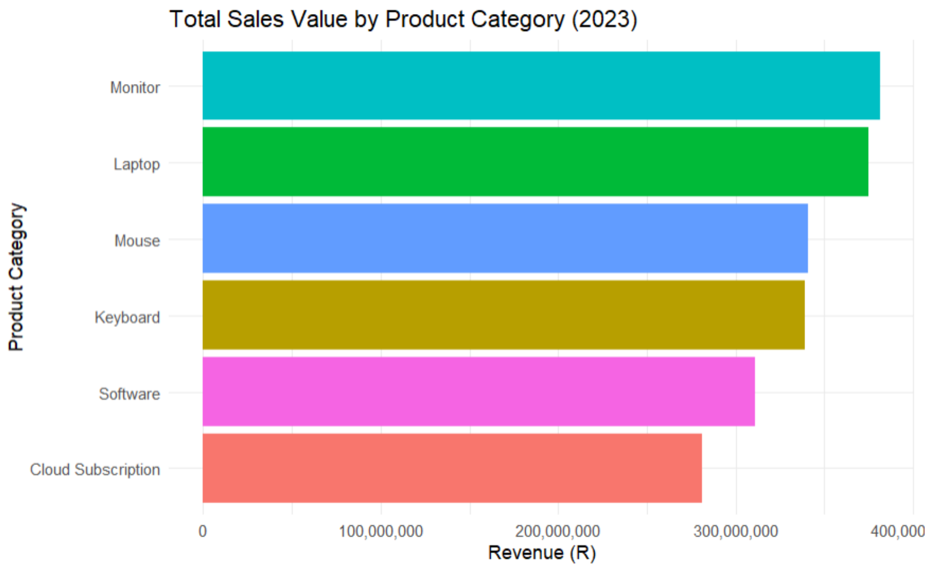
Figure 7: Line graph - Monthly Revenue Trend



Monthly revenue grew significantly into early 2022, where it stayed between R200-215M before dropping sharply in December 2022 until January 2023. This correlates with the seasonal drop in sales in January and December. Although monthly revenue climbed once again in February 2023, a significant decrease in revenue can be seen between 2022 and 2023.

Sales Value by Product Category

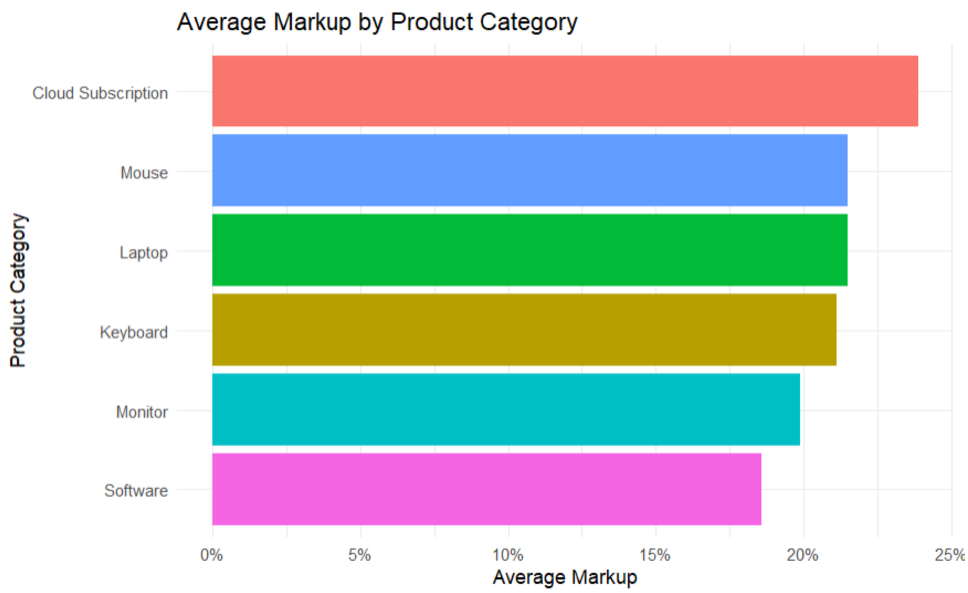
Figure 8: Total Sales Value by Product Category (2023)



The 2023 sales data can be seen to display a fairly balanced revenue distribution across product categories, with each product category contributing relatively evenly to overall revenue. Monitor and laptop sales do, however, lead overall revenue and cloud subscriptions and software sales are the lowest performers. Overall, the business can be seen to have a diversified and stable product mix, without overreliance on a single product category. This diversification could buffer the business against market changes, or production problems compromising the sales of a single product category.

Average Markup by Product Category

Figure 9: Average Markup by Product Category



The plot of average markup per product category reveals that although total revenue was fairly evenly distributed across product categories, profitability has more noticeable trends. Although cloud subscriptions were responsible for the lowest amount of revenue in 2023, they are the business's most profitable product, with an average markup of around 23%. This strong markup could be a factor contributing to software sales being the lowest revenue product category in 2023, and the business could therefore consider reducing this steep markup to appeal to a larger demand. Interestingly, monitors are the business's 2nd-lowest marked-up product, despite being the most profitable product category. This is likely due to a strategic pricing decision to maximise sales and capitalise on demand. Software finds itself as the least profitable product as well as the product category with the 2nd-lowest revenue in 2023. The business could therefore consider investing less into software products or even consider removing them from their product line to cut costs without affecting overall revenue and profitability greatly. These 2 plots together reveal that the company's product mix is balanced in sales volume and strategically positioned for profitability.

1.3 Analysis

Overall, this data analysis reveals that the business's sales performance is consistent and well diversified across both customer demographics and product categories. Order values are generally relatively low, with occasional high-value orders that are outliers in the data. This suggest that most transactions are routine, and a small percentage of sales that are higher in value are responsible for a large portion of total revenue. Delivery performance was revealed to have declined towards the end of both years analysed, likely due to potential capacity constraints in high-demand seasons. The seasonal and temporal patterns in the data illustrate that sales and customer activity remain relatively stable throughout the year, with demand peaking predictably during peak working hours. These insights provide a strong foundation for further analysis, and optimization of this business's operations through subsequent process control and hypothesis testing analysis.

3. Statistical Process Control

Control charts were plotted for each product. This allows for the out-of-control datapoints and process variation to be identified.

The data was ordered by year, month, day, then picking hours and formed into subgroups of 24 delivery hours. The first 30 subgroups were used to determine centre lines, outer control limits, the 2-sigma control limits and the 1-sigma-control limits for the X-Bar charts and S-Charts for each product. The charts generated are shown in the appendix and the results are discussed below.

