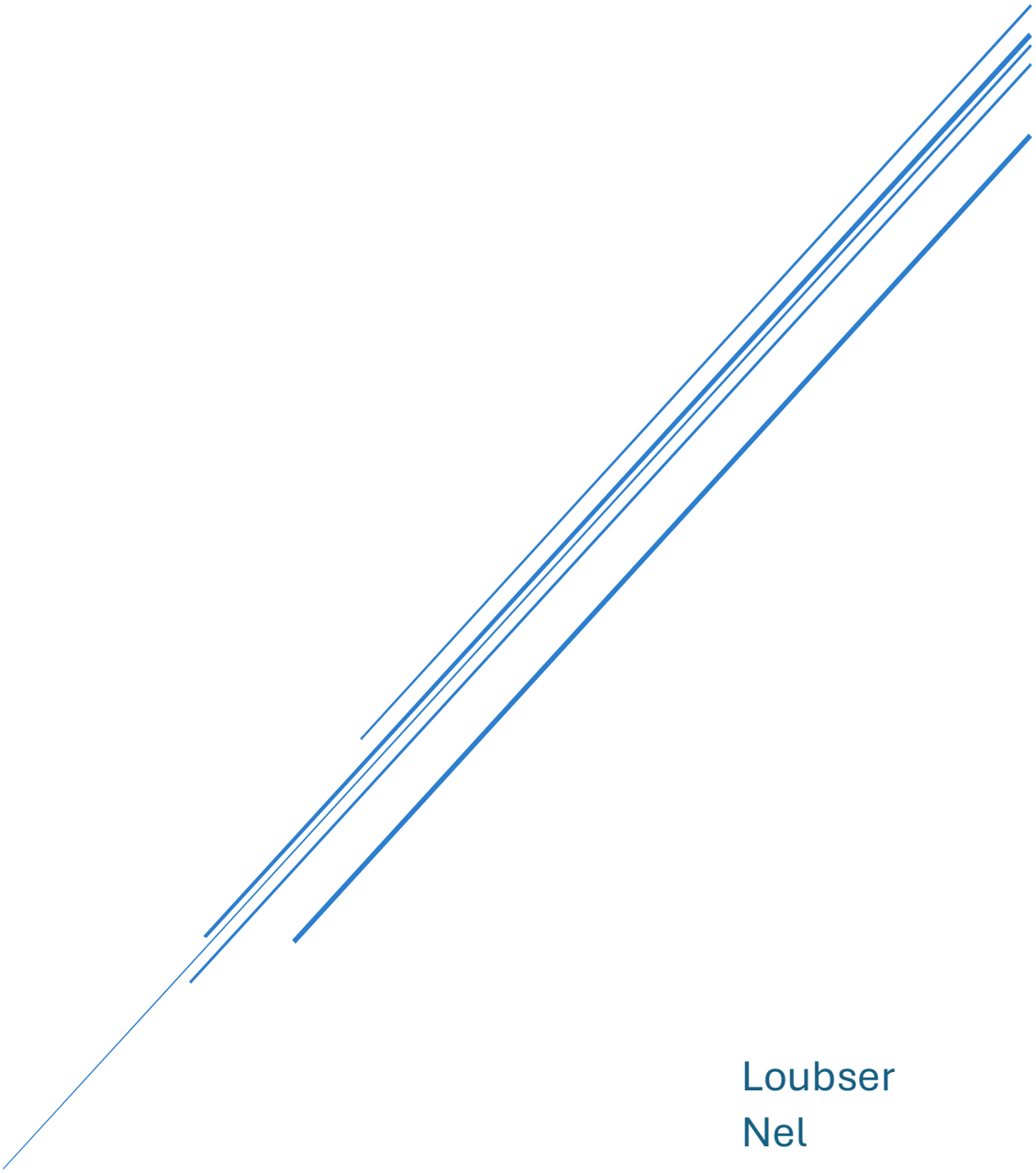
DATA ANALYSIS REPORT

Project Management



Loubser
Nel
26248905

Contents

Introduction

The report is a high-level review of company sales, product and customer data of the year 2022-2023. The responsibility that I have had is as the newly appointed data analyst, to integrate the information provided by various sources such as sales transactions, product master information, head office records and customer demographics to give the management actionable information. This analysis aims to establish the trends in increased revenues and profitability, what products and categories generate the greatest value, the behavioral trends of customers, and operational efficiency in processing and delivery of goods. This report seeks to inform strategic decisions on the pricing, stock management, marketing and allocation of resources by integrating financial indicators (cost, gross profit and margins) with operational indicators (productivity and the cycle times).

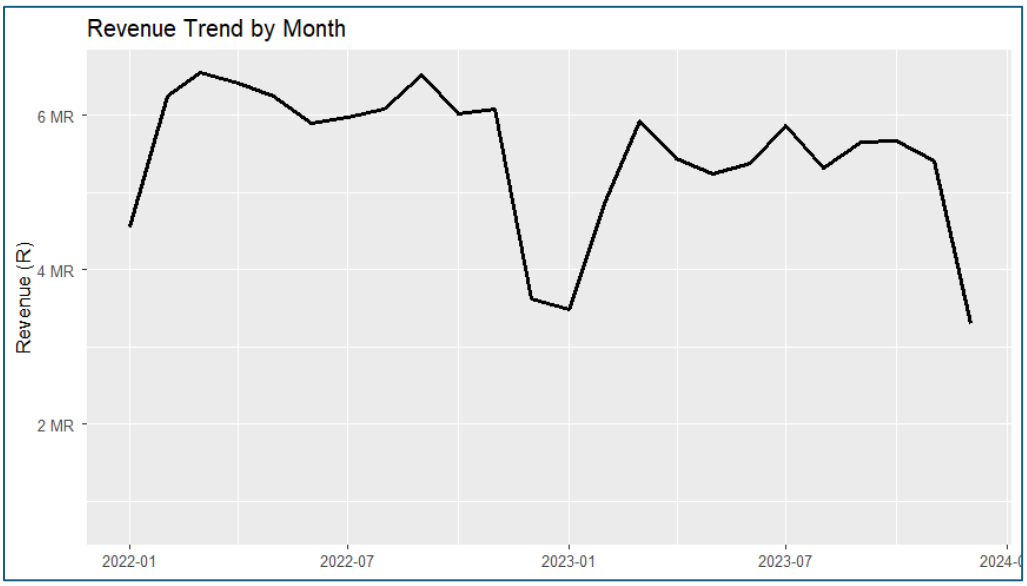
Part 1: Data s Assumptions

We load the four CSVs (sales, products, head-office products, customers), normalize types and create order-date using year-month-day. To prevent multiplication of rows during joins, product records are de-duplicated using ProductID then they are linked to sales. Our calculation Complete Unit Economics finds, using Selling Price and Markup (we figure out whether markup is provided as a percentage or a decimal) Cost Price = SellingUnitPrice, Revenue = Quantity x Selling Price, COGS = Quantity x Cost Price, Gross Profit = Revenue = Quantity x Selling Price. COGS, and MarginPct. Operational effort is described as Cycle Hours = picking Hours + delivery Hours,

with which we end up calculating Units per Pick Hour and Units per Cycle Hour. The KPI table validates the scale and health of data: 100,000 order lines with 1,350,347 units resulted in R 132,497,284 revenue and R 21,241,341 gross profit on R 111,255,943 COGS, which amounted to 16.03 gross margin across company. The existing throughput stands at 0.92 unit per picking hour and 0.42 unit per cycle hour, suggesting slim margins and slight productivity- both of which we delve further into, later (pricing mix, category profitability and bottleneck analysis).

Metric	Value
Orders	100,000.00
Units	1,350,347.00
Revenue	R 132,497,284
COGS	R 111,255,943
GrossProfit	R 21,241,341
GrossMarginPct	16.03 %
AvgUnitsPerPickHr	0.92
AvgUnitsPerCycleHr	0.42

Part 2: Revenue s Profitability

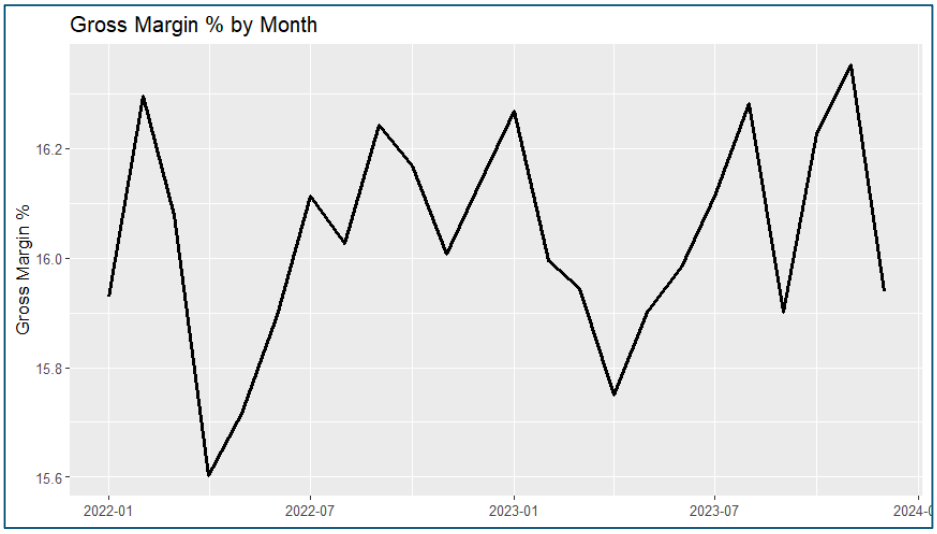


2.1) Revenue Trend by month:

