Quality Assurance 344

ECSA Final Deliverable

Stellenbosch University

Paul Jozua Joubert

27497232

1 October 2025

I. Table of Contents

II. Deliverable Part 1: Data Analysis

1. Introduction

This analysis provides an initial overview of the company's customer base, product catalogue, and sales activity for the years 2022 and 2023. The aim of this report is to analyse the data sets, identify key trends, understand customer data, evaluate profitability and highlight any problems. This will help to improve the business in future developments. Through this first part Data tables and Graphs will be used to achieve this.

2. Customer Analysis

n_customers	mean_age	sd_age	min_age	median_age	max_age	mean_income
5000	51.5538	21.21609601	16	51	105	80797
sd_income	min_income	median_income	max_income			
33150.10674	5000	85000	140000			

Table 1: Customer Overview

Var1	Freq
Female	2432
Male	2350
Other	218

Table 3: Gender Distribution

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

Table 2: Top Cities by Customer

The customer dataset consists of 5000 individuals with average age of 51.5 years. Income varies from R 5000 to R140 000, with median of R 85 000, indicating most customers falls in middle to upper income groups.

Most customers are located in major metropolitan areas which shows high demand in densely populated areas. This makes it ideal for future expansion.

No major difference in customer gender. This implies that products appeal to both genders equally.

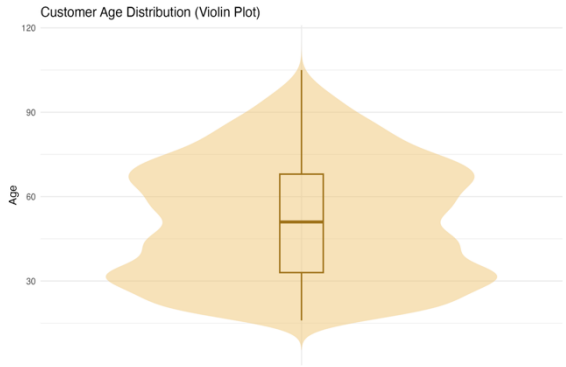


Figure 1: Customer Age Distribution

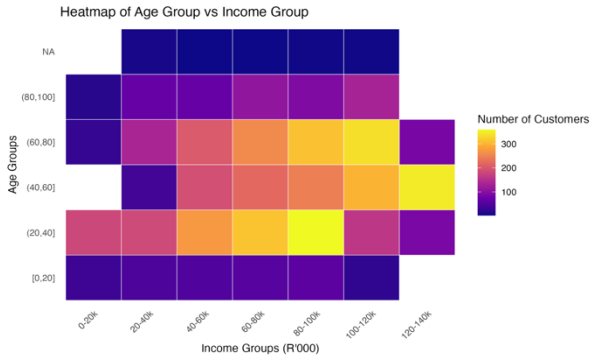


Figure 2: Age vs Income Heatmap

With the help of the Violin Plot (Figure 1) we can see that most Customers are between 35 and 70 years old, with highest density in 45 – 60 years.
The age distribution is wide but is mostly centered around middle-aged adults making them their primary market segment, as they likely have stable incomes.

Figure 2 strengthens this by with the heatmap indicating that higher income is in the areas between 40 and 60 years old, confirming previous statements. It also gives us a idea of which age groups to focus on more as the age group 40 – 60 has highest income compared to other age groups.

3. Product Analysis

n_products	mean_price	sd_price	min_price	median_price	max_price
60	4493.592833	6503.77015	350.45	794.185	19725.18

mean_markup	sd_markup
20.46166667	6.072597849

Table 4: Product Summery

Category	n_products	mean_price	sd_price	mean_markup	mean_cost_pct	mean_cost_abs
Laptop	10	5217.545	7315.414265	20.623	4066.572856	5196.922
Monitor	10	5014.17	6983.932943	20.727	3979.128151	4993.443
Keyboard	10	4638.172	7141.548814	20.161	3771.097394	4618.011
Mouse	10	4585.465	7094.587988	20.668	3538.667356	4564.797
Software	10	3814.344	6143.554926	20.038	3244.1416	3794.306
Cloud Subscription	10	3691.861	5812.224378	20.553	3030.092025	3671.308

Table 5: Product Data by Category

Laptops and monitors have highest average selling prices and highest revenue, making them premium products. Markups are consistent across all categories.

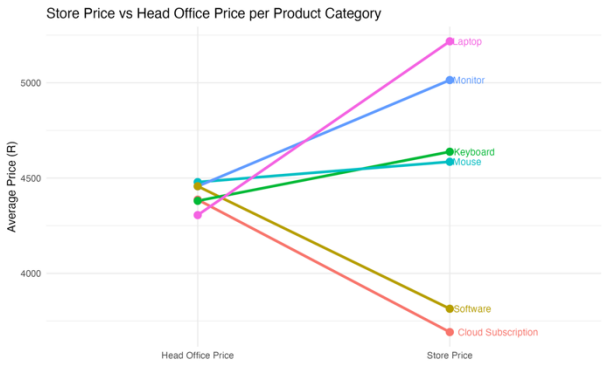


Figure 4: Store vs Head Office Prices

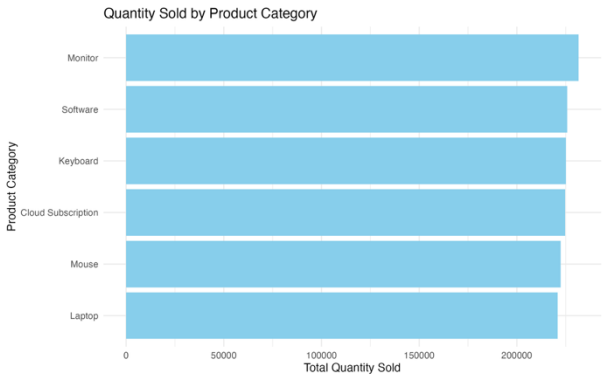


Figure 3: Quantity Sold per Category

As shown in Figure 4, the prices that products are sold at and the prices that Head Office has given said products do not align. This shows inconsistency between head office and stores.

Monitors and software are top sellers, but all sales volume is very close in amount (Figure 3).