3.1 X-Bar charts

Table 1: X-bar Charts

	UCL <dbl>	U2Sigma <dbl>	U1Sigma <dbl>	CL <dbl>	L1Sigma <dbl>	L2Sigma <dbl>	LCL <dbl>
SOF	1.137718	1.076585	1.015451	0.9543181	0.8931846	0.8320512	0.7709178
KEY	22.796531	21.594762	20.392992	19.1912222	17.9894525	16.7876828	15.5859131
CLO	22.793606	21.577533	20.361461	19.1453889	17.9293166	16.7132444	15.4971721
MOU	22.757601	21.583392	20.409182	19.2349722	18.0607626	16.8865529	15.7123432
MON	23.146348	21.916399	20.686449	19.4565000	18.2265506	16.9966012	15.7666518
LAP	23.168448	21.958678	20.748908	19.5391389	18.3293693	17.1195997	15.9098302

Figure 10: X-Charts SOF samples 31:end

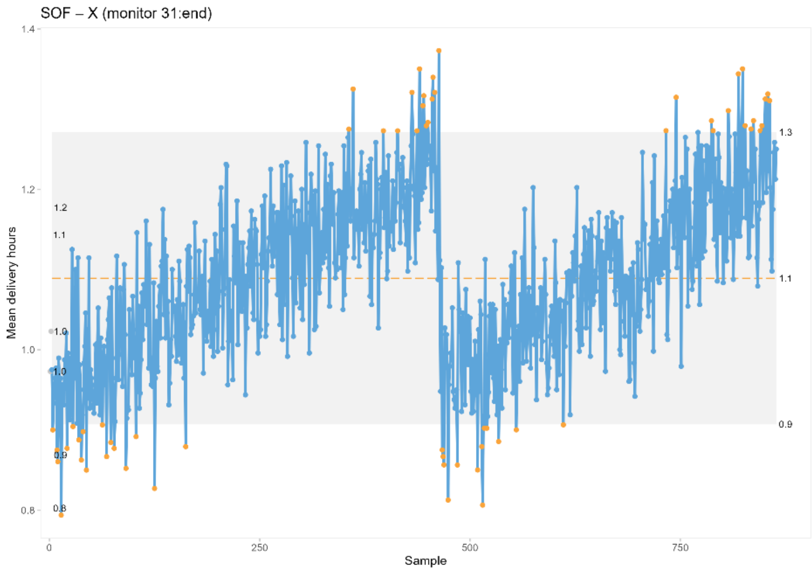


Figure 11: X-Charts KEY samples 31:end

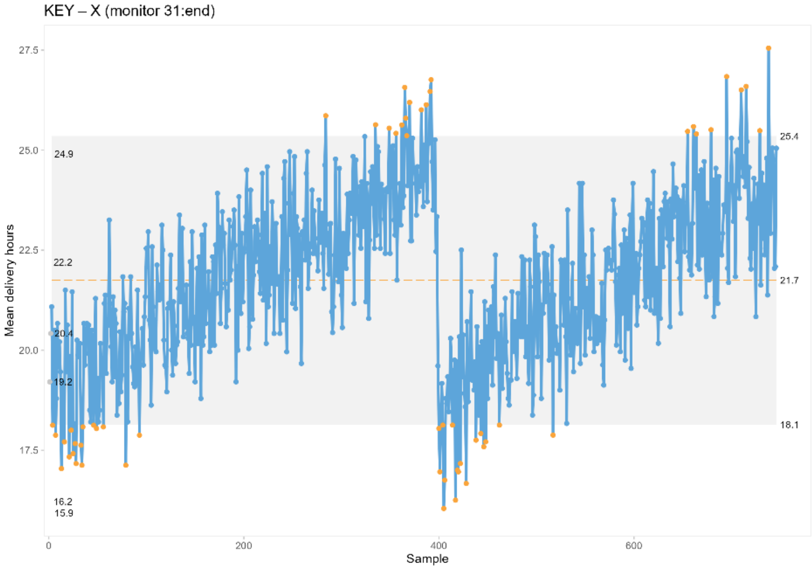


Figure 12: X-Charts CLO samples 31:end

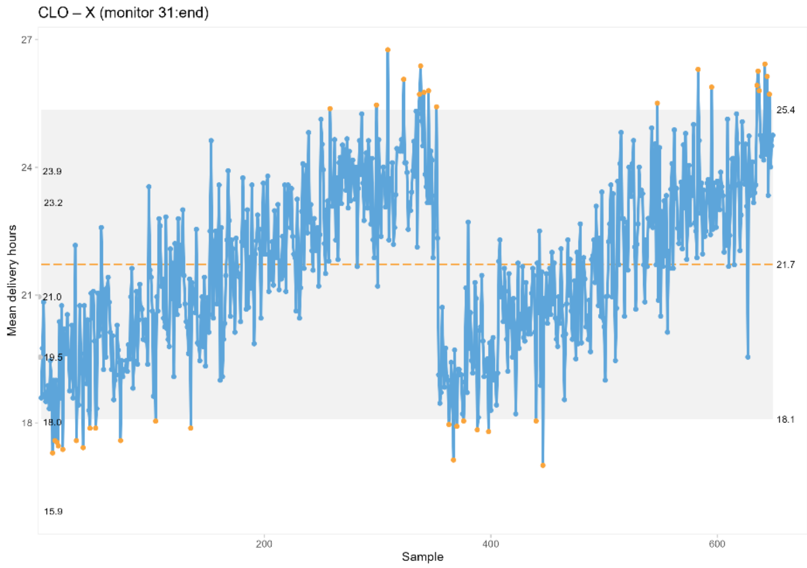


Figure 13: X-Charts MOU samples 31:end

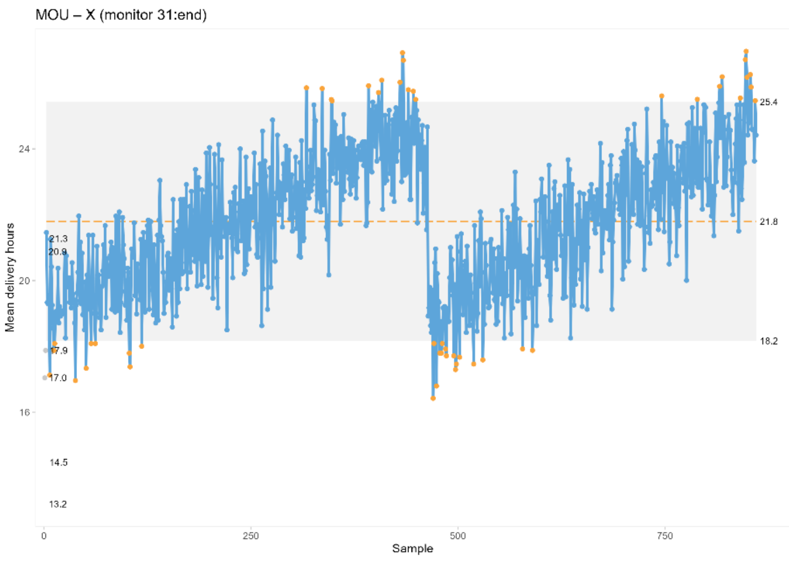


Figure 14: X-Charts MON samples 31:end

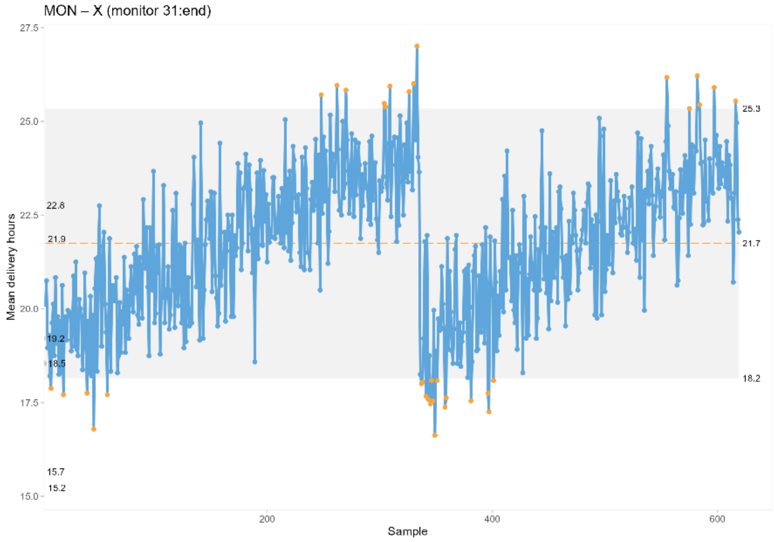
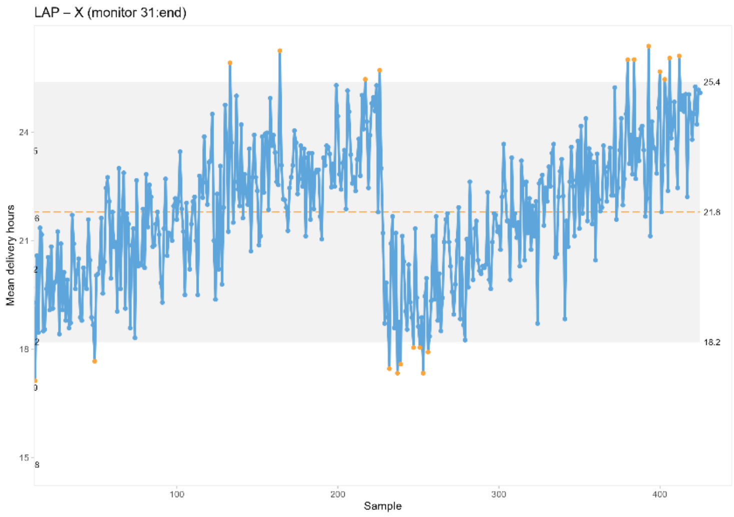


Figure 15: X-Charts LAP samples 31:end



3.2 S-Charts

Table 2: S-Charts

	UCL <dbl>	U2Sigma <dbl>	U1Sigma <dbl>	CL <dbl>	L1Sigma <dbl>	L2Sigma <dbl>	LCL <dbl>
SOF	0.4279903	0.3840785	0.3401666	0.2962547	0.2523429	0.208431	0.1645191
KEY	8.4134950	7.5502692	6.6870434	5.8238176	4.9605918	4.097366	3.2341402
CLO	8.5136261	7.6401268	6.7666276	5.8931283	5.0196291	4.146130	3.2726306
MOU	8.2205493	7.3771198	6.5336902	5.6902607	4.8468312	4.003402	3.1599721
MON	8.6107789	7.7273117	6.8438446	5.9603774	5.0769103	4.193443	3.3099760
LAP	8.4695014	7.6005293	6.7315573	5.8625852	4.9936132	4.124641	3.2556691