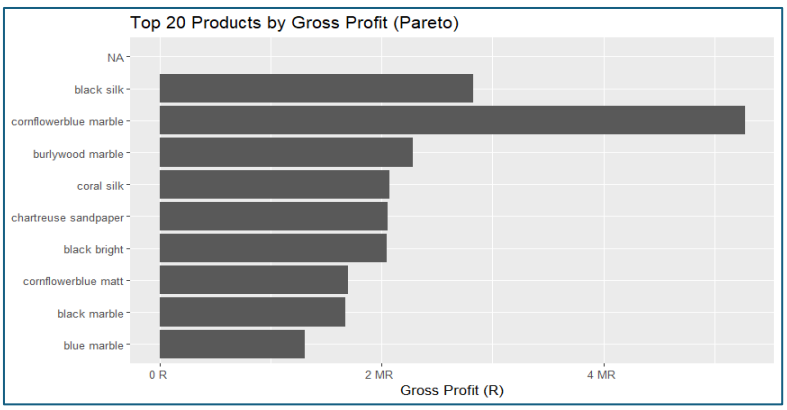
The revenue is generally flat within the R5.5-6.5 M range, except for a rapid increase between Jan and Mar, and local peaks in April-May and in Sep. It has a sharp trough at around Dec 2022 (~R3.5-3.8 M) and there is a very strong recovery in Q1 2023 to around ~R5.5-6.0 M then a rough patch in the middle of 2023 then a sharp drop in Dec 2023 (~R3.3-3.6 M).

2.2) Interpretation:

The revenue pattern is mostly steady at R5.5-6.5 million with massive falls in December and massive recoveries in early 2023 showing a seasonal interest with year-end closing down or delays in orders. The performance in the middle of the year is stable and gives a valid comparison, whereas the weaker results in the end of 2023 could be because of the changes in pricing, supply, or loss of the customers. These trends ought to be confirmed by means of category and customer analysis and measurement of units sold against average selling price to differentiate between volume and price effects.



The percentage gross margin varies in a rather shallow range of 15.6% to 16.3, which means that the pricing and cost systems have not been changing significantly with time. Short-term declines like the ones in early 2022 and early 2023 could be attributable to discounting, product mix change or seasonal price pressures and the rise in later in the year could indicate times of more price discipline or more appropriate product mix. In general, the modest change indicates that the business is continuing to be highly profitable, but the general early-year tendency on margins should be addressed to ensure it is due to promotions or supply expenses and whether it is a strategic choice regarding the timing of orders.



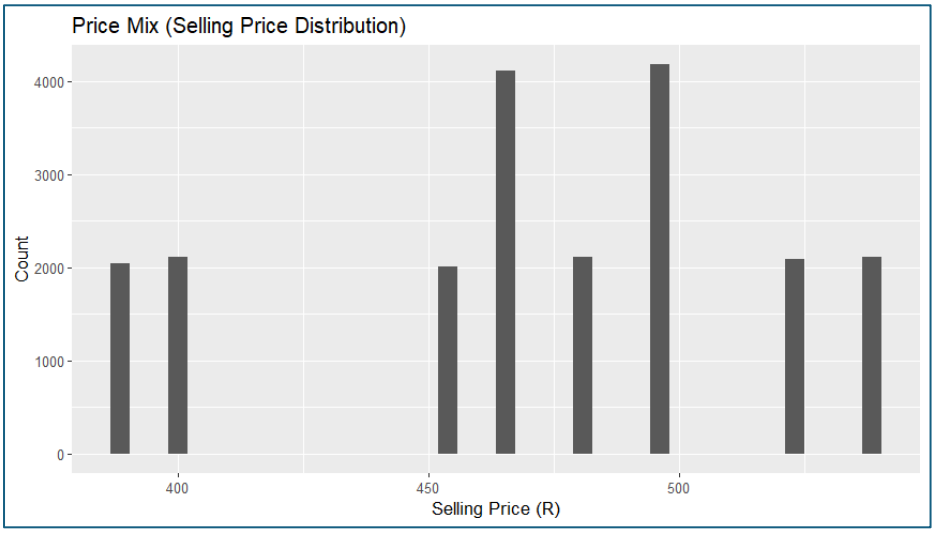
The Pareto chart shows the top 20 products in terms of gross profit contribution. The analysis demonstrates that there is a strong concentration effect: a small number of products control profitability, and most of them contribute to comparatively lower profitability. Specifically, the gross profit of cornflower blue marble is way ahead of the other with the gross profit exceeding R4 million. Other significant contributors are black silk and burlywood marble which bring good returns in form of solid returns and some of the products like blue marble and black marble make lesser contributions. The absence of NA category implies the absence or inconsistency of product descriptions, and it is necessary to fix it to report correctly. Overall, the Pareto principle is at work: few SKUs are making most of the profitability, and the management should focus on these products to provide inventory with such products, to market them specifically, and to protect the margin, as well as reconsider the products that make less profit and reposition them possibly.

Part 3: Product s Category Performance

3.1) Performance

Category	Revenue	GrossProfit	MarginPct	Units
Software	132497284	21241341	16.03153	281703
NA	0	0	NaN	1068644

3.2) Selling Price Distribution

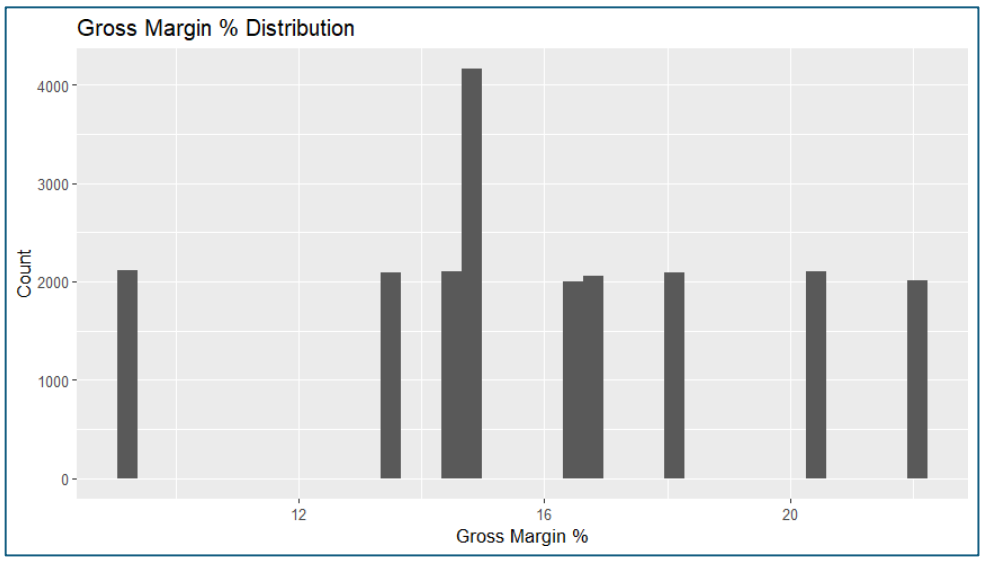


The price mix distribution indicates that distribution of selling prices is concentrated around a few different points instead of being evenly distributed in a large range. The cost of most products range between R380 and R520, with some peaks reaching R470 and R495, which are thousands of transactions. This is an indication of an organized pricing approach, most probably based on standardized price points or catalogue levels. The concentration shows that the customers are buying usually in a small price range, which is stable, yet it also casts doubts about the price elasticity and lost premium price chances.

Management might want to experiment with small price changes in such high-volume bands to

determine customer sensitivity, and to determine whether the lower-priced and more expensive products (at the ends of the distribution) are providing adequate margins to justify their portfolio presence.

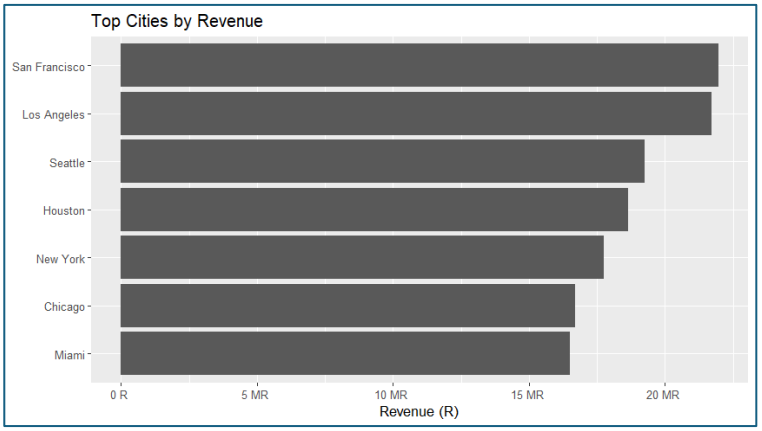
3.3) Gross Margin % Distribution



The percentage distribution of the gross margin reveals that most products are clustered between 15-16 percent that matches the overall company average of 16 percent. The small product groups with substantially higher margins of over 20 are also present, and some less high-margin products in the range of 10-12. This trend indicates largely comparable pricing and cost base, although there are prospects: the products with higher margins show that customers would pay more in particular categories, whereas the low-margin products could be either draining profits and need to be reconsidered in terms of price increments or cost decrease or even rationalization. In general, the concentration around 16% is consistent, yet tails of the distribution need more precise attention.