4. Sales Performance Analysis

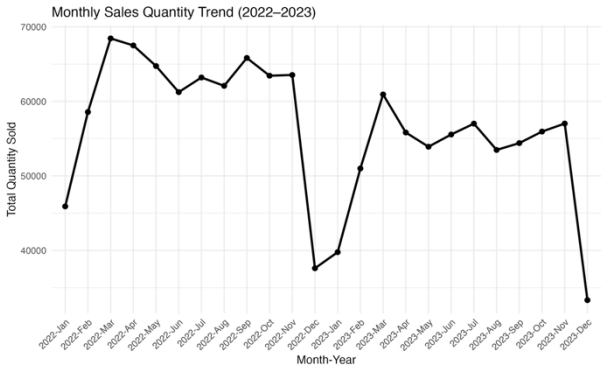


Figure 6: Monthly Sales Trend

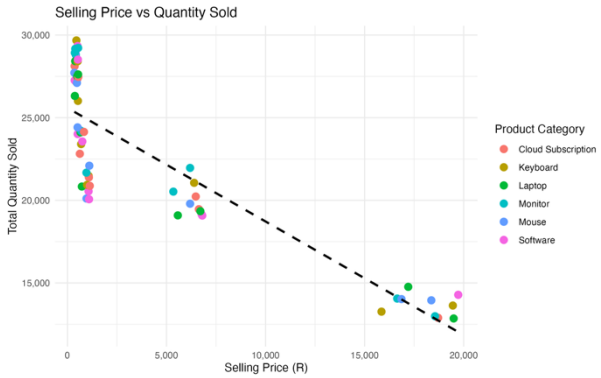


Figure 5: Selling Price vs Quantity Sold

Year	Total_Sales
2022	2,320,410,018
2023	2,032,177,660

Table 6: Total Sales Value

As seen in Figure 6 sales increase sharply at beginning of each year and peaks around March to April. There are very big drops in Dec to Jan showing demand is very low for those two months, possibly due to Holiday. Overall, 2022 shows higher sales volumes then 2023.

A clear negative trend is visible in Figure 5, showing the higher the products price the lower the amount sold. This is good as it confirms price sensitivity through different categories. But prices of same types of products vary a lot also which does not make sence.

5. Operational Efficiency

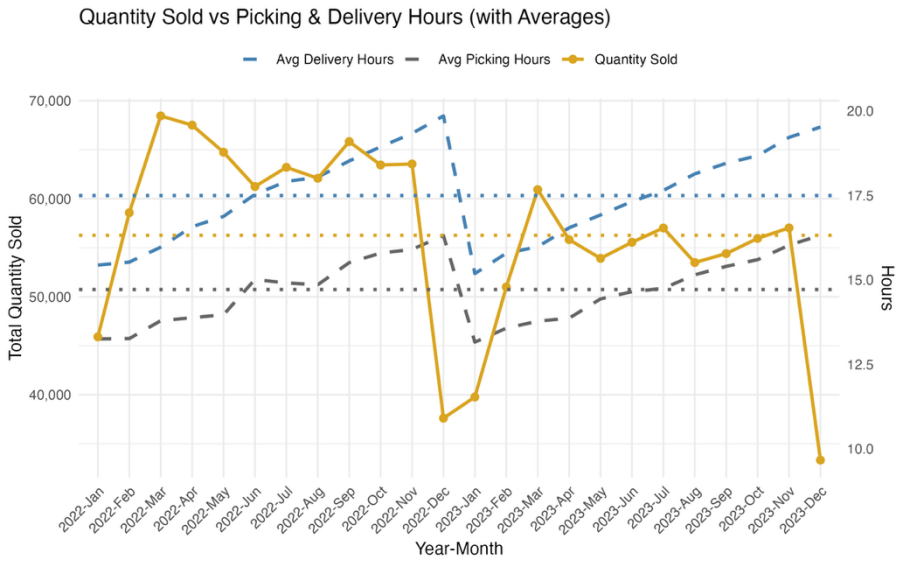


Figure 7: Quantity Sold vs Avg Picking & Delivery Hours

Picking and delivery times show an upward trend, but delivery's curve stays much higher than picking. Months with higher sales does not necessarily have lower picking and delivery times, suggesting operational delays do not immediately influent sales volume. As sales rise picking and delivery times also rise but reacts to sales volumes with slight delay.

6. Conclusion For Part 1

Customer insights: Majority of customers are middle aged and have a middle to high income. They are mainly located in large metropolitan areas and indicates they should be the focus.

Product performance: Products share fairly the same wait in revenue as higher-end products tend to have higher markups, but lower-end products tend to sell more.
Prices between product stores and head office very greatly and need to be addressed.

Sales behavior: Sales volumes peak in the first and last quarters of the year but drastically decreases over Dec and Jan period due to holidays. Figure 5 also shows us the difference in sales of higher value products compared to lower value products.

Operational efficiency: Delivery times are much higher than picking times, this tells us that the delivery process is not very efficient and should be improved.

Overall, the company has a strong foundation but through minor adjustments profit can significantly be improved.

III. Deliverable Part 3: SPC

1. Introduction

This section analyzes delivery time process stability using X and S control charts for the different Product Types. The first 30 subgroups of 24 deliveries were used as a foundation to establish a baseline for center lines and control limits. The remaining subgroups were evaluated using SPC rules. Then process capability indices were calculated using first 1000 deliveries.

2. Data Preparation

- Data form sales were sorted by Year-Month-Day-Hour.
- Product type was created with first 3 letters of Product ID.
- Data was grouped into subgroups of 24.

p3_product_counts

ProductType	n_rows	n_subgroups	has_phase1_30
SOF	20749	864	TRUE
MOU	20662	860	TRUE
KEY	17920	746	TRUE
CLO	15598	649	TRUE
MON	14864	619	TRUE
LAP	10207	425	TRUE

Table 7: Product Count

3. Phase 1 - Control Limit Calculations

- Calculated CL, LCL and UCL for X and S using first 30 subgroups per type.

p3_phase1_limits_summary

ProductType	n_rows	n_subgroups	Xbar_CL	Xbar_LCL	Xbar_UCL	S_CL	S_LCL	S_UCL
SOF	20749	864	0.9556375000000000	0.7906837384411540	1.1205912615588500	0.2985784224642470 0	0.1804656190886980 0	0.4166912258397970 0
MOU	20662	860	19.239138888888900	16.086064893781500	22.392212883996300	5.707295489811360	3.449581538378830	7.965009441243900
KEY	17920	746	19.194000000000000	15.949771895364400	22.43822810463560	5.87229112232588	3.5493075625265700	8.1952746821251900
CLO	15598	649	19.125944444444400	15.888720229210900	22.363168659678000	5.859613568149350	3.5416450441030500	8.177582092195660
MON	14864	619	19.425944444444400	16.17954756191880	22.672341326970100	5.876216769579740	3.5516802871064000	8.200753252053070
LAP	10207	425	19.523861111111100	16.26244771335250	22.785274508869700	5.903397764949660	3.568108919545410	8.238686610353910

Table 8: Phase 1 Limits

4. Phase 2 – Monitoring (SPC Rules)

Three Checks were Applied:

- Rule A (S chart): any subgroup S > S-UCL.
- Rule B (S chart): longest run of S within ±1σ of S-CL (good control).
- Rule C (X̄ chart): any window of four consecutive subgroup means above +2σ.

p3_rules_summary

ProductType	A_S_over_UCL	B_longest_S_run_in_pm1sigma	C_any_4_Xbar_over_plus2sigma
MOU	3	14	292
SOF	0	12	298
KEY	0	12	253
CLO	0	17	252
MON	0	29	200
LAP	0	18	131

Table 9: Rules A - C Summery

p3_ruleA_s_over_UCL

ProductType	A_first3	A_last3	A_total
SOF			0
MOU	129, 483, 592	129, 483, 592	3
KEY			0
CLO			0
MON			0
LAP			0

Table 12: Rule A

p3_ruleB_longest_run_s_within_pm1sigma

ProductType	B_longest_run_len	B_start_subgroup
SOF	12	538
MOU	14	238
KEY	12	178
CLO	17	476
MON	29	238
LAP	18	116

Table 11: Rule B

p3_ruleC_four_xbar_over_plus2sigma

ProductType	C_first3	C_last3	C_total
SOF	205, 206, 207	862, 863, 864	298
MOU	197, 226, 227	858, 859, 860	292
KEY	102, 115, 116	744, 745, 746	253
CLO	125, 173, 174	647, 648, 649	252
MON	137, 174, 175	613, 618, 619	200
LAP	122, 132, 133	423, 424, 425	131

Table 10: Rule C

Rule A: flags potential special cause variability, If **Rule B** runs long, it suggests stable spread, **Rule C** flags mean shifts.