Figure 16: S-Charts SOF samples 31:end

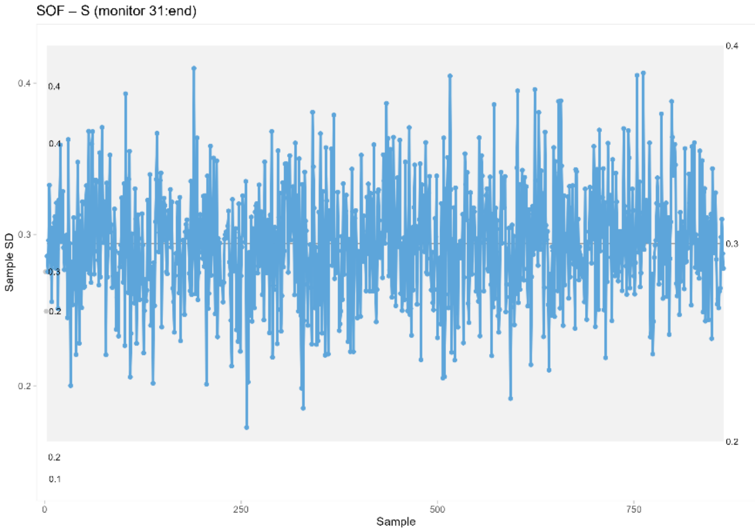


Figure 17: S-Charts KEY samples 31:end

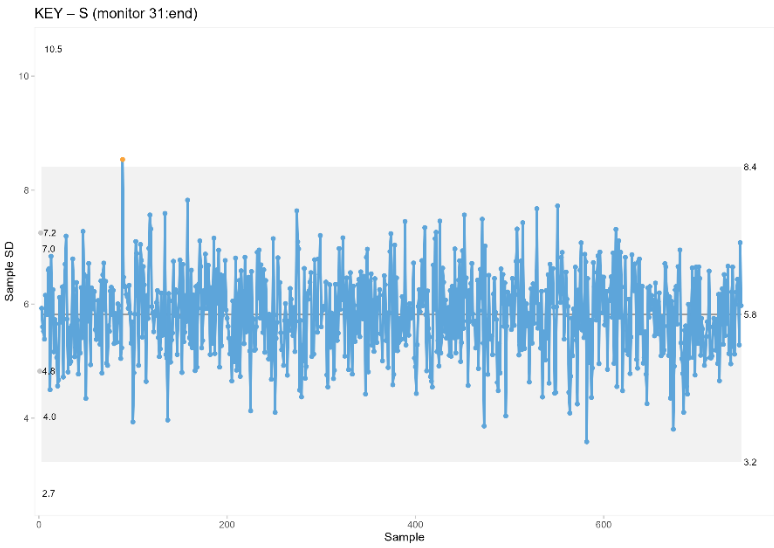


Figure 18: S-Charts CLO samples 31:end

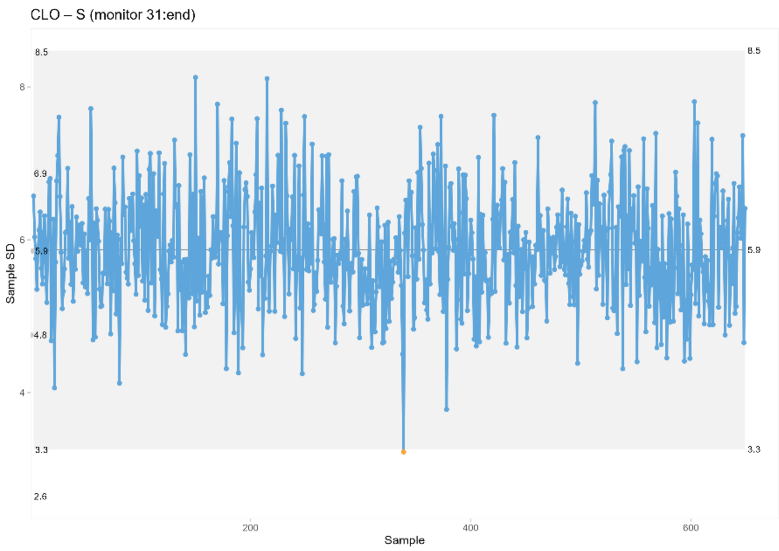


Figure 19: S-Charts MOU samples 31:end

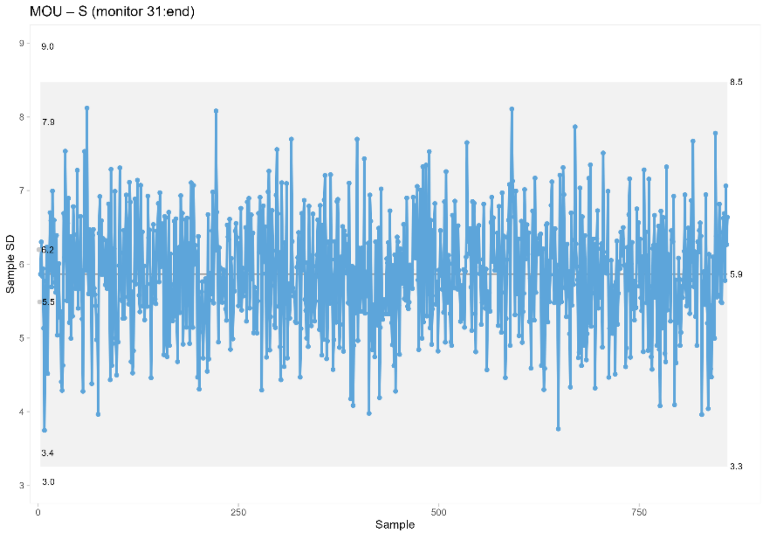


Figure 20: S-Charts MON samples 31:end

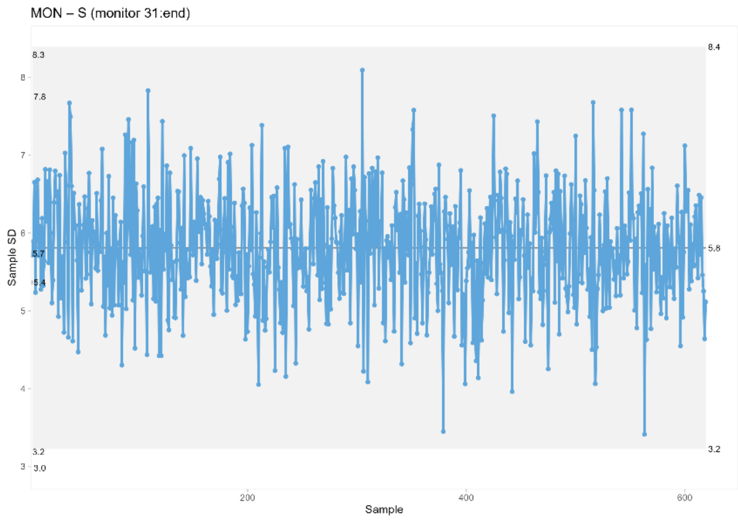
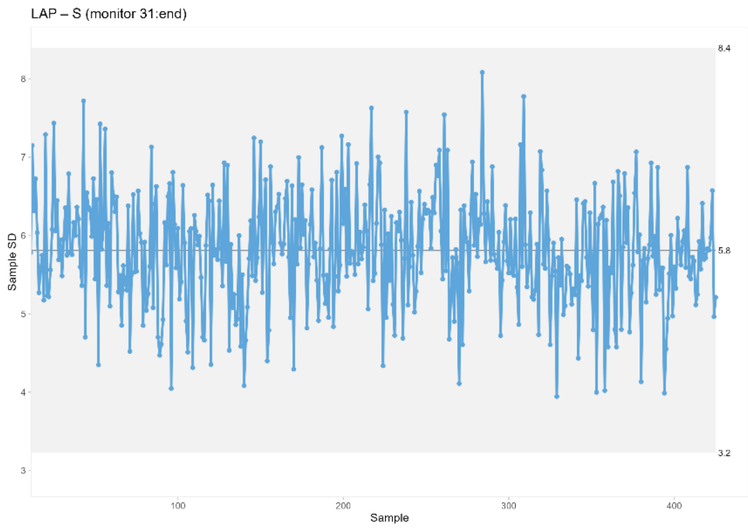


Figure 21: S-Charts LAP samples 31:end



3.3 Process Control Capabilities

Table 3: Process Control Capabilites

Product ID	**Cp**	**Cpk**	**Cpu**	**Cpl**
CLO	0.897	0.717	0.717	1.077
KEY	0.917	0.730	0.730	1.104
LAP	0.899	0.670	0.697	1.101
MON	0.890	0.700	0.700	1.080
MOU	0.915	0.725	0.725	1.105
SOF	18.155	1.087	35.223	1.087

The general rule of thumb is that Cpk must be greater than or equal to 1 for a process to be considered capable of meeting the VOC. Accordingly, only SOF is therefore capable of meeting the VOC.

3.4 Process Control Performance

The X-bar and S-Charts were examined based on the control limits established from the first 30 samples, which were then plotted with the remaining samples. The X-bar Charts display corresponding cyclic patterns, where the mean delivery times first steadily rise in a linear fashion before undergoing a sharp drop, and then rising in a similar fashion as before. This pattern is displayed across all product types, suggesting a larger system-wide effect rather than a