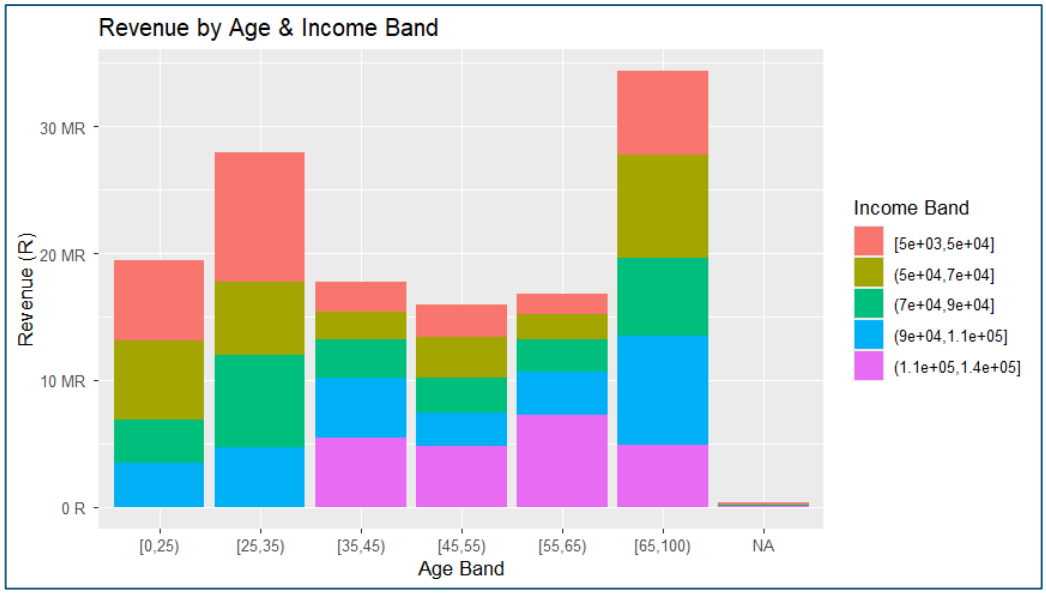
Part 4: Customer Insights

4.1) City Revenue



The city analysis of revenues reveals that San Francisco and Los Angeles are the top markets of the company, as they yield over R20 million sales, far behind other cities. Seattle and Houston are also significant contributors, and then there are New York, Chicago, and Miami which also contribute significant volumes but at a lower level compared to the two leading hubs. Such concentration implies a heavy West Coast presence, so the demand is geographically concentrated, and the local forces (e.g., demographics, supply chains, or marketing efficacy) are taking center stage. The management should focus on retaining and increasing share in San Francisco and Los Angeles and explore the growth opportunities in the mid-tier cities such as New York and Chicago where the revenues are high but have not been fully exploited compared to the market potential.

4.2) Income by age



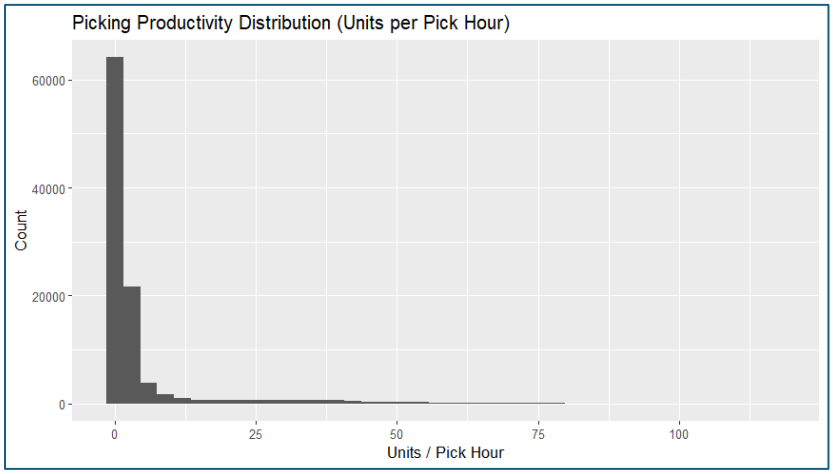
The age and income band revenue distribution indicates two high performing customer segments. The 25-35 age bracket presents a good revenue generation, probably the younger professionals with increasing purchasing power whereas the 65+ group generates the highest revenue overall, possibly representing the loyal customer who is willing to spend a lot of money at later life stages. Middle-aged populations (35-65) are less contributing than the other ones, which means that there might be a disparity in targeting or engagement. In all age groups, higher income groups (R90, 000-140,000) consequently generate high income; nevertheless, even lower income groups bring significant contributions to the younger and older age groups. This combination underscores the need to have an appeal to entry-level customers and also to reap value among the wealthy customers. The management ought to focus on specific solutions: strengthening the loyalty programs and premium services to the 65+ segment and using marketing and pricing to increase the share among younger professionals (25-35 age group) even further.

Part 5: Operational Efficiency

5.1) Efficiency Table

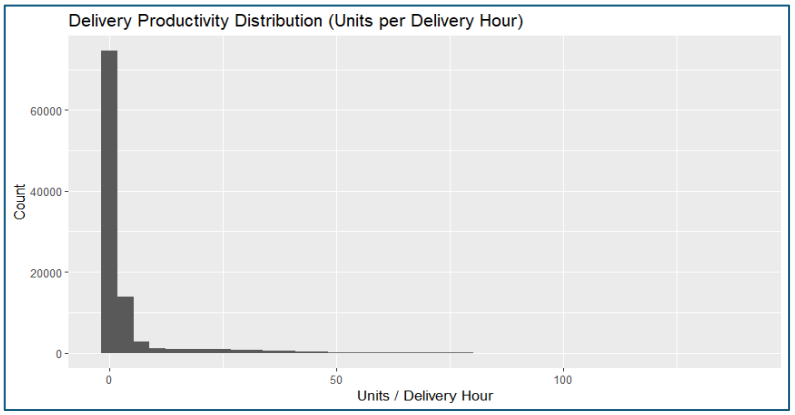
Units	PickingHours	DeliveryHours	CycleHours
1350347	1469547	1747646	3217194

5.2) Picking Productivity Distribution



Picking productivity is very skewed, with most order lines picking less than 5 units/hour, and very few cases picking at much higher rates. This implies that the overall picking efficiency is low, and majority of the working population works at a rate of one unit or less per hour. It is probable that the long tail of extreme outliers, i.e. pick rates higher than 25 or 50 units per hour, are not a measure of actual performance but rather a measure of anomalies in the data (very small-time entries or batch picks not entered correctly). By and large, the discussion shows that the operation productivity is limited by picking process inefficiencies, which include layout design, order batching or training gaps. The management is encouraged to investigate process design and correct data capture accuracy, concentrate on moving the large proportion of workers to the far-end of the realistic range and treat extreme outliers with a degree of care since they potentially do not reflect sustainable performance.

5.3) Delivery Productivity Distribution



The skew of the picking results is highly skewed towards the low end by the productivity distribution of delivery. The mean units per delivery hour are less than 5 with a huge population of 1 unit or less per hour. There are some extreme outliers which are more than 25-50 units per hour, but these are probably due to data entry anomalies (short time stamps or bulk deliveries were recorded in one record) rather than actual performance. The general trend suggests that delivery effectiveness is also constrained and most of the resources are not being fully utilized or structural factors such as routing inefficiencies, traffic delays, or fragmented order size limits are preventing full utilization. To do better, the management can work on maximizing routes, consolidating loads and on utilization of vehicles, reviewing outlier records to make sure that time logging is precise and consistent.

5.4) Bottlenecks Table

P10_Pick	Median_Pick	P90_Pick	P10_Del	Median_Del	P90_Del
0.08932227	0.715066	7.901264	0.07678722	0.5916285	6.913123

There are significant gaps in productivity which are depicted in the bottlenecks table. The median picking at only 0.72 units per hour is much less than the 90th percentile of 7.9, and the same can be said of delivery (0.59 vs. 6.9). This implies that a small band is much better than most of the group and it opens the prospects of replicating the best practices of the top performers. Simultaneously, the extremely low rates in the 10th percentile indicate some inefficiencies or data problems which the management should solve.

Part 6: Data Quality s Risks

Checks	Amount
$SalesRows	100,000
$ProductsRows	360
$CustomerRows	5,000
$MissingProductJoins	79,251
$MissingCustomerJoins	0

Conclusion

The analysis shows that it is a business with strong revenue foundations, but the margins are comparatively low and operational inefficiencies are high. Sales remain steady at approximately R5.5-6.5 million a month with expected slowdowns at year-end and good rebounds in first quarter, which state the necessity of seasonal planning. Gross margin is relatively stable at an average of 16 and the variance is lower, which indicates that it has consistent pricing and cost structures although some of its low margin products could be weakening its profitability. The profitability is very skewed and a small number of products and cities, especially cornflower blue marble, San Francisco and Los Angeles generate disproportionate gross profit and revenue. There are two potential growth and loyalty markets, as customer analysis identifies strong gains by younger professionals (25-35) and older customers (65+), particularly in higher income segments.