5. Control Charts (X and S)

A. Product Type: CLO

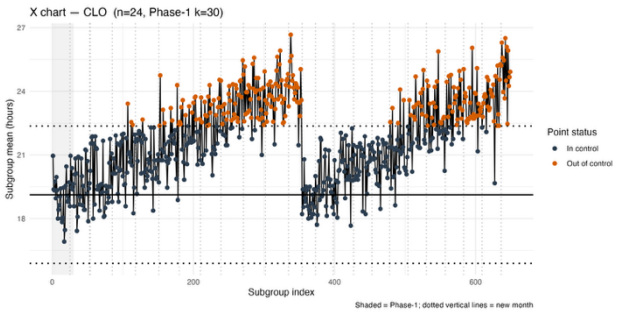


Figure 9: X Chart: CLO

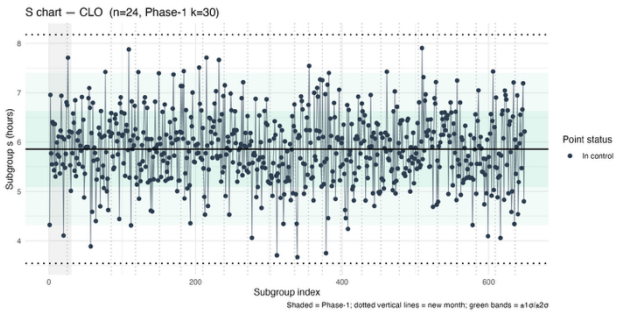


Figure 9: S Chart: CLO

B. Product Type: KEY

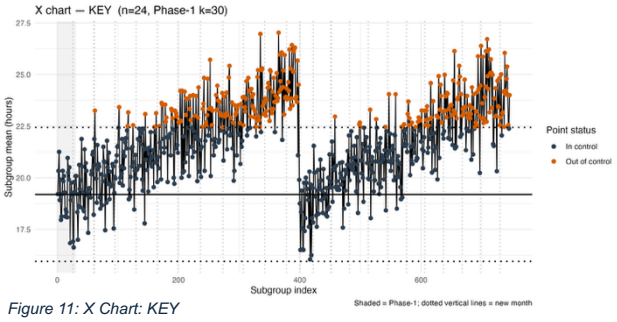


Figure 11: X Chart: KEY

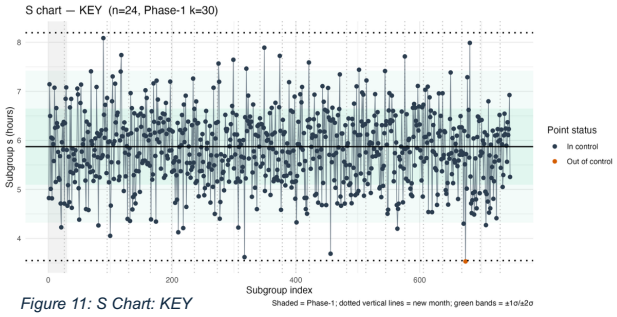


Figure 11: S Chart: KEY

C. Product Type: LAP

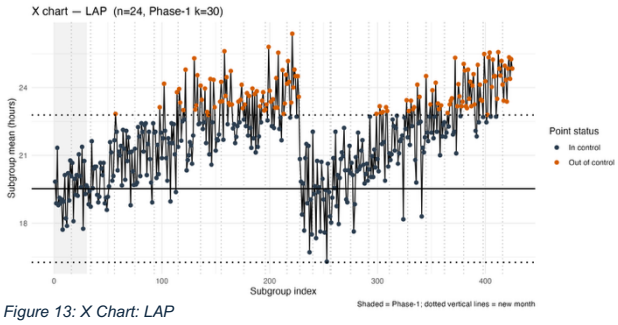


Figure 13: X Chart: LAP

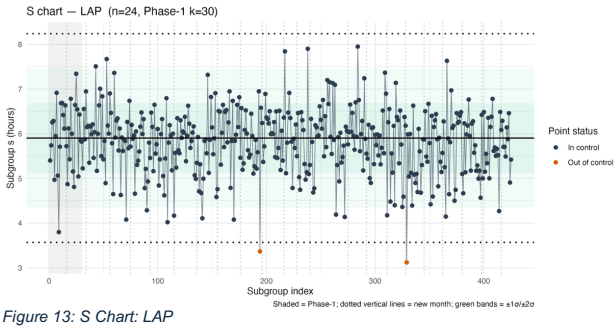


Figure 13: S Chart: LAP

D. Product Type: MON

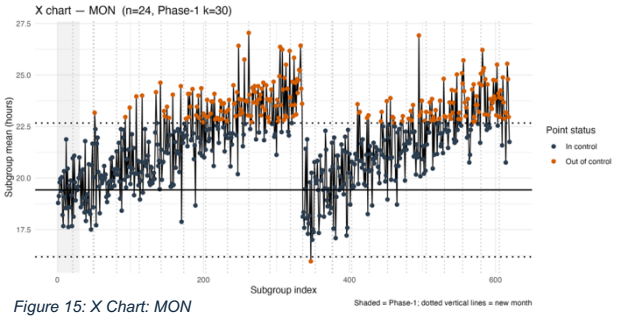


Figure 15: X Chart: MON

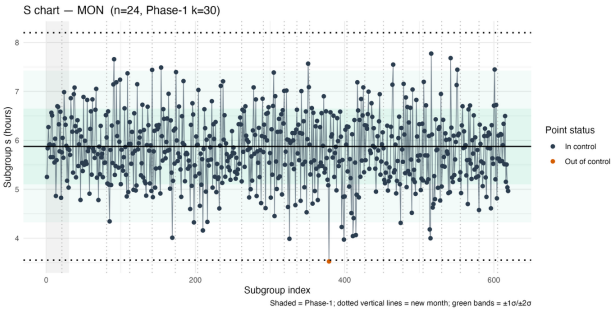


Figure 15: S Chart: MON

E. Product Type: SOF

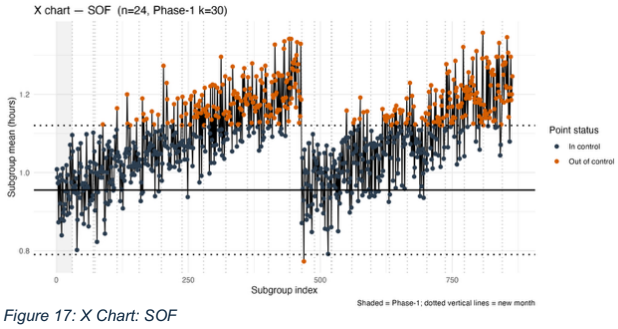


Figure 17: X Chart: SOF

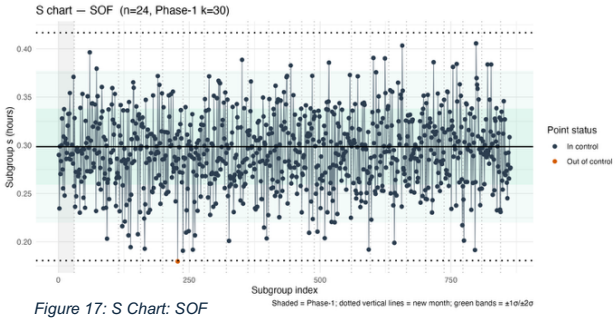


Figure 17: S Chart: SOF

F. Product Type: MOU

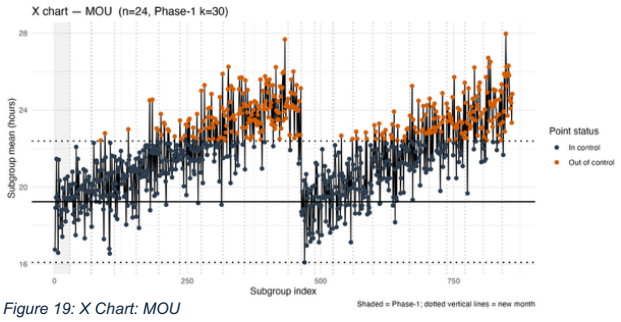


Figure 19: X Chart: MOU

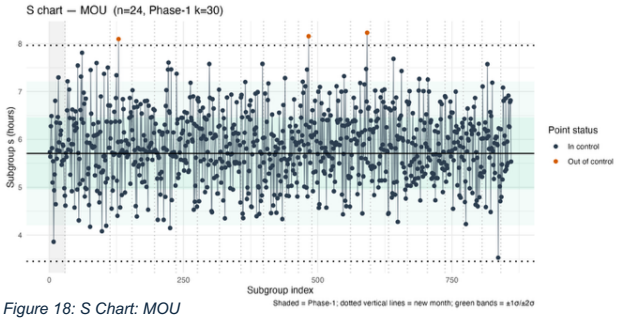


Figure 18: S Chart: MOU

The X Charts shows how average delivery time of subgroups changes over time.
All of them has a shift upward as time pass and has a lot of points that crosses the UCL.
This shows that the process are unstable.

The S Charts display the variation in each subgroup.
Products Type CLO is only one that does not exceed control limit, this says CLO has a stable variation and that only mean is shifting.
All other product types exceed the control limits but is very limited, this tells us that there variation is mostly stable.

6. Process Capability

Process capabilities were calculated using first 1000 deliveries. LSL = 0 hours and USL = 32 hours.

p3_capability_indices

ProductType	mu	sigma	Cp	Cpl	Cpu	Cpk	CapabilityFlag
SOF	0.954925	0.2935930793994840	18.165732462911400	1.084182004509110	35.24728292131380	1.084182004509110	Marginal
KEY	19.276000000000000	5.815194965991440	0.9171374931578150	1.1049213948818800	0.7293535914337520	0.7293535914337520	Not capable
MOU	19.297500000000000	5.827602311302550	0.9151848476326900	1.1037987248244900	0.7265709704408900	0.7265709704408900	Not capable
CLO	19.226	5.940805431014320	0.8977458351842930	1.0787538392033300	0.7167378311652600	0.7167378311652600	Not capable
MON	19.410000000000000	5.998919155233840	0.8890490428897000	1.078527620155570	0.6995704656238330	0.6995704656238330	Not capable
LAP	19.606	5.933958869748580	0.8987816482050040	1.1013445621692100	0.696218734240801	0.696218734240801	Not capable

Table 13: Capability Indices

Although SOF is only with Cpk > 1 it is still below 1.33.

All other Product Types have Cpk < 1, showing their delivery times are to variable and will not meet VOC consistently.

7. Conclusion

For most product types of the process remained in control with only a few out-of-control points, this is mostly from the X Charts as the mean keeps shifting, while variability stays fairly constant.