product-specific trend. The X-bar charts display points above the control limit just before this sharp drop, and below the control limit at the bottom of this drop across product types. All product types display samples out of the control limit at the end of the monitoring period too. None of the product charts display extended sequential runs within the ±1 sigma region, indicating that no over-adjustment behaviour is evident and confirming natural variation. All product types display clusters of 4+ points above the 2-sigma bound around the 1st and 2nd peaks of the 2 upwards trends in the X-bar charts. This confirms that these points are not in statistical control.

4. Risk, Data Correction and Optimising for Maximum Profit

4.1 Estimating Type 1 Errors

<u>Rule A</u>

α (Rule A) = 0.2699796%

α (Rule B) = 6.911344%

α (Rule C) = 0.0004286034%

Figure 22: Screenshot of R code

```
# Rule A
PA <- 2 * (1 - pnorm(3))

# Rule B (7 consecutive within ±1σ)
PB <- (pnorm(1) - pnorm(-1))^7

# Rule C (4 consecutive outside ±2σ)
PC <- (2 * (1 - pnorm(2)))^4

PA*100
PB*100
PC*100
```

4.2 Estimating Type 2 Errors

Figure 23: Screenshot of R code

```
mu1 <- 25.028
sigma1 <- 0.017
LCL <- 25.011
UCL <- 25.089

zL <- (LCL - mu1)/sigma1
zU <- (UCL - mu1)/sigma1
beta <- pnorm(zU) - pnorm(zL)
power <- 1 - beta
```

β = 84.11783%

Therefore, there is an 84.1% chance of not detecting the out-of-control condition, illustrating that the control limits are too wide for such small mean shifts.