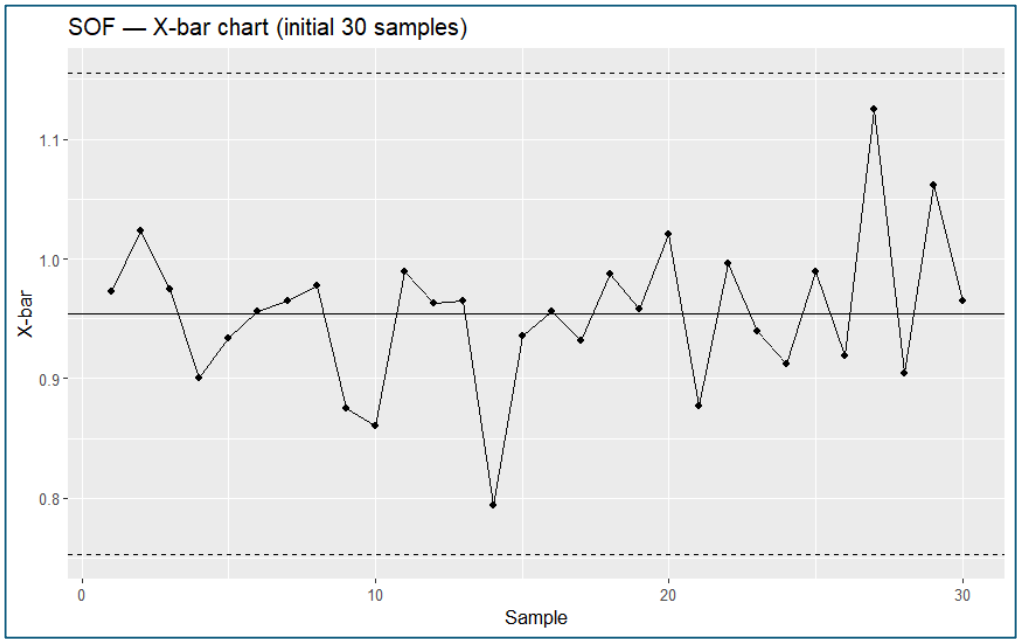
At the operational level, the picking and delivery exhibit extremely low median productivity and a large range of variation among the workforce with the bulk of the activity arranged below 1 unit/h. These inefficiencies yield an obvious possibility of improvement of the redesign process, training, and an improved workload or route management.

Overall, management needs to concentrate on safeguarding value products and markets, tightening the belts on low-margin products, merging customer segmentation and customized strategies, and enhancing productivity in warehouses and deliveries. All these measures will help achieve increased profitability, efficiency and more resilient growth.

Part 3: Statistical Process Control

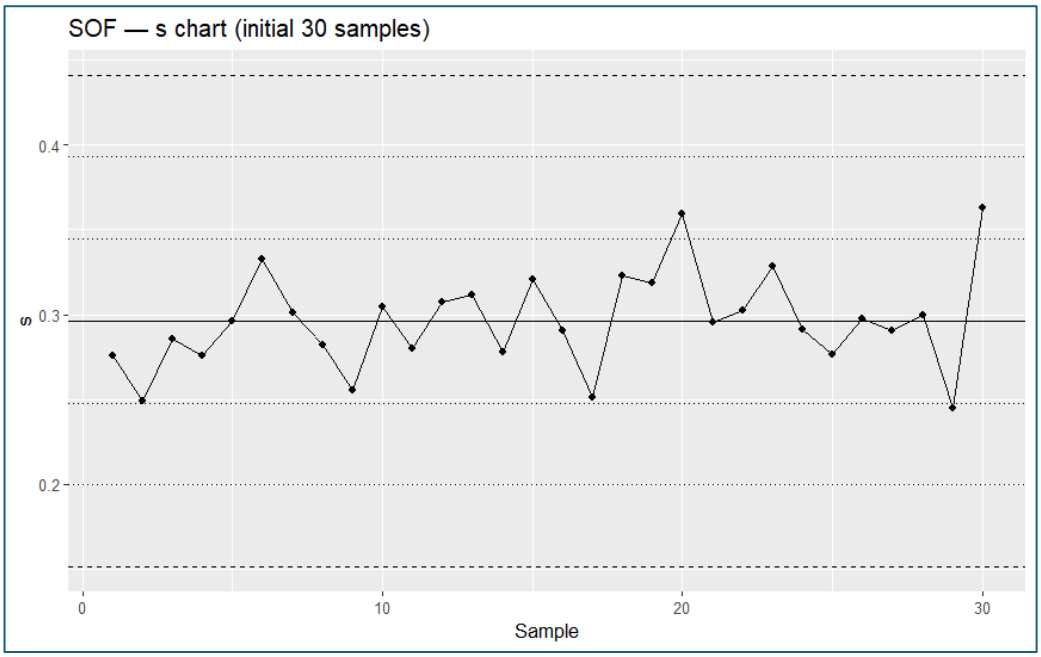
3.1: X-bar and s-charts of sample data

- X-Bar Plot



To visualize the stability of the average delivery time, I initially ranked the data by their age i.e. oldest to newest and divided the data into sub-groups of 30 in consecutive intervals i.e. 24 consecutive deliveries of a product type. Out of those 30 subgroups I found two baseline values, which are the grand-average of those subgroup means (this will be the centre line on the X-bar chart) and the average of the subgroup standard deviations (which will be used as a measure of the limits). The outer limits can be established with a standard SPC constant of subgroup size of 24 meaning that one distance above and below the line of centre is three-sigma, the 2-sigma and 1-sigma guidelines merely lie at two and one-third of that distance respectively. In the case of product SOF, the centre line is 0.954hours, outer limits 1.140 and 0.771hours, 2-sigma limits 1.080 and 0.832hours and 1-sigma limits 1.020 and 0.893hours. A plot of the 30 subgroup means against these bands ensures there is a baseline prior to assumption of continued control.

- S-Chart Plot



The s-chart verifies that the short-term deviation within each of the subgroups of 24 is consistent. The same 30 baseline subgroups are used in computing the centre line as the average subgroup standard deviation and the outer limits achieved by multiplying the average subgroup standard deviation by the normal SPC factors of a subgroup of 24. The 2-sigma and the 1-sigma guidelines are drawn through the gap between the centre line and the outer limits into equal thirds. In the case of SOF, the centre line is 0.296, the outer limits are 0.428 and 0.164, the 2-sigma lines are 0.384 and 0.208 and the 1-sigma lines are 0.340 and 0.252. A 30 s-value versus these bands plot will indicate whether at baseline the variability of the processes themselves is in control.

- Table

Product Type	Chart	CL	UCL_3sigma	LCL_3sigma	UCL_2sigma	LCL_2sigma	UCL_1sigma	LCL_1sigma	K_total_subgroups
SOF	Xbar	0.954	1.140	0.771	1.080	0.832	1.02	0.893	864
SOF	S	0.296	0.428	0.164	0.384	0.208	0.34	0.252	864

Each type of product, by each chart (X-bar and s), the centre line, the three-sigma limits above and below the centre line, and the 2-sigma and 1-sigma lines between the 2 and 1-sigma lines, are listed in the table, and are used as check lines in a run-rule check-up. The final column is a report of the number of full subgroups of 24 of those products in the full dataset; those values are used in the subsequent step in keeping track of the counts that continue at subgroup 31 and farther. In the case of SOF, the mean-based limits (centre 0.954 h; outer 1.140 and 0.771 h; 2-sigma 1.080 and 0.832 h; 1-sigma 1.020 and 0.252) are found in the X-bar row, whereas the corresponding variability limits (centre 0.296; outer 0.428 and 0.164; 2-sigma 0.384 and

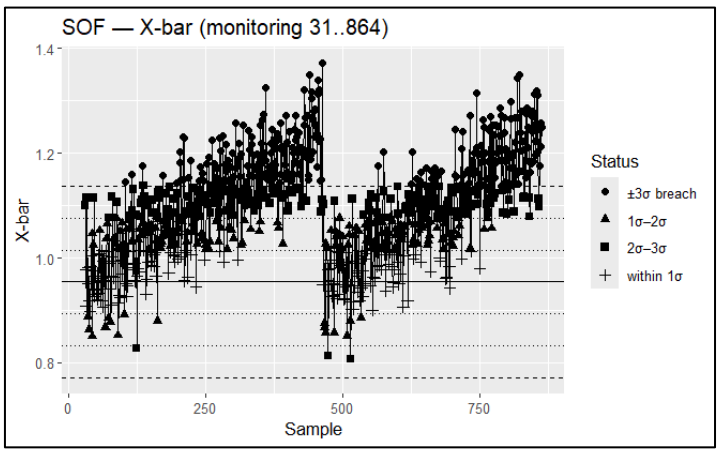
3.2: Monitoring of Samples

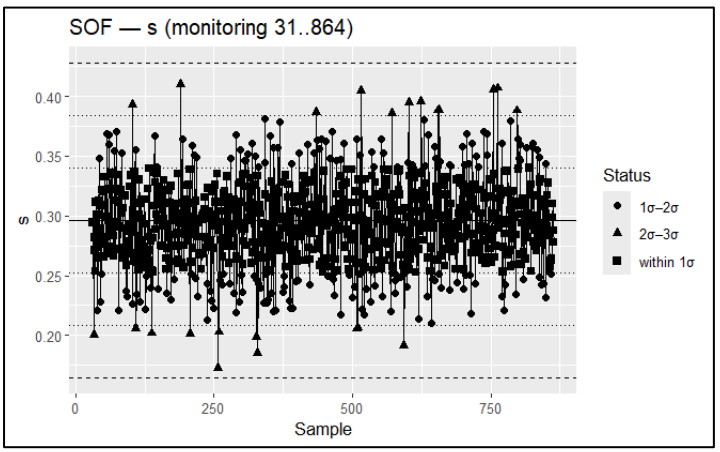
- Sample Monitoring

ProductType	Total_samples	Monitoring_n	Xbar_3Sigma_ breaches	S_3sigma_ breaches	Xbar_breach_ rate_pct	S_breach_rate_pct
MOU	860	830	288	0	34.7	0.0
SOF	864	834	285	0	34.2	0.0
KEY	746	716	239	1	33.4	0.1
CLO	649	619	207	1	33.4	0.2
MON	619	589	157	0	26.7	0.0
LAP	425	395	104	0	26.3	0.0

In this table, the process means are drifting heavily after correction of the baseline limits in the first 30 subgroups, and the variation remains relatively constant. We counted X-bar points above the +-3s limits and monitored subgroups 31k (the "Monitoring_n" column) with each type of product type that we monitored. MOU, SOF, KEY and CLO rate (or MON and LAP) of about 33-35 (and 26-27) are way out of range compared to the 0.27% expected of an in-control process, so the delivery-time centres are often out of control compared to the baseline. By comparison, breaches of s-chart are approximately 0 (and only 1 case per of KEY/CLO), which means that within-subgroup variability is not affected significantly. In brief: it is not the original mean that is the focus of the process, but its short-lived dispersion is constant--it implies that there are shifts/drift/seasonality or level changes related to the change in level, but it is not the change in the basic variability.