From SPC Rules:

Rule A, variability was not very high, as the highest was MOU.

Rule B, good consistency was found especially from MON and LAP

Rule C, was biggest issue as mean shifts was main problem, this suggests a gradual performance drift as time passed.

Prosses capability showed that only one product type marginally meets the VOC requirement, whereas all other product types fell below acceptable levels.

IV. Deliverable Part 4

1. (4.1) Type I Error

A Type I error happens when it looks like a process is out of control, but it isn't.

Rule A (S > UCL)

In a stable process there is a 0.00135 or 0.135% chance that one subgroup standard deviation is randomly above the UCL.
This means that a process can be fine but we can expect about 1 false alarm in every 740 samples.

Rule B (long run within ±1σ)

This rule only checks for good control and low variation, there is no alarm for a out-of-control process.
This means there is no Type I error.

Rule C (4 points above +2σ on $\bar{X}$ chart)

For four subgroup means to be above +2 standard deviations in a row, while process is still in control, has a extremely low probability of 0.00000027. This means that if Rule C is triggered it is almost guaranteed there is a mean shift.

2. (4.2) Type II Error

A Type II error is when we fail to pick up mean shift, even if one happened, because the sample mean stays between the control limits.

In this example:

The process originally was centered at 25.050L & control limits are 25.011 and 25.089.

The process means shifts to 25.028L and standard deviation to 0.017L.

If we calculate the possibility that sample means produced is still inside the control limits, we get 0.84 or 84%.

This shows us that there is 84% chance that the shift will go unnoticed, since shift is small X values stay mostly inside control limits.

3. (4.3) Updated Data

A. Customer Analyses:

As customer data was not influenced by any of the changes that were made in product and product head office, there will be no difference in any of the analysis.

Therefore, no need to compare any customer data.

B. Product Analyses (New vs Old)

New Product Summary:

products_summary

n_products	mean_price	sd_price	min_price	median_price	max_price	mean_markup	sd_markup
60	4493.592833333330	6503.770149844410	350.45	794.185	19725.18	20.461666666666700	6.072597848760310

Table 14: New Product Summary

Old Product Summary:

n_products	mean_price	sd_price	min_price	median_price	max_price
60	4493.592833	6503.77015	350.45	794.185	19725.18

mean_markup	sd_markup
20.46166667	6.072597849

Table 15: Old Product Summary

Product Summaries has not changed, this confirms that the data correction did not change the number of products or basic price structure.