4.3 Updated Data Analysis

Figure 24: Updated Total Sales Value by Product Category

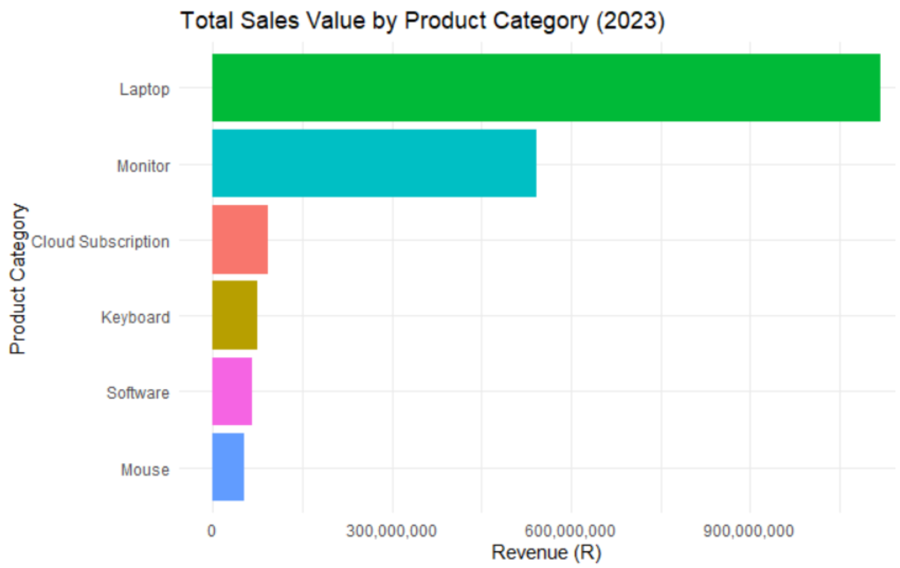
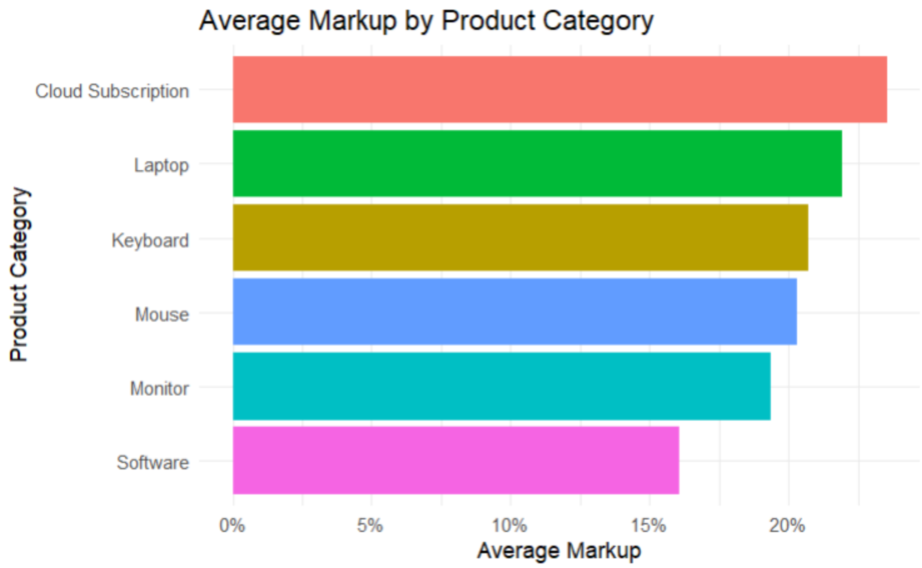


Figure 25: Updated Average Markup by Product Category



The updated data analysis on revenue and profitability by product class reveals a far more defined distribution across product categories compared to the old data set. Laptops now greatly dominate total sales in 2023, generating over R900 000 in revenue. Monitors brought in just under R600 000 in revenue, and the remaining product categories sit far below these 2 hardware categories, contributing far less to total revenue. This reveals that the business is actually more concentrated in high-value hardware products and is heavily reliant on these sales rather than being evenly-distributed across product categories.

Profitability analysis revealed similar insights to the old data, with cloud subscriptions remaining the most profitable product category with a markup of around 23%. This highlights a strong opportunity for the business to strengthen its profitability by expanding its high-margin digital product offerings. Monitors remain the company's 2^{nd} least profitable product, despite being responsible for the 2^{nd} highest amount of revenue, likely due to a strategic pricing decision.

Overall, the corrected data analysis reveals that the business does not have a diversified product mix for consistent revenue, but rather a dual performance profile: high revenue is driven by key hardware sales (monitors and laptops) and high profitability is driven by its digital products. This could be further strengthened and optimized by further capitalizing on the high profitability of the digital product market. Expanding its digital segment, particularly cloud services, could help in achieving a stable balance of growth and profitability.

While the updated datasets corrected inconsistencies and errors in product pricing and markup calculations, these calculations only affected this revenue and profitability analysis. The remainder of the data analysis will remain unchanged and will therefore reveal no new insights not already discussed in the first data analysis of this report.

5. Profit Optimization Problem

Method

In this analysis, the operational data from two separate coffee hops was analysed to evaluate the impact of different staffing levels on daily service capacity and profitability. The data of each coffee shop was first cleaned and grouped by staffing level (2-6), and then descriptive statistics were calculated such as the mean, median and 95^{th} percentile. Using these averages, the expected number of services that can be delivered within an 8-hours working day was estimated for each staffing level. Daily revenue was derived by multiplying the average services per day by the assumed margin of R30 per customer order, and profit was calculated by subtracting the barista fee of R1000 per barista per day.

Results

Table 4: KPI's per Barista Level: Shop 1

baristas	n_orders	mean_s	median_s	p95_s	avg_services_per_day	revenue_R
<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2	3556	100.17098	100	112	287.5084	8625.253
3	12126	66.61174	67	77	432.3562	12970.686
4	29305	49.98038	50	59	576.2261	17286.784
5	56701	39.96183	40	48	720.6876	21620.629
6	97895	33.35565	33	41	863.4220	25902.661

median_s	p95_s	avg_services_per_day	revenue_R	staff_cost_R	profit_R
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
100	112	287.5084	8625.253	2000	6625.253
67	77	432.3562	12970.686	3000	9970.686
50	59	576.2261	17286.784	4000	13286.784
40	48	720.6876	21620.629	5000	16620.629
33	41	863.4220	25902.661	6000	19902.661

Table 5: KPI's per Barista Level: Shop 2

baristas	n_orders	mean_s	median_s	p95_s	avg_services_per_day	revenue_R
<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2	8859	141.51462	141	154	203.5125	6105.376
3	19768	115.44091	116	126	249.4783	7484.348
4	35289	100.01527	100	109	287.9560	8638.681
5	54958	89.43597	89	98	322.0181	9660.543
6	78930	81.64272	82	89	352.7565	10582.695

mean_s	median_s	p95_s	avg_services_per_day	revenue_R	staff_cost_R	profit_R
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
141.51462	141	154	203.5125	6105.376	2000	4105.376
115.44091	116	126	249.4783	7484.348	3000	4484.348
100.01527	100	109	287.9560	8638.681	4000	4638.681
89.43597	89	98	322.0181	9660.543	5000	4660.543
81.64272	82	89	352.7565	10582.695	6000	4582.695