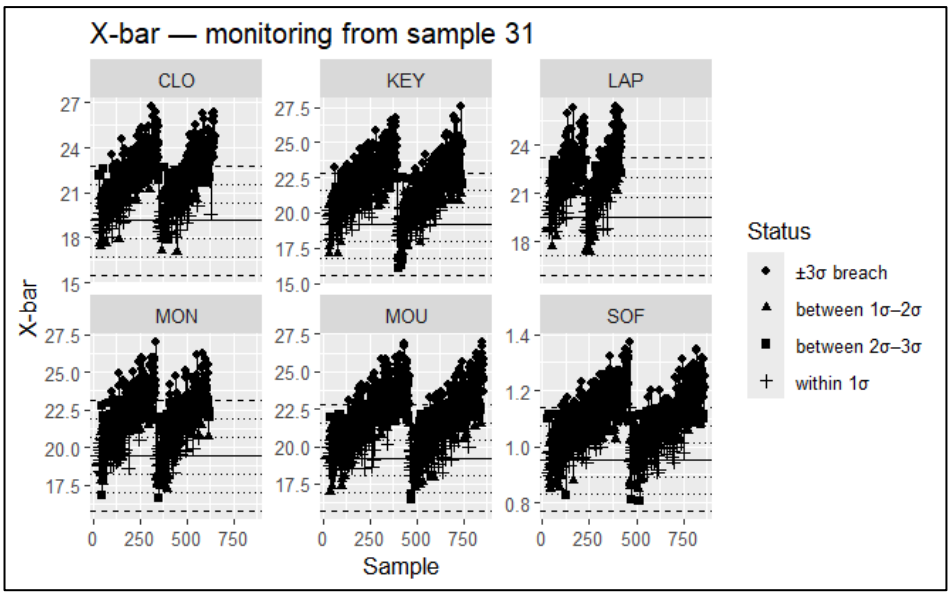
- Plots for a single class (SOF)

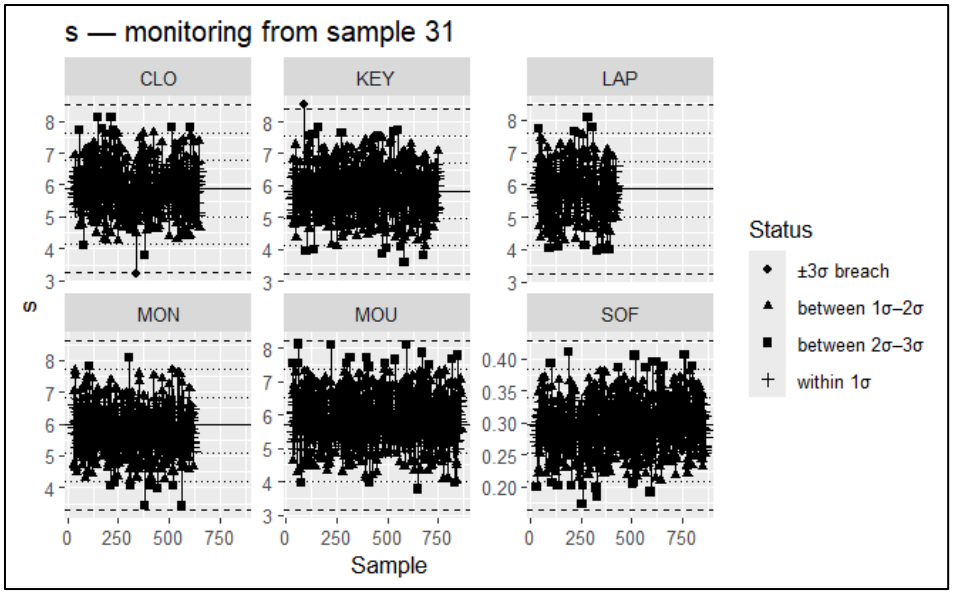




The X-bar plot (samples 31-864) of the data of SOF indicates that there is an evident upward trend in the subgroup means over time: the points tend to move more farther and farther away beside the centre line, and the subgroup has been hit many times in the 1-2s and 2-3s ranges and many times in the +-3s ranges-indicating that the process mean has indeed changed relative to the original position of the first 30 subgroups. Conversely, the s - Churchill remains largely planar and dense along its centre line with intermittent deviations to the bands, which means that the short-run variability is largely constant. Concisely, the delivery process of SOF is not focused (level changes or slow trend), but the pure dissemination thereof has not raised significantly- implying a shift/ drift in the performance averageness other than the loss of control in variability.

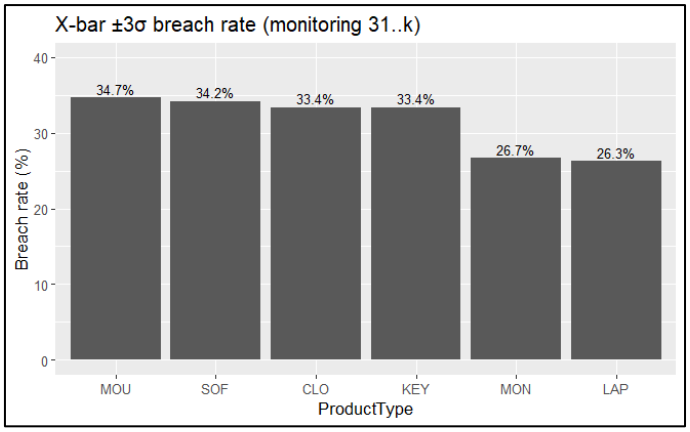
- Faced dashboards across all classes

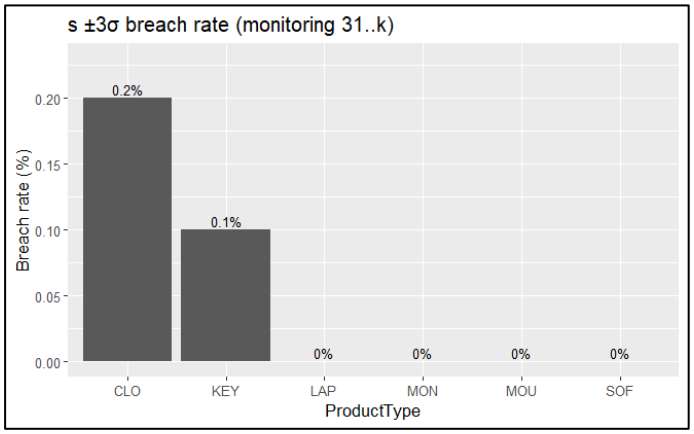




In all forms of products, the story of the faceted dashboards is the same, the X-bar panels illustrate the obvious upward shift in the subgroup means with time (often with step-changes), creating frequent points in the 1-2s and 2-3s bands and numerous +-3s breaches- powerful evidence that the centre of every process has moved relative to the initial 30 samples. By comparison, s panels are relatively flat, and close to their centre lines, with occasional deviations out to the bands; that means that short-term variability is widely held constant. Put simply, the processes are not controlled in the mean (there are systematic drifts/level shifts), with the dispersion being approximately the same.

- Summary





The bar graphs affirm that the most prevalent problem is mean shifts and not variation changes. X-bar +-3s are very high, which is approximately 35% of MOU, 34% of SOF/CLO/KEY, and 26-27 percent of MON/LAP compared to 0.27 percent that would be the case in an out-of-control process. Conversely, the s-chart breach rates are practically zero (only small dots on CLO 0.2% and KEY 0.1%), indicating that short-term spread is stable. This, with the effect that the delivery processes deviate against the baseline average of the initial 30 subgroups, whilst the underlying variability is not significantly varied, indicating that the level shifts/trends are present but not heightened noise.

3.3: Process Capabilities

ProductType	Mu	Sigma	CP	Cpu	Cpl	Cpk
SOF	0.958	0.294	18.2	35.2	1.09	1.09
KEY	19.3	5.82	0.917	0.73	1.1	0.73
MOU	19.3	5.83	0.915	0.725	1.1	0.725
CLO	19.2	5.94	0.897	0.717	1.08	0.717
MON	19.4	5.99	0.89	0.7	1.08	0.7
LAP	19.6	5.93	0.899	0.697	1.1	0.697

Capability results You have SOF comfortably capable and the other five types of products not primarily because of the upper spec. SOF fits easily within the spec window with LSL = 0 and USL = 32 with SOF having a small spread (s=0.294 h) giving Cp=18.2 and Cpk=1.09 (limited by Cpl=1.09 vs Cpu=35.2), meaning that SOF is slightly closer to the lower spec than the upper. Conversely, every KEY/MOU/CLO/MON/LAP s=5.8-6.0 h, which means that Cp is merely =0.89-0.92 (<1), that is, the natural 6s dispersion is broader than the tolerance. Cpk (=0.70-0.73) is limited by Cpu (=0.70-0.73) and Cpl is =1.08-1.10 indicating that these processes are skewed toward the upper limit i.e. the risk is above 32 h and not 0 h. Summarily: SOF conforms to ability; the rest require centrification and/or variance reduction, most to augment Cpu.