New Products by Category:

products_by_category

Category	n_products	mean_price	sd_price	mean_markup	mean_cost_pct	mean_cost_abs
Laptop	10	18086.429	1357.4277979853400	18.43	14765.367589	18067.999
Monitor	10	6310.525	501.8677101532280	23.868	4817.4816804	6286.657
Cloud Subscription	10	1019.062	118.31961477286800	19.956	816.7359817	999.106
Keyboard	10	644.66	107.23235892210900	23.981000000000000	489.68263880000000	620.679
Software	10	506.183	44.467830082231600	16.04	426.1071797	490.143
Mouse	10	394.698	33.844281906664500	20.495	314.32431300000000	374.203

Table 16: New Products by Category

Old Products by Category:

Category	n_products	mean_price	sd_price	mean_markup	mean_cost_pct	mean_cost_abs
Laptop	10	5217.545	7315.414265	20.623	4066.572856	5196.922
Monitor	10	5014.17	6983.932943	20.727	3979.128151	4993.443
Keyboard	10	4638.172	7141.548814	20.161	3771.097394	4618.011
Mouse	10	4585.465	7094.587988	20.668	3538.667356	4564.797
Software	10	3814.344	6143.554926	20.038	3244.1416	3794.306
Cloud Subscription	10	3691.861	5812.224378	20.553	3030.092025	3671.308

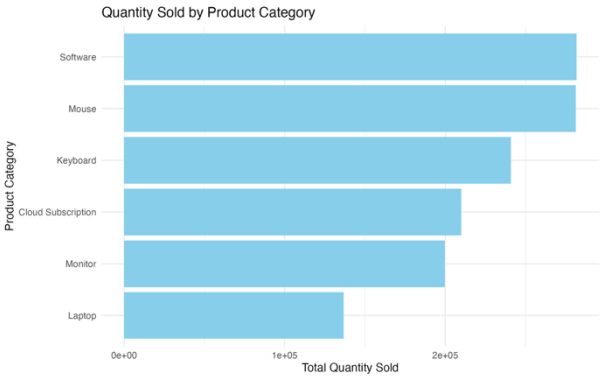
Table 17: Old Products by Category

The new data sets show clear changes:

- Categories like Laptop and Monitor's mean prices has got much higher, which aligns with more realistic values for their product types
- Smaller categories like mouse and software also have more realistic prices.

New files overall prices look much better.

New:



Old:

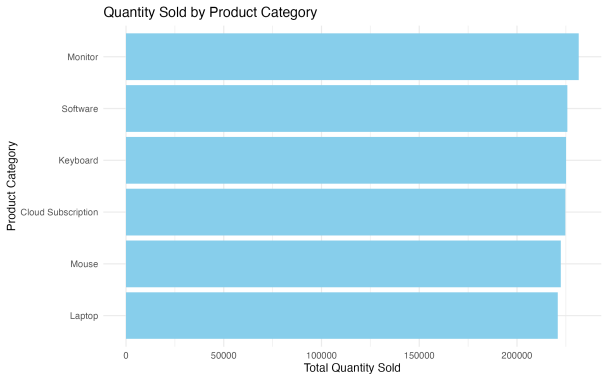
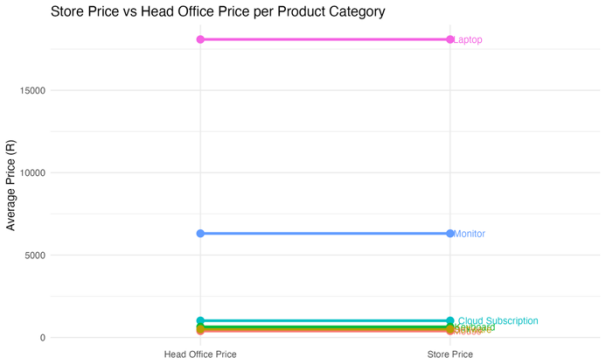


Table 18: New vs Old Quantity by Product Category

New graph for quantity also looks improved, as it follows a more realistic sales pattern. Products like laptop and mouse almost had same number of sales in old graph, which is not likely as products like laptops is much more expensive than products like mouse or keyboard.

New Graph. Is definitely more reliable.

New:



Old:

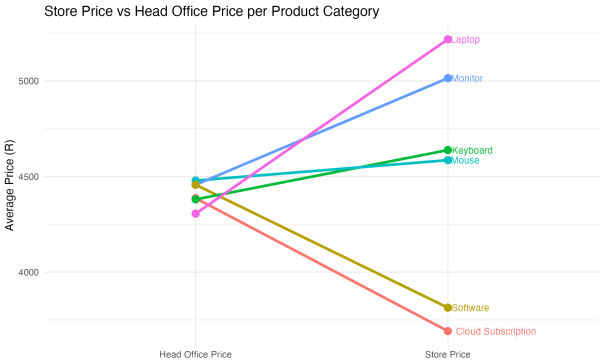


Table 19: Store vs Head Office Prices

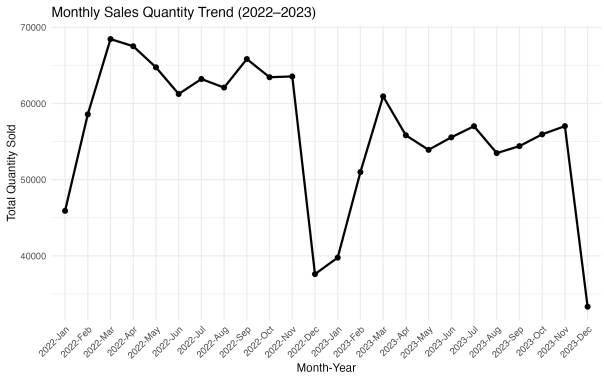
Old slope chart showed large differences between head office and product, indicating then there was a fault in data.

New graph shows no difference between the two.

This confirms that the rebuild of the data was successful.

C. Sales Analyses

New:



Old:

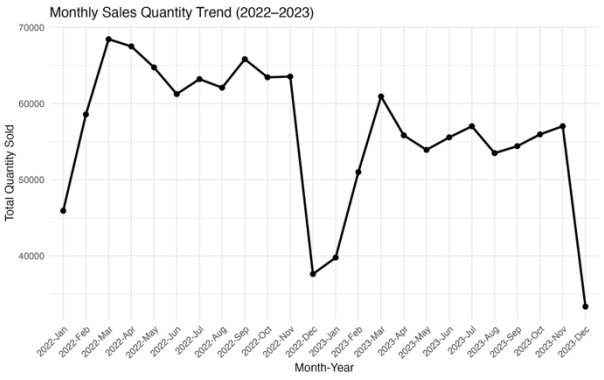


Figure 20: New vs Old Monthly Trend

Both graphs are identical, this is good as confirms 4.3 did not influence the quantities

New

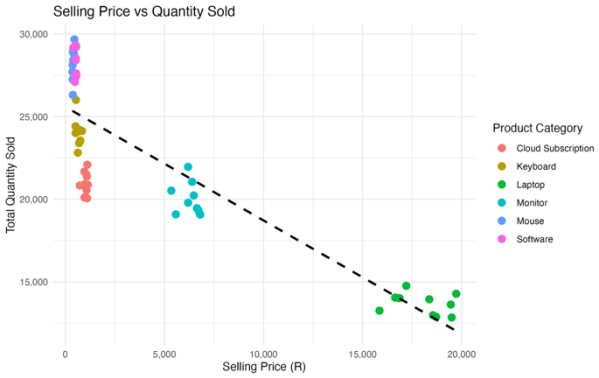
sales_by_year

Year	Total_Sales
2022	2,320,410,018
2023	2,032,177,660

Old

Year	Total_Sales
2022	2,320,410,018
2023	2,032,177,660

New:



Old:

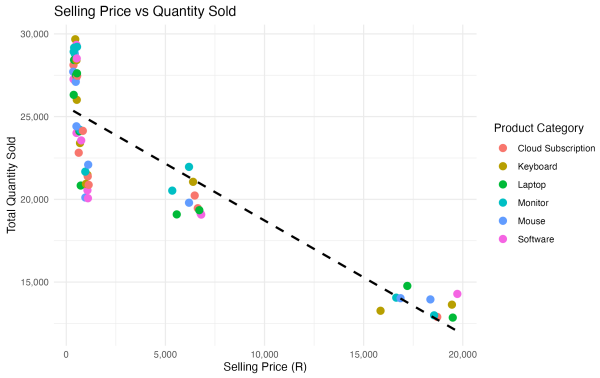


Figure 21: New vs Old Selling Prices

Still shows a negative relationship, as expected.

The new plot look much better as categorical groupings are now closer to each other and not scattered all over graph.

D. Conclusion:

The updated data corrected all category and pricing problems between the files.
This improved the accuracy of data by a huge margin and makes data easier to understand and Analyze.

V. Deliverable Part 5: Optimization

1. Descriptive Analysis

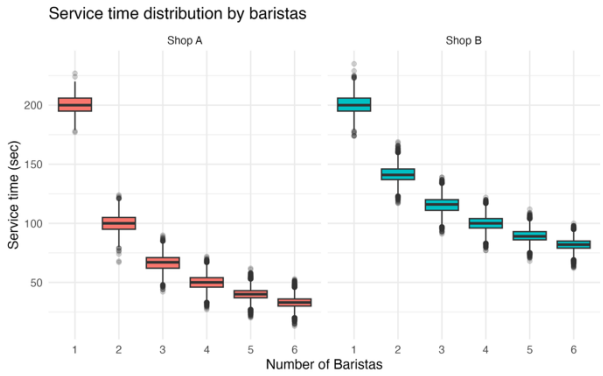


Figure 23: A. Service Time Distributions

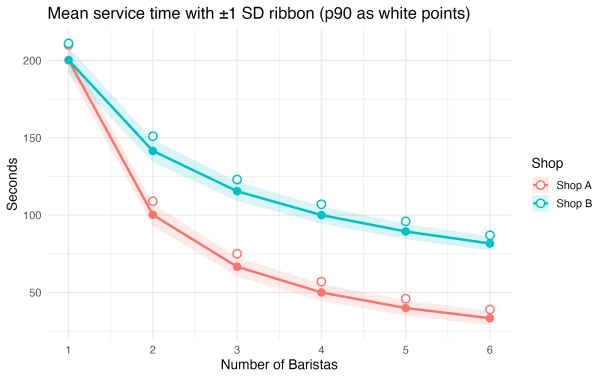


Figure 23: B. Mean Serves Time

Figure 23: A show as number of baristas increases the average service time and variability decreases. Shop A maintains faster service times over all 6 barista counts compared to Shop B; this indicates that Shop B is more efficient in their workflow.

Figure 23:B , both shops maintain a similar negative slope, but shop a achieves a much lower mean over time, this suggests faster throughput even with equal staff members.

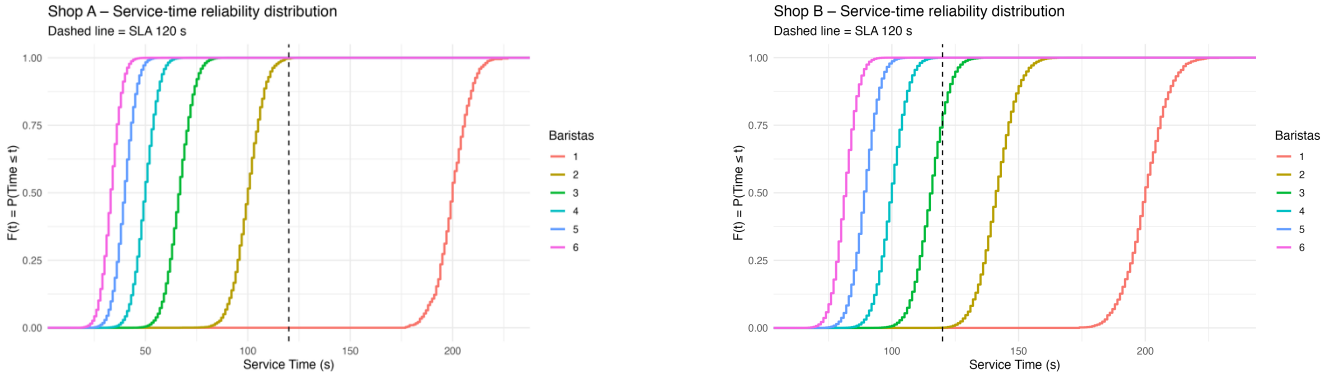


Figure 24: Shop A & B Reliability Curves

The ECDF reliability curves show that shop A reaches 120s SLA from around 4 baristas, while shop B reaches full reliability at 5 - 6 baristas. Shop A tighter spread curves show more reliable serves where Shop B's wider curve is less reliable.

As baristas are added throughput increases, but Shop A serves more customers per day then Shop B. This happens because Shop A has lower average service times in 10h workday compared to Shop B. This reduces number of customers Shop B can serves and reduces revenue and profit, as will be shown in later analysis

throughput_tbl

Shop	Baristas	Mean_Service_s	Customers_per_Day	Reliability_pct	Profit_R
Shop A	1	200.2	180	0	4396
Shop A	2	100.2	359	99.7	8782
Shop A	3	66.6	540	100	13213
Shop A	4	50	720	100	17608
Shop A	5	40	901	100	22026
Shop A	6	33.4	1079	100	26378
Shop B	1	200.2	180	0	4395
Shop B	2	141.5	254	0.1	5632
Shop B	3	115.4	312	79.3	6355
Shop B	4	100	360	100	6798
Shop B	5	89.4	403	100	7076
Shop B	6	81.6	441	100	7228

Table 20: Throughput Table A & B

2. Profit Optimization

A. Combined results:

summary_tbl

Shop	Baristas	n	mean_sec	median_sec	p90_sec	reliability_pct	throughput_raw	revenue_R	personnel_cost_R	profit_R
Shop A	1	417	200.2	200	210	0	179.9	5396	1000	4396
Shop A	2	3556	100.2	100	109	99.7	359.4	10782	2000	8782
Shop A	3	12126	66.6	67	75	100	540.4	16213	3000	13213
Shop A	4	29305	50	50	57	100	720.3	21608	4000	17608
Shop A	5	56701	40	40	46	100	900.9	27026	5000	22026
Shop A	6	97895	33.4	33	39	100	1079.3	32378	6000	26378
Shop B	1	2196	200.2	200	211	0	179.8	5395	1000	4395
Shop B	2	8859	141.5	141	151	0.1	254.4	7632	2000	5632
Shop B	3	19768	115.4	116	123	79.3	311.8	9355	3000	6355
Shop B	4	35289	100	100	107	100	359.9	10798	4000	6798
Shop B	5	54958	89.4	89	96	100	402.5	12076	5000	7076
Shop B	6	78930	81.6	82	87	100	440.9	13228	6000	7228