Figure 26: Average Services per 8-hour Day vs Baristas for Both Shops

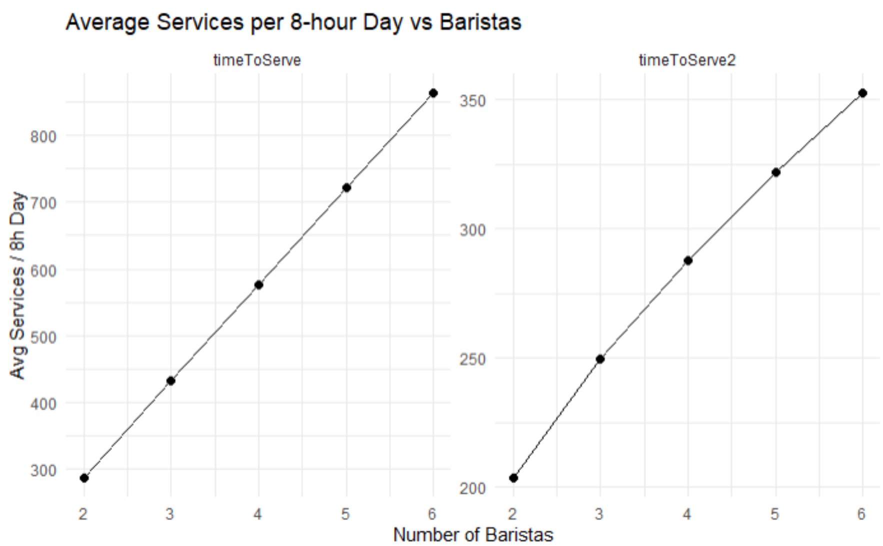
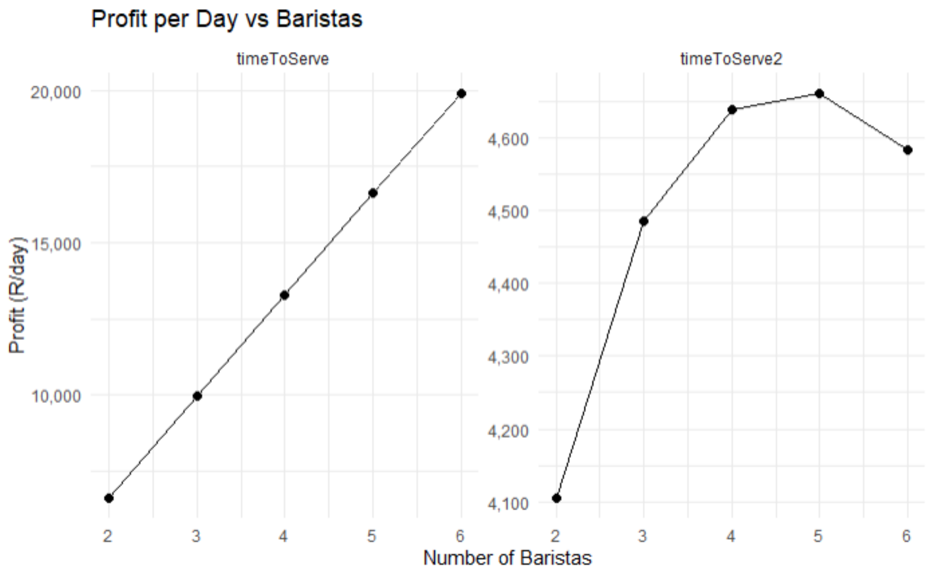


Figure 27: Profit per Day vs Baristas for Both Shops



Analysis

Shop 1 (timeToServe)

The profit per day for shop 1 rises in an almost-linear fashion as the number of baristas staffed increases. Optimizing profit for shop 1 is therefore straightforward: hire the maximum number of baristas that the shop is prepared to. Due to the high sales volume of shop 1, the daily profit will likely continue to increase as barista level increases for very long, as the extra R1000 salary per day will not greatly influence the daily profit. In this case, 6 baristas is the most profitable for shop 1. No information is given on demand, and it is therefore assumed for the sake of this problem that demand cannot be exceeded. Staffing 6 baristas results in a median service time of 33 seconds, and a profit of R19 903 per day.

Shop 2 (timeToServe2)

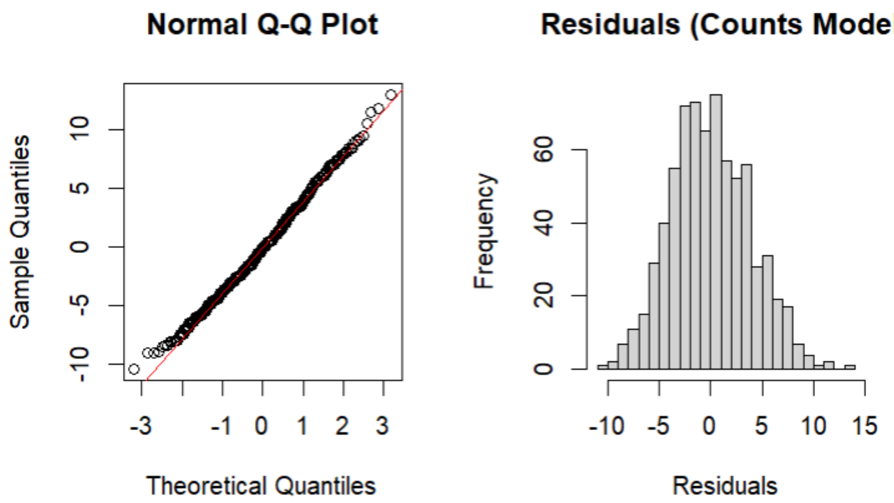
Shop 2 optimizes very differently to shop 1, as the daily profit vs number of baristas hired plot can be seen to slope off and begin to decrease after 5 baristas. This is due to shop 2 having less than half of shop 1's average services per day. This is likely largely due to shop 2's median service times being far longer than shop 1's for the same barista levels, which greatly restricts the number of services they can complete in a day. Therefore, the additional fee of R1000 per barista hired begins to influence the daily profit sooner as the number of baristas increases, until it will eventually result in the salaries of the hired baristas being greater than the revenue generated daily. The optimal number of baristas for shop 2 to maximise its daily profit is therefore 5, where median service time is 82 seconds and expected daily profit is R4583.

6. ANOVA Analysis

Analysis of Variance (ANOVA) is a statistical method that is used to compare the means of three or more groups for a single variable. It compares the variability within groups to decide whether observed differences are statistically meaningful or due to random variation. A 2-Way ANOVA will therefore be applied to evaluate whether there is a statistically significant difference in the number of monthly sales, and between the two product categories of laptops and monitors. These products were selected for analysis, as significant variation in their sales volume would have a major impact on overall business performance. This is because they are responsible for the majority of the business's total revenue generation. The application of ANOVA will therefore reveal whether trends observed in the analysis of this data reflect seasonal trends or are simply due to natural variation in sales.

6.1 Monthly Variation of Count of Laptop and Monitor Sales: 2-Way ANOVA Test

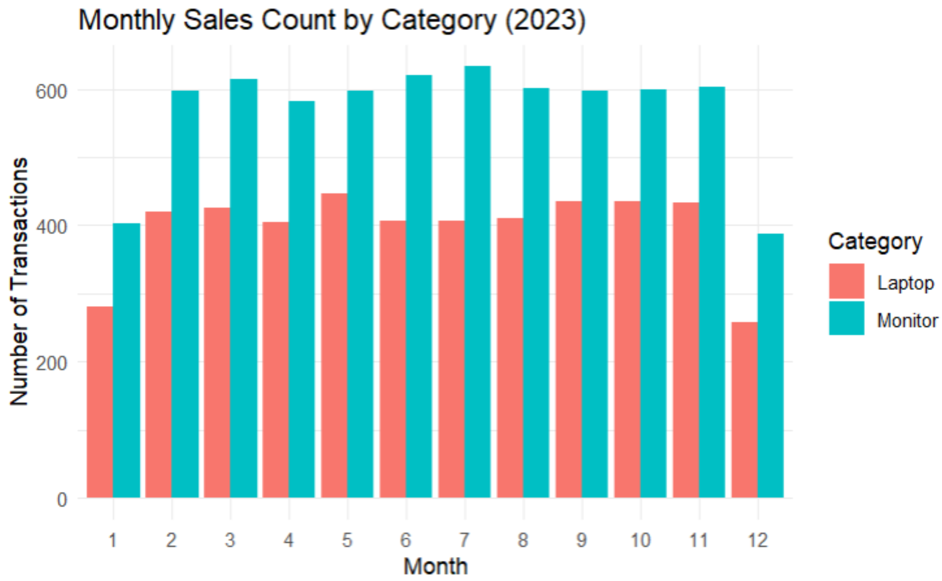
Figure 28: Normal Q-Q Plot and Residuals Plot