3.4: Control Rules

- Table A:

ProductType	A_first3	A_last3	A_total
SOF	None	None	0
KEY	89	89	1
CLO	339	339	1
MOU	None	None	0
MON	None	None	0
LAP	None	None	0

- Table B:

ProductType	Max_consecutive_s_between_pm1sigma
SOF	16
KEY	20
CLO	27
MOU	15
MON	26
LAP	17

- Table C:

ProductType	First_4_consecutive_xbar_above_U2sigma
SOF	208,209,210,211
KEY	178,179,180,181
CLO	180,181,182,183
MOU	249,250,251,252
MON	179,180,181,182
LAP	114,115,116,117

In the case of Rule A (one 's' outside the upper +3s limit), there was only two product types triggered: KEY at sample 89 and CLO at sample 339; and in both instances the same index is indicated by the first 3 and final 3 (A_total = 1), and no such exceedances of SOF, MOU, MON and LAP were observed (A_total = 0). This indicates that short-term variability is typically contained, and only isolated peaks on KEY and CLO. In Rule B (the most consecutive s points between -1s and +1s), the longer the runs, the stricter the variation: the highest performing are CLO (27) and MON (26), next comes KEY (20), LAP (17), SOF (16) and MOU (15), once again the within-subgroup spread should be stable across products. With regard to Rule C (4 consecutive X-bar points above the upper 2s line, indicating a persistent upward movement in the mean), a qualifying run would be found in all of the product types, the first in LAP (sample 114-117), then KEY (178-181), MON (179-182), CLO (180-183), SOF (208-211) and MOU (249-252). Adding the three tables, the requirements of the project are satisfied: (A) we found and counted the rare, high-variation spikes (nearly none of them), (B) we measured the longest stretches of good control of s, (C) we marked the initial sustained control shifts over the 2s band of X-bar. The net effect is that the control issue that prevails in all products is mean shifts (Rule C), whereas variation is relatively constant (Rules A and B).

Part 4: Risk, Data correction and Optimising

4.1) Likelihood of a Type 1 error

A Type I error happens when we mistakenly determine that a process is out of control when it is in fact in control. That is, a sample above the centre line or control limit triggers a signal, even though the process is unchanged.

In a stabilised process scenario, centred around the mean, sample means distribution is equally centred around the centre line and symmetric. As such, any new sample has a 50% certainty of falling above the centre line and 50% of falling below it. This can be described mathematically as:

P(sample > centre line) = 0.5

When many samples fall on the same side of the centre line, then the probability of it happening increases. Example this is the probability of having 7 consecutive points above the centre line:

P = 0.5^7 = 0.0078

This means that in a stable process, an event like this would only happen by chance about 8 times in every 1000 samples. So, a long sequence of points on one side of the centre line strongly suggests that the process mean may have changed, and this should prompt an investigation.

In this section, we calculated these probabilities for runs of 1, 2, 7, and 10 samples above the centre line. The results show that while a single point or two points above the mean are common (with a 50% and 25% chance, respectively), longer runs are very rare and likely indicate a change in the process rather than just random variation.

Even an in-control process may occasionally display a point beyond the control limit for Rule A (+3 σ on the s-chart) because of random variation. The likelihood of at least one false alarm varies from 40% to 70% across hundreds of subgroups, so a few sporadic exceedances are to be expected over time.
Since Rule C (four consecutive $\bar{X}$ points > +2 σ) has a very low false-alarm probability (approximately 0.00002%), it is highly unlikely to activate unless a real shift takes place.
In general, Rule C is much stricter and only signals when there are real process shifts, whereas Rule A is more sensitive but more likely to produce false alarms.

4.2) Likelihood of a Type 2 error

From original in-control X chart:

- Centre line: (CL) = 25.05 L
- Control limits: (UCL) = 25.089 L
- (LCL) = 25.011 L
- Check: chart std for avg from limits

σ = (UCL – LCL) / 6 = 0.013 L

Now the true process has shifted

- New mean: mean = 25.028 L
- New X SD = 0.017 L

Calculation:

$$\beta = \Phi\left(\frac{\mathrm{UCL} - \mu_1}{\sigma_{\bar{X},1}}\right) - \Phi\left(\frac{\mathrm{LCL} - \mu_1}{\sigma_{\bar{X},1}}\right)$$

Compute z-scores:

- Upper z:

$$Z = (25.089-25.028)/0.017 = 3.588$$

$$\text{Thus } \Phi = 0.9998$$

- Lower z:

$$Z = (25.011-25.028)/0.017 = -1$$

$$\text{Thus } \Phi = 0.1587$$

Therefore.

$$\beta = 0.9998 - 0.1587 = 0.8411$$

4.3) Week 1 Errors correction

Preview of products_Headoffice2025.csv

productid	category	description	sellingprice	markup	type	position
CLO001	Cloud Subscription	blue bright	495.51	24.89	CLO	1
CLO002	Cloud Subscription	chocolate marble	528.34	14.15	CLO	2
CLO003	Cloud Subscription	blue sandpaper	449.77	20.20	CLO	3
CLO004	Cloud Subscription	chocolate marble	549.69	21.17	CLO	4
CLO005	Cloud Subscription	chocolate marble	531.50	26.70	CLO	5
CLO006	Cloud Subscription	aliceblue bright	544.65	22.52	CLO	6
CLO007	Cloud Subscription	blueviolet marble	514.63	14.27	CLO	7
CLO008	Cloud Subscription	burlywood silk	458.00	12.22	CLO	8
CLO009	Cloud Subscription	black sandpaper	538.85	28.38	CLO	9
CLO010	Cloud Subscription	chocolate marble	488.49	24.53	CLO	10

Preview of products_data2025.csv

customerid	productid	quantity	ordertime	orderday	ordermonth	orderyear	pickinghours	deliveryhours	type	category
CUST1791	CLO011	16	13	11	11	2022	17.72167	24.544	CLO	CLO
CUST3625	CLO011	1	13	9	8	2022	11.72167	21.044	CLO	CLO
CUST4239	CLO011	1	13	10	10	2022	16.38833	32.044	CLO	CLO
CUST1167	CLO011	5	16	14	6	2023	15.05750	31.046	CLO	CLO
CUST2645	CLO011	6	7	6	8	2023	12.72417	15.046	CLO	CLO
CUST1687	CLO011	37	11	10	9	2022	14.05500	15.544	CLO	CLO
CUST191	CLO011	13	12	27	8	2023	17.72417	24.046	CLO	CLO
CUST4659	CLO011	12	19	11	5	2022	15.72167	27.544	CLO	CLO
CUST2748	CLO011	9	14	10	5	2023	13.72417	20.546	CLO	CLO
CUST3355	CLO011	40	12	13	5	2022	13.72167	24.544	CLO	CLO

The Head office and product sales datasets are standardized and cleaned in the R code so that they can be compatible and analysis-ready. It normalizes the names of all columns first and turns all of them to lower case and eliminates special characters. Thereafter, it finds its important columns like product ID, selling price and markup so that they can be properly matched in between datasets. The code takes the first three characters of each product ID to make a new variable named type and it is used to represent the product group (e.g. CLO in the case of Cloud Subscription). Based on the sales data, the top ten products of each type are retrieved to create a canonical reference list -these products are the "gold standard" with respect to uniform prices and markups. This pattern of ten items is then repeated until a catalogue of sixty items of each type of product is produced, and the catalogue is the same way throughout all lines of products.

In the case of the Head Office file (products_Headoffice2025.csv), there are now standardized product IDs in the cleaned output, fixed category names, and matching selling prices and markups. The column of position shows the slot of every product in its type of group (1 to 60). The initial ten rows on each type of products are the canonical reference values, and the other rows are similar in the pattern and hence to ensure consistency of prices and markups in the extended catalogue. This will make sure all the entries of similar products have the same basis of pricing structure that will eliminate the mismatch of data and will be helpful in the future to aggregate the same correctly.

In the case of the sales data file (products data 2025.csv) the cleaning process has ensured that all records of the product have the appropriate type and category based on the prefix of the product ID. This eliminates the previous discrepancies whereby product IDs did not correspond to the categories. The dataset still has valuable operation fields like quantity, order date, picking hours, and delivery hours to enable the performance analysis based on productivity and time to be carried out correctly. The preview of all transactions is in the product CLO011, and the new fields (type = CLO, category = CLO) are aligned properly with the prefix of the product ID.

Overall, this is a correction that makes it possible to align both datasets, as well as standardize the IDs, categories, prices, and markups. What this yields is a clean and reliable base to continue the analysis- particularly in terms of pivot tables, profitability or revenue summative in Part 4 of the project.