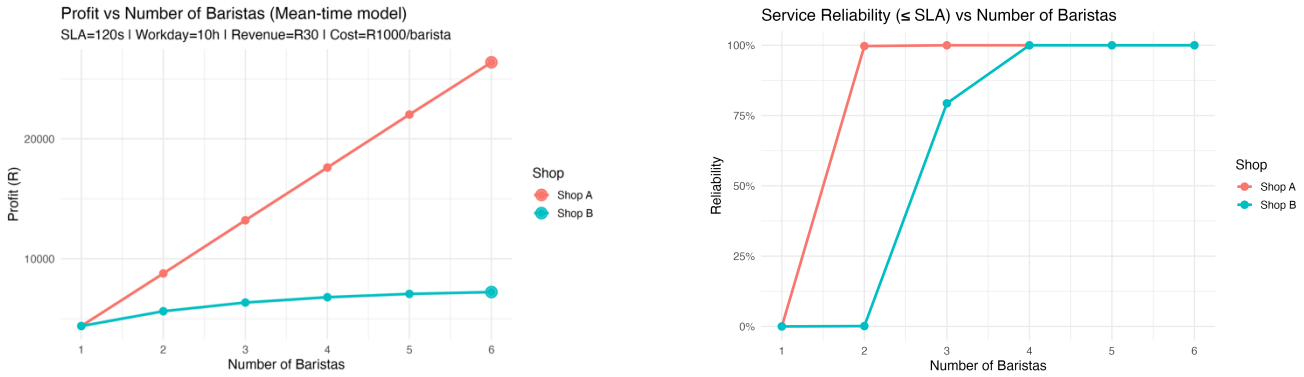


Figure 25: Profit + Reliability Figures

B. Shop A Results:

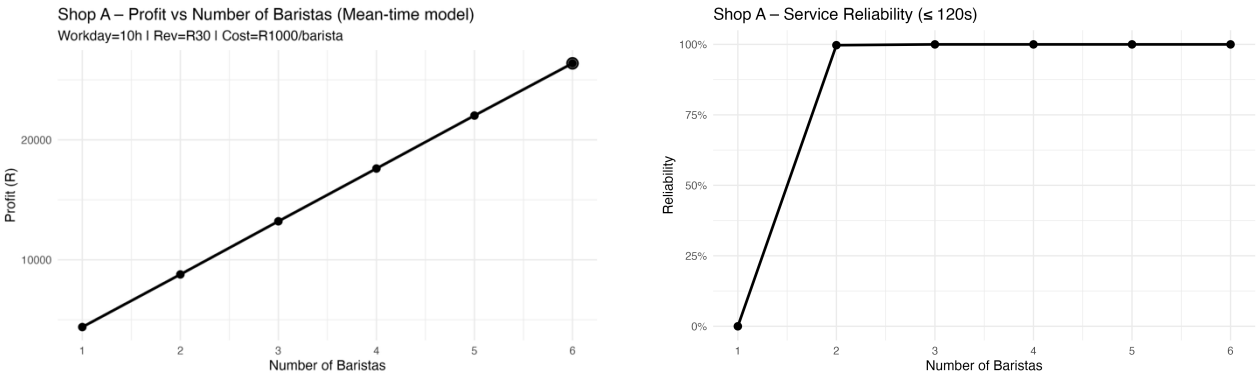


Figure 26: Shop A Profit and Reliability

Shop A's reliability reaches 100% at around 2 baristas, this means that all customers can be served within 120s. The profit does keep rising as the model assumes faster average service times increase total throughput which allows more customers to be served. This suggest that reliability does not need to improve, but by adding baristas profit will rise. With this in mind, to maximize the total profit for Shop A, 6 baristas should be taken. If shop only wants to hit 100% reliability only 2 baristas are required.

C. Shop B Results:

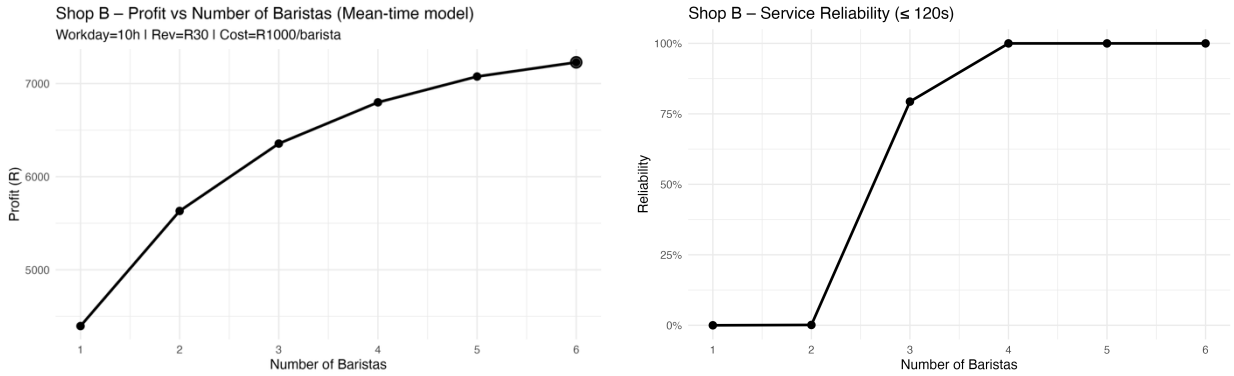


Figure 27: Shop B Profit & Reliability

Shop B reliability improves gradually until it reaches 100% at around 4 baristas.
The profit continues to rise but starts to slow more at around 4 baristas as labor costs start to offset smaller gains. Although its starts to slow profit gains, 6 baristas still allow for more profit as the same happens here ass with Shop A where more staff increases throughput rate. This then allows us to serves more customers increasing profit. For Shop B to maximize profit, 6 Baristas are needed, but if Shop B only wants to be 100% reliable only 4 baristas will be sufficient.

final_rec

Shop	Optimal_Baristas	Profit_R	Reliability_pct	Customers_per_Day	Mean_Service_s
Shop A	6	26378	100	1079	33.4
Shop B	6	7228	100	441	81.6

Table 21: Final Barista Count for Profit

VI. Deliverable Part 6

1. ANOVA Results

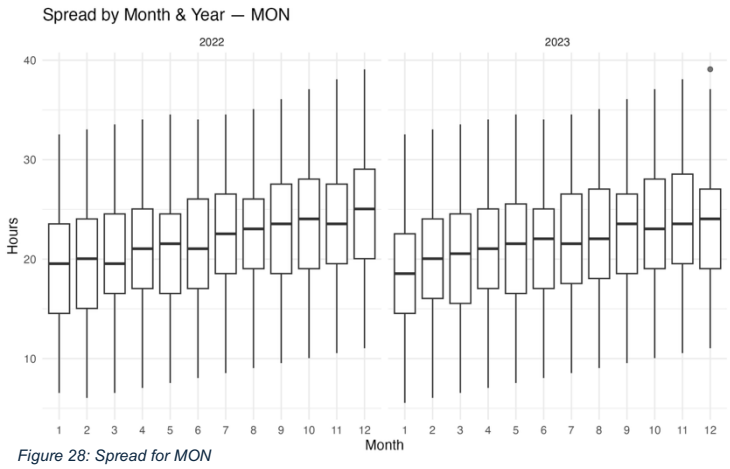


Figure 28: Spread for MON

The boxplots show the distribution of delivery times for each month of years 2022 and 2023. The delivery hours increase throughout the year from around 19h to almost 25h.
This suggests that deliveries become slower as the year progress and can be from higher workloads.