The residuals diagnostics for the ANOVA on count of sales by month for the 2 product categories shows that the model's assumptions were met. The normal Q-Q plot shows that the residuals follow the theoretical normal line very closely, and the histogram of the residuals show a symmetric bell-shaped distribution. Therefore, the model errors are normally distributed, and variance is stable, validating the use of ANOVA for these variables.

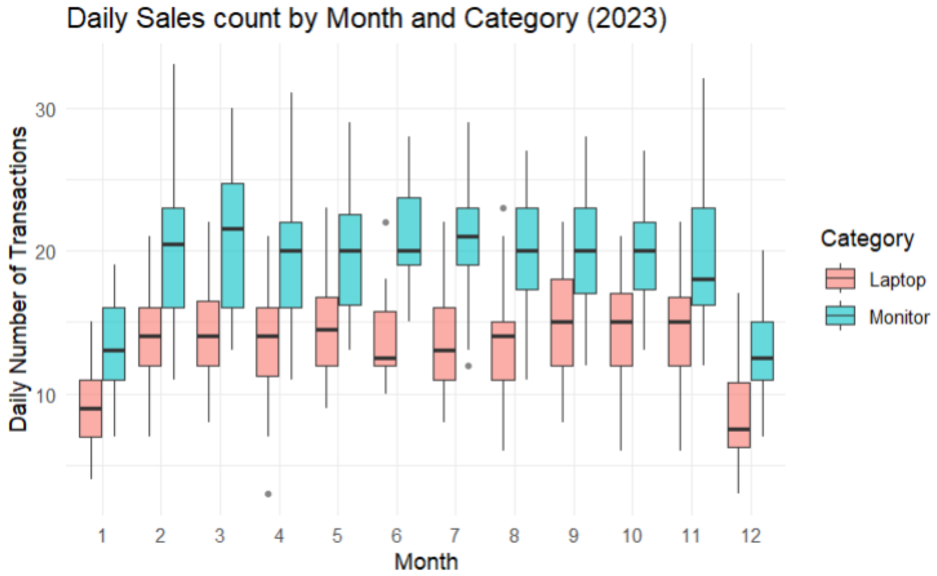
- H0: There is no significant difference in the number of sales across months or between categories
- H1: There is a significant difference in the number of sales across months and/or between categories

Figure 29: Monthly Sales Count by Category (2023)



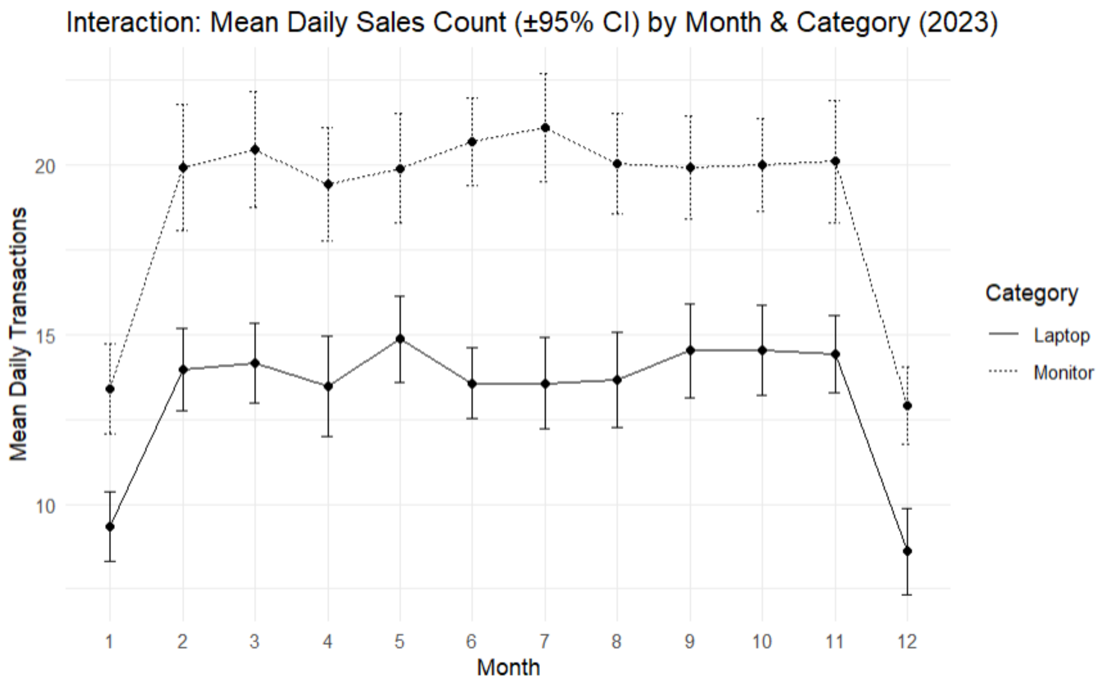
It can be seen from the bar chat that the number of sales for monitors was greater every month than those of laptops. Sales activity for both categories were consistent from February through November, but both recorded drops in January and December. This aligns with the analysis done on total count of sales by month, which displayed the same seasonality. This pattern is likely due to seasonal factors such as the post-holiday slowdown and year-end operational closes. These insights show that monitors contribute to sustained sales volume, and laptops, although recording less sales monthly, remain the higher-revenue product due to the higher unit value.

Figure 30: Box Plots of Daily Count by Month and Category (2023)



Overall, it can be seen that monitor sales consistently exceed laptop sales on a daily basis, affirming the insights from the monthly totals. Daily sales for monitors typically range between 15-25 sales, while laptops average between 10-18 sales per day. The same seasonality of consistency from months February – November, with dips in months December and January. The relatively narrow spread of each box suggest that daily sales volumes are relatively stable and without extreme fluctuation. Together with the plot of total sales count, sales patterns can be seen to be predictable and stable.

Figure 31: Mean Daily Sales Count by Month & Category (2023)



This interaction plot illustrates how the average daily number of sales transactions for laptops and monitors changed from month to month in 2023, with 95% confidence intervals indicating the reliability of these averages. The near-parallel nature of these lines indicates that there is no strong effect of interaction between the 2 product categories throughout the year, as both categories experience similar seasonal trends.