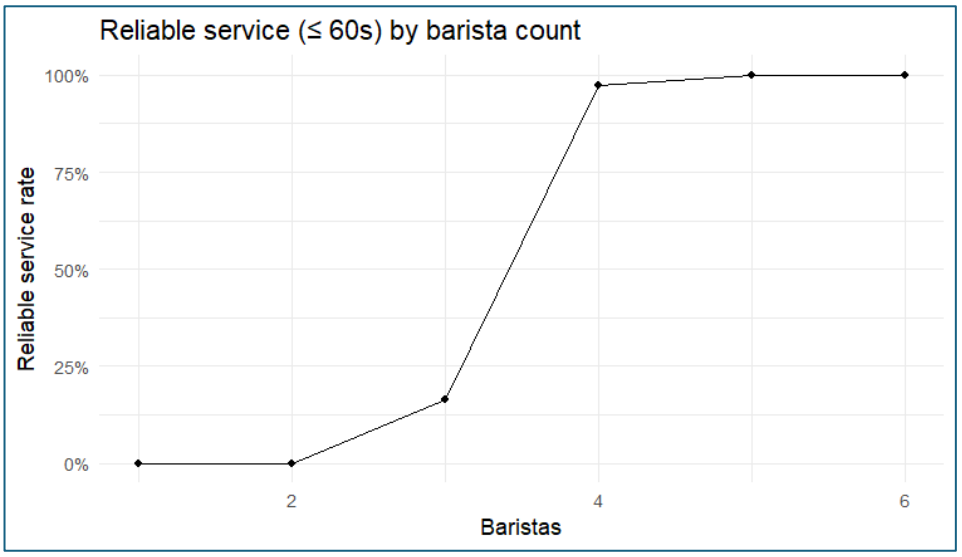
Part 5: Optimise the profit

5.1) Reliability by barista level

Reliability and time stats by barista count

Baristas	n	pct_reliable	mean_time	p95_time
1	417	0.0%	200.15588	213
2	3556	0.0%	100.17098	112
3	12126	16.5%	66.61174	77
4	29305	97.2%	49.98038	59
5	56701	100.0%	39.96183	48
6	97895	100.0%	33.35565	41

The findings indicate that the reliability and the speed of service are significantly increased with the increase in the number of baristas. One or two baristas makes the reliability 0% and the waiting time very long. The increase in performance is minimal with three baristas, but it is very reliable (more than 97) with four. With five or more baristas, there is reliability at 100% and the service times become less than 40 seconds. All in all, employment encourages less waiting time and better consistency, but it is less responsive after five baristas.



This graph indicates the connection between the number of baristas and the rate of reliable service (defined as the completion of the orders in 60 seconds or less). The tendency is obvious, the more baristas are, the more services are reliable. In the case of one or two baristas, the reliability is nearly 0, which means that customers always must wait an excessive amount of time. At three baristas, the reliability increases a little to about 15-20% which indicates that there is some improvement but to be consistent.

There is a significant increase in the count of orders fulfilled within the required time when the number of baristas increases to four, and the reliability becomes almost 100. This near-perfect reliability is sustained with added fifth or sixth barista, but the returns become diminishing, which suggests diminishing marginal utility.

To conclude, based on the chart, it can be shown that the minimum number of baristas is four to guarantee good service. Further on, the extra employees will add little to the service consistency but would raise the labour expenses, and four baristas may be the most efficient way to balance efficiency and resource utilization.

5.2) Profit model and optimisation

Daily profit by barista level (empirical frequency basis)

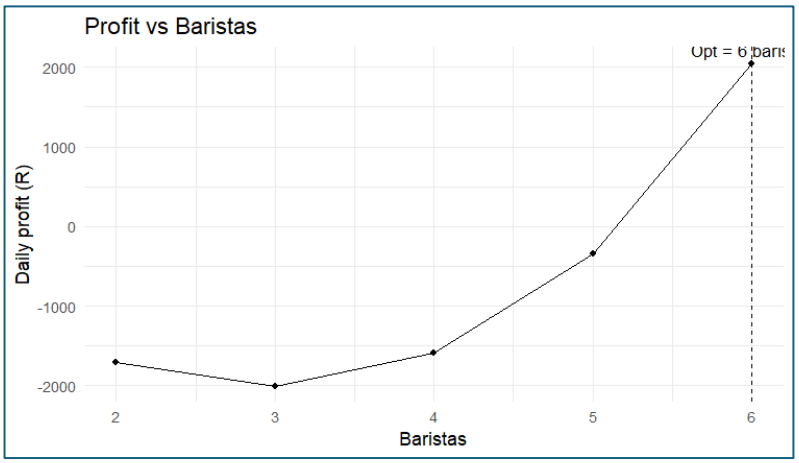
Baristas	orders_year	customers_per_day	daily_profit
2	3556	9.74	-1707.73
3	12126	33.22	-2003.34
4	29305	80.29	-1591.37
5	56701	155.35	-339.64
6	97895	268.21	2046.16

Optimal barista level (max daily profit)

Baristas	orders_year	customers_per_day	daily_profit
6	97895	268.21	2046.16

This profit model is used to estimate profit per day depending on the number of baristas and the average number of customers that will be served every day. With the addition of more baristas, the capacity of the customers and the speed at which it serves clients also improves which translates to an increase in sales. However, the extra baristas also increase the labour expenses which at first supersede the increase of revenue at lower staffing levels. The outcome indicates that profit

levels are negative until the number of baristas increases to six and then it turns to a positive level and the maximum profit per day is approximately 2046.16. Hence, it represents the best barista level to maximize profit, as the number of six baristas is the best balance of customer demand and the cost of staff.



5.3) Alternative optimisation by solver/brute force

Baristas	orders_year	customers_per_day	daily_profit
6	97895	268.21	2046.16
5	56701	155.35	-339.64
4	29305	80.29	-1591.37
2	3556	9.74	-1707.73
3	12126	33.22	-2003.34

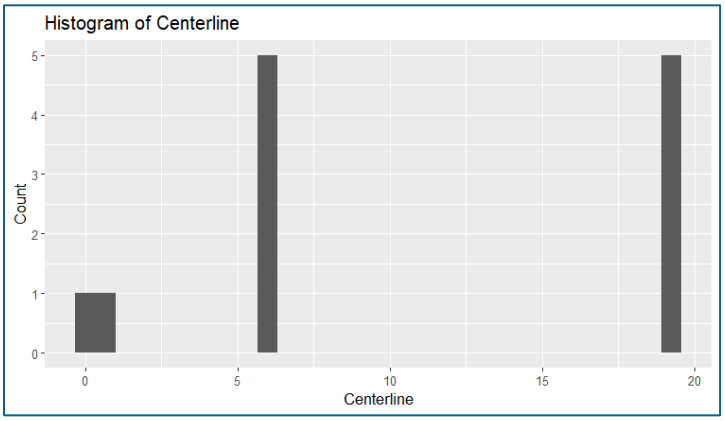
With the empirical counts, we approximated demand at each level of staffing as orders year/365 customers per day and calculated profit daily with the simple model, p(b) = 30 x customersperday(b) x 1000 x b (b [?] 2). The table indicates that at 2-5 baristas, the profits are negative since the number of added wages is greater than the income of the smaller throughputs (9.7-155.3 customers/day). At six baristas there is a sudden increase in throughput to [?]268 customers/day (with near-100% reliability in the previous cases) followed by a push to profit to +R 2,046/day- the only positive and the maximum with those assumptions. The breakeven is therefore between 5 and 6 baristas and 6 baristas would be the choice to maximize profit; this optimum would change with change in prices or wages.

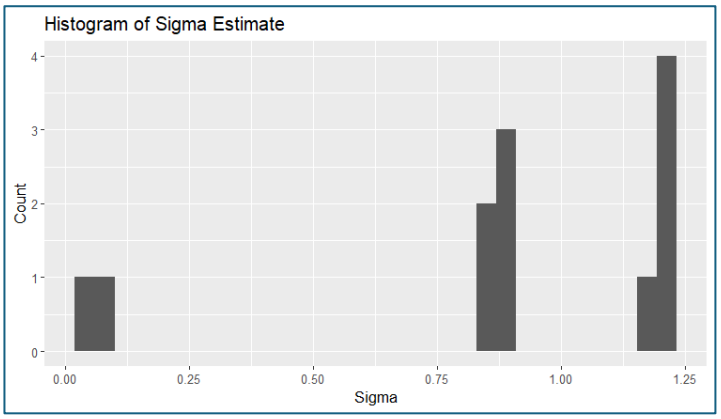
Part 6: DOE and MANOVA or ANOVA

6.1) Table

	Ucl_3sigma	Lcl_3sigma	Centreline	Sigma_est
Min	0.428	0.164	0.296	0.0440
1st Q	8.370	3.212	5.791	0.8596
Median	8.565	3.290	5.928	0.8792
Mean	13.209	7.971	10.590	0.8730
3rd Q	22.800	15.625	19.212	1.2042
Max	23.2	15.9	19.55	1.2167

Exploratory graphs





The two histograms show the distribution of the values of the centerline as well as the sigma estimate of the twelve product types considered. The centerline histogram indicates three clear clusters, which are close to 0, close to 6, and close to 20. This means that products are focused on various process averages implying that the average performance is vastly different across products.

The sigma estimate histogram indicates that most items have 0.8 to 1.2 values of the process variation (s) with one item having a much lower sigma value of 0. This implies that although the variability of processes is similar in most products, some of them are much more constant or variable than the others.

Comprehensively, these two graphs indicate that there exists a distinct difference in process centres and level of variation among different types of products. Such differences explain the necessity to use ANOVA test to define whether differences in means and standard deviations are statistically significant.