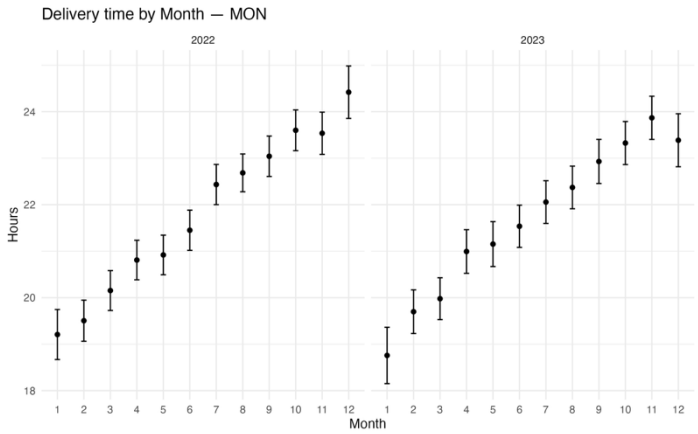


Figure 29: Mean delivery Time

Both years follow a upward trend, indicating that later parts of the year have higher means delivery times. This shows that the months has a significant effect on delivery performances. There is not much change from year 2022 to 2023, except for moths Jan and Dec.

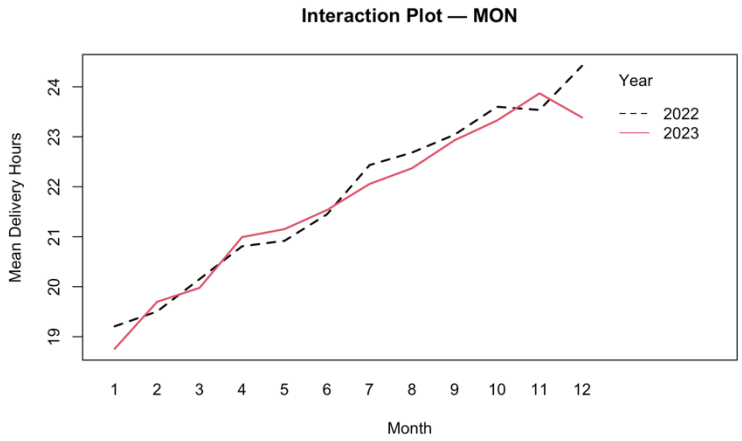


Figure 30: Interception

The graphs show that there is no major difference in delivery times between the 2 years, which suggest no major pattern changes happens between years and seasonal pattern only repeats.

2. Conclusion

Delivery times for product MON show variation on a month-to-month bases, but no major changes happen between years. This points to seasonal operational inefficiency rather than any variation between years. By improving scheduling and better capacity planning for end of the year periods can help to stabilize delivery performances.

VII. Deliverable Part 7

part7_reliability_summary

metric	value
Reliable days (of 397)	366
Empirical reliability	0.92191435768262
Expected reliable days/year	336.498740554156

Table 22: Reliability Summery

This summery shows us that serves was reliable for 92.2% of the recorded days.
When this is scaled to a full year it shows to be reliable for 336 days where the remaining days likely experience services delays or understaffing.

part7_profit_optimisation

N	Prob_Problem	Problem_Days	Annual_Loss	Hire_Cost	Total_Cost
14	1	365	7300000	-9E+05	6400000
15	0.7286	266	5318778	-6E+05	4718778
16	0.389601	142	2844085	-3E+05	2544085
17	0.163769	60	1195512	0	1195512
18	0.057205	21	417596	3E+05	717596
19	0.017273	6	126093	6E+05	726093
20	0.004637	2	33853	9E+05	933853
21	0.00113	0	8249	1200000	1208249
22	0.000254	0	1853	1500000	1501853

Table 23: Profit

This model shows us how reliability and costs change as number of workers are changed. As more staff is introduced the reliability increases to where it almost reaches 100% at around 21 members, but this also increases total cost. However, with around 18 workers the amount of costs is at its lowest while only having 21 problem days.

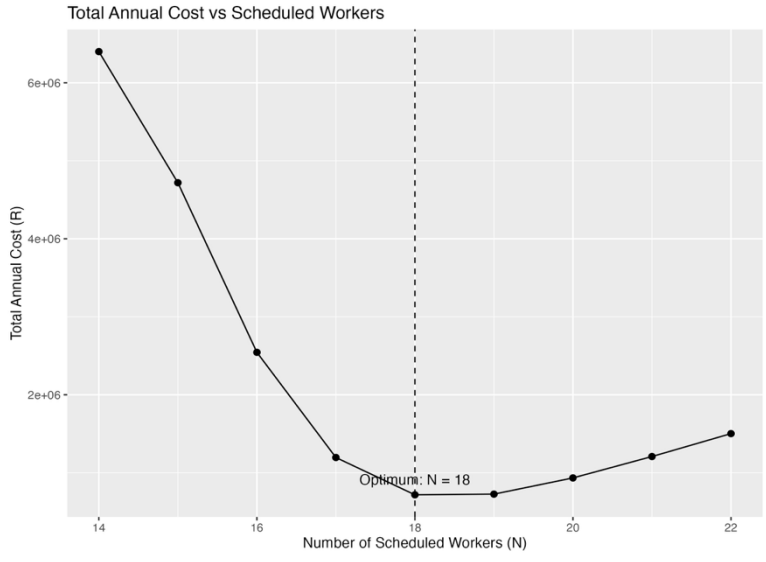


Figure 31: Cost vs

The curve shows how cost decreases as workers are introduced and decreases until it hits 18 workers. After these costs start to rise again. This indicates that **optimal number of workers is 18**. After 18 workers hiring cost take over and only increase total annual costs

VIII. Bibliography

TileStats (2021) *Two-way ANOVA – the basics | interaction | two-way vs one-way*. YouTube, 18 Sep. Available at: https://www.youtube.com/watch?v=9yIobRrZAyg (Accessed: 28 October 2025).

Scribbr (2020) *Two-Way ANOVA: Examples & When to Use It*. Available at: https://www.scribbr.com/… (Accessed: 28 October 2025).