Figure 32: Post-hoc Comparisons of Laptop vs Monitor Daily Sales Counts by Month

```
P value adjustment: tukey method for comparing a family of 12 estimates

Category differences within each month:
month_f = 1:
 contrast         estimate   SE  df t.ratio p.value
 Laptop - Monitor    -4.07 1.02 696  -3.994  0.0001

month_f = 2:
 contrast         estimate   SE  df t.ratio p.value
 Laptop - Monitor    -5.97 1.02 696  -5.860  <.0001

month_f = 3:
 contrast         estimate   SE  df t.ratio p.value
 Laptop - Monitor    -6.30 1.02 696  -6.187  <.0001

month_f = 4:
 contrast         estimate   SE  df t.ratio p.value
 Laptop - Monitor    -5.97 1.02 696  -5.860  <.0001

month_f = 5:
 contrast         estimate   SE  df t.ratio p.value
 Laptop - Monitor    -5.03 1.02 696  -4.943  <.0001
month_f = 6:
 contrast         estimate   SE  df t.ratio p.value
 Laptop - Monitor    -7.13 1.02 696  -7.006  <.0001

month_f = 7:
 contrast         estimate   SE  df t.ratio p.value
 Laptop - Monitor    -7.53 1.02 696  -7.399  <.0001

month_f = 8:
 contrast         estimate   SE  df t.ratio p.value
 Laptop - Monitor    -6.37 1.02 696  -6.253  <.0001

month_f = 9:
 contrast         estimate   SE  df t.ratio p.value
 Laptop - Monitor    -5.40 1.02 696  -5.303  <.0001

month_f = 10:
 contrast         estimate   SE  df t.ratio p.value
 Laptop - Monitor    -5.47 1.02 696  -5.369  <.0001

month_f = 11:
 contrast         estimate   SE  df t.ratio p.value
 Laptop - Monitor    -5.67 1.02 696  -5.565  <.0001

month_f = 12:
 contrast         estimate   SE  df t.ratio p.value
 Laptop - Monitor    -4.30 1.02 696  -4.223  <.0001
```

The results of the Tukey post-hoc test reveal consistent and statistically significant differences in daily sales counts between laptops and monitors across all months in 2023, as $p < 0.0001$ for every month. The average difference in sales ranges between 4-7 more transactions for monitors per day.

Based on the ANOVA and Tukey post-hoc results, all p-values were well below 0.05, and the null hypothesis is therefore rejected in favour of H1. There are therefore statistically significant differences in the sales count of monitors and laptops across all months of the year.

7. Reliability of Service

Reliable service is expected when the number of workers present is 15 or more. Therefore, in these 397 days there were 366 days where reliable service can be expected, corresponding to a reliability of 92.19%.

Optimisation of Profit

There are assumed to be 16 staff on the roster each day, and therefore a maximum staffing of 16 workers before hiring extra staff. Service is reliable if more than 14 workers are staffed on a given day, otherwise an average loss of R20 000 is incurred per unreliable day. Hiring an extra person costs R25 000 per month per person, or R300 000 per year per person. The objective is therefore to choose the number of rostered staff (n) to minimise expected annual cost.

Method

Let $X \sim$ Binomial ($n=16$, p) be the number present. Each worker is assumed to have an independent probability p of being present on a given day. Using the given data, an estimated p value as calculated for each worker count by solving the binomial probability formula. These individual p values were then combined through the use of a weighted average, where categories with more observed days contributed more heavily to the overall mean. This ensure that categories with greater reliabilities were weighted heavier. This weighted p value was then used to model the likelihood of having at least 15 workers on a given day, and therefore the reliability. This probability was then used to estimate the number of unreliable days per year, and the corresponding penalty that incurred. Where additional employees are hired, a staffing cost is incurred per extra staff member.

Results

Table 6: Estimated Cost Calculation for Different Staffing Levels

n	p	reliability	expected_bad_days	annual_loss_R	extra_staff_cost_R	total_expected_cost_R
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
15	0.9761922	0.6966746	110.7137772	2214275.544	0e+00	2214275.5
16	0.9761922	0.9454690	19.9038259	398076.518	0e+00	398076.5
17	0.9761922	0.9928550	2.6079411	52158.823	3e+05	352158.8
18	0.9761922	0.9992478	0.2745379	5490.757	6e+05	605490.8

The results in table 6 illustrates the effect of increasing the number of workers on the roster on overall reliability and total expected annual cost. With only 15 workers (the minimum amount to achieve reliability), the company achieves a reliability of 69.7%, resulting in an estimated 111 unreliable days in a year. This would result in an estimated annual loss of R2 214 276, and this is therefore financially unfeasible

If the company employs 16 workers (n=16), reliability is greatly improved to 94.5% and the expected annual bad days drops to 20. Annual loss is estimated at R398 077, and the total annual cost of the company drops to R398 077. Increasing the roster to 17 workers results in the optimal solution for minimizing this company's expected costs and maximizing their profit. With 17 workers on the roster, the reliability of the company rises further to 99.3%, resulting in an estimated 3 annual bad days. The total annual loss of the company is R52 159, and the total expected cost is R352 158.

While increasing the roster to 18 workers can be seen to improve reliability, further decrease the expected annual bad days and annual loss, the cost of hiring another employee results in the total expected costs of the company increasing to R605 491. Therefore, hiring 17 workers achieves the optimal profit for the company by achieving the best balance between cost and service reliability.

Conclusion

This report analysed and interpreted sales and operational data for a business selling various hardware and software products, with the aim of identifying potential quality issues and performance improvement opportunities. The data was first cleaned through a structured wrangling process to remove invalid and duplicate entries, resulting in a reliable dataset for analysis. Descriptive statistics were then applied to investigate key patterns such as revenue distribution, customer demographics, delivery performance, and product profitability.

The results revealed that revenue was relatively balanced across product categories, with monitors and laptops contributing the largest share of sales. However, while cloud subscriptions exhibited the highest markup, their high pricing may have limited demand. Process capability and statistical process control analyses showed that delivery times were not well-centred and displayed signs of instability, particularly for certain product classes. This indicates a need for improved process consistency and scheduling control.

A two-way ANOVA was performed to test for significant differences in sales across months and between major product categories. The analysis confirmed that variations in sales volume were statistically significant, implying that both time-based factors and product type influence sales performance. Finally, a reliability analysis assessed staffing and service delivery consistency, identifying potential risks related to maintaining adequate staffing levels to meet service standards.

Overall, this study provided data-driven recommendations aimed at improving process efficiency, pricing strategy, and service reliability. Implementing these findings could enhance the company’s operational performance, customer satisfaction, and long-term profitability.

References

- Field, A. (2013). *Discovering Statistics Using R.* London: SAGE Publications.
- Montgomery, D.C. (2019). *Introduction to Statistical Quality Control* (8th ed.). Hoboken, NJ: John Wiley & Sons.
- The R Foundation for Statistical Computing. (2024). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/
- Wickham, H., François, R., Henry, L. and Müller, K. (2023). *dplyr: A Grammar of Data Manipulation.* R package version 1.1.3. Available at: https://CRAN.R-project.org/package=dplyr
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* 2nd ed. New York: Springer-Verlag.
- Kassambara, A. (2023). *R Statistical Tutorials: ANOVA and MANOVA in R.* Available at: http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance
- Hothorn, T., Bretz, F. and Westfall, P. (2008). *Simultaneous Inference in General Parametric Models. Biometrical Journal*, 50(3), pp.346–363. (Referenced for Tukey and post-hoc testing concepts.)
- R Documentation. (2024). *qcc: Quality Control Charts.* Available at: https://CRAN.R-project.org/package=qcc