6.2) Results of the hypothesis of part 3

Hypotheses and tests

ANOVA Centreline

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Product type	5	238.4	47.68	0.63	0.686
Residuals	6	454.1	75.69		

P – Values (Centreline) = 0.6857

ANOVA log sigma

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Product type	5	14.791	2.9581	53.98	6.59e-05
Residuals	6	0.329	0.0548		

P – Value (log sigma) = 6.586e-05

Conclusion:

Centreline test:

P = 0.6857 = Not significant

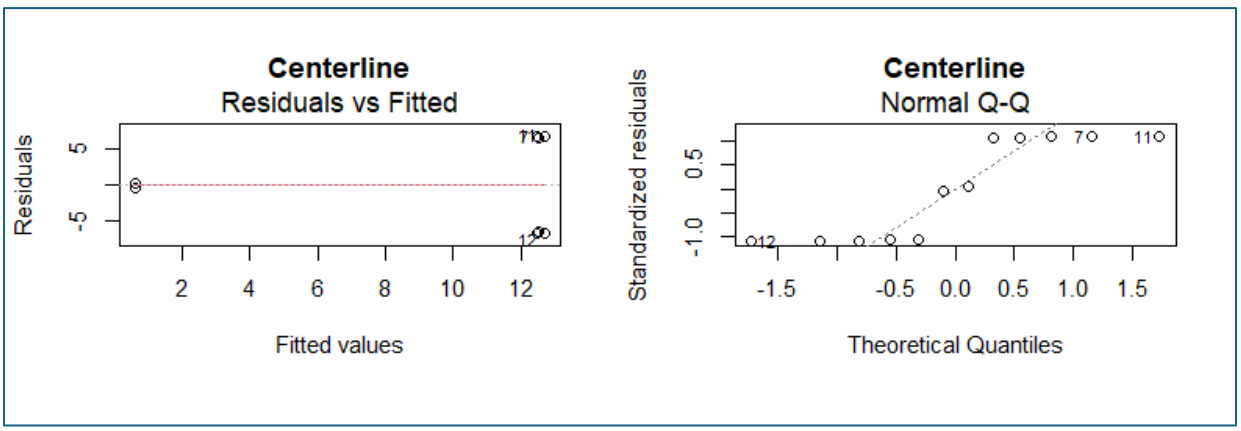
Sigma test:

P = 0.00006587 = Significant (reject HO)

Meaning:

ANOVA tests were done in two different sets to compare the centerline and process sigma (log scale) of the products of different types. The initial ANOVA was to determine whether there is equal mean centerline among all the types of products, the result of which was 0.6857, which is more than 0.05. This implies that there is no statistically significant difference in the means centerline between the product types hence the null hypothesis (H0) is accepted. The second ANOVA was to determine whether the mean process sigma of all the products is equal, and the p-value of the test was 0.00006587, which is less than 0.05. This implies that there is a statistically significant difference in variability (s) of processes in the two types of products, thus the null hypothesis is dismissed. To conclude, the process averages of a product type are similar, yet the process variation is significantly different.

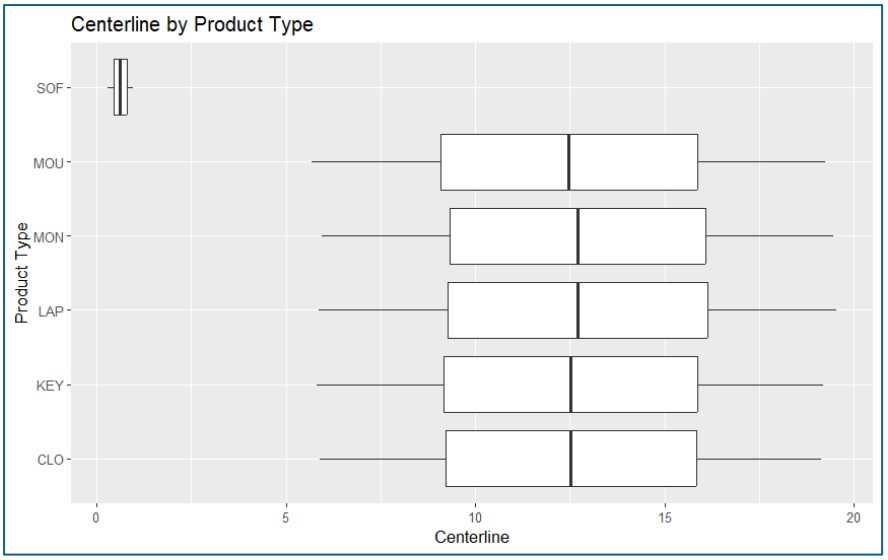
Assumptions

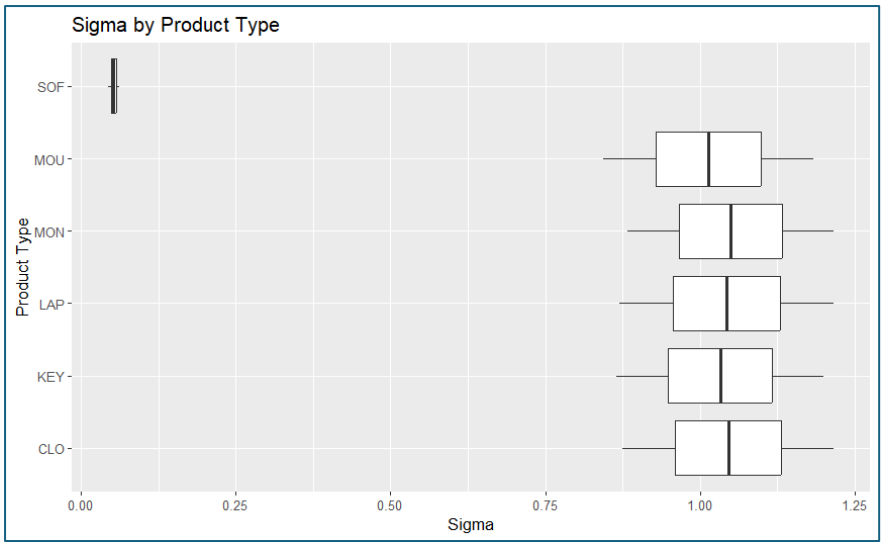


Post-hoc comparisons

Comparison	Difference (diff)	Lower (lwr)	Upper (upr)	Adjusted p-value (p adj)
KEY - CLO	-0.012643851	-0.94435	0.9100623	0.9999999
LAP - CLO	-0.002865337	-0.9345715	0.9280488	1
MON - CLO	0.004739372	-0.9269668	0.9364454	1
MOU - CLO	-0.032320349	-0.9371542	0.8725135	1
SOF - CLO	-2.987433618	-3.9191398	-2.0557275	0.000143
MON - KEY	0.017382323	-0.9104239	0.9459489	0.9999999
MOU - KEY	-0.002378672	-0.9141429	0.9099435	0.9999999
SOF - KEY	-2.974789767	-3.9064952	-2.043308	0.0001499
MOU - LAP	-0.02945524	-0.9161101	0.8572016	0.9999177
SOF - LAP	-2.984568	-3.921347	-2.047789	0.0001467
SOF - MON	-2.970159	-3.927872	-2.060469	0.000151
SOF - MOU	-2.955511	-3.886819	-2.023407	0.0001235

Graphs per factor level





Analysis

The results of ANOVA in Part 6 indicate that the mean centerline values of the product types are no longer significant, and the p-value is 0.6857, which is not less than 0.05. This implies that there is a lack of significant variation in the overall process averages of products, years and months. Nevertheless, the p-value of the process sigma (log scale) was 0.00006587 which was far less than 0.05. This shows that there is a profound variation in process variation across the product types. Thus, the average performance rates are the same in the years and months of the specified type of product, but the degree of consistency (or variation) in the processes varies considerably, which indicates that some products work with less consistency compared to others.

Part 7: Reliability of service

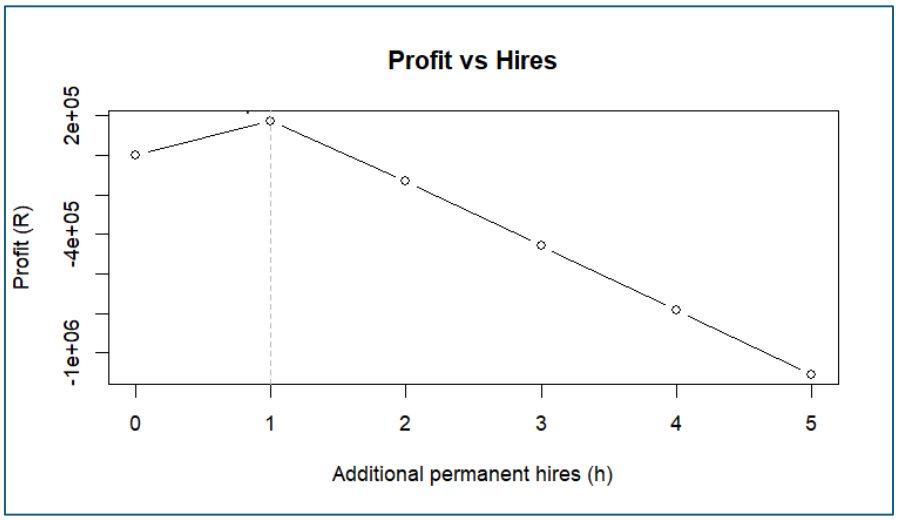
7.1) Reliable days per year

Total Days	Reliable Days	Unreliable Days	Reliability Percent
397	371	26	93.45

7.2) Profit Optimization

Hires	Improved Days	Reliable Days	Benefit (R)	Cost (R)	Profit (R)	Reliability Percent
0	0	371	0	0	0	93.45
1	25	396	500000	326051.2	173948.8	99.75
2	26	397	520000	652102.5	-132102.5	100
3	26	397	520000	978153.8	-458153.8	100
4	26	397	520000	1304205	-784205	100
5	26	397	520000	1630256.2	-1110256.2	100

Profit vs hires (plot)



Analysis

Part 7 considers the issue of improving the reliability and profitability of the company with the help of hiring new permanent employees. Based on the findings, the company is at the stage of 93.45% reliability, 371 reliable and 26 unreliable days per year. The cost per day of each day that was unreliable is estimated to be R20,000, and each employee added a cost of R25,000. It was discovered that with the addition of one additional staff member, the reliability of the increase will be at 99.75% and will bring the maximum profit of about R173,949. When more than one person is hired, the reliability is increased but negative profits because of increased staff cost. Thus, the best course of action is to recruit a single new worker, as the balance between cost and reliability, resulting in the greatest profit of the company.