ECSA PROJECT

De Souza, S, Mr [26875039@sun.ac.za]

QUALITY ASSURANCE

Contents

Introduction

The purpose of this report is to demonstrate data analysis, quality control, and process optimisation skills required by the Engineering Council of South Africa (ECSA) Graduate Attribute 4 (GA4). The report integrates statistical, computational, and engineering techniques to evaluate the quality, consistency, and performance of operational processes based on various provided datasets. These data sets include head office product information, customer information, product sales, delivery time, and service reliability records by various classes of product and time frames.

With the assistance of R programming, data were cleansed, joined, and analysed to identify patterns, trends, and potential inconsistencies. Standard summary descriptive statistics were used to uncover the structure and variability of the data sets, and control charts ($\bar{x}$–s charts) to monitor process stability and signal out-of-control conditions. Analysis also includes computing Process Capability Indices (Cp, Cpk) to determine whether each process has the capability of producing VOC specifications, and to detect potential risks using Type I and Type II error probabilities.

As a complement to SPC analysis, the report further elaborates on Design of Experiments (ANOVA) to establish differences in service performance and profitability across product categories and periods. In the last stage, optimisation models are developed with the aim to maximise daily profits in service environments and minimise unreliability costs in staffing problems, reflecting the integration of data analytics with real decision-making. The results are framed in an industrial engineering context to provide actionable recommendations for improving process capability, product quality, and overall operational performance.

Data Quality Report

Data report overview – Head Office Product Data

The dataset examined has the following dimensions:

Feature	Result
Number of observations	360
Number of variables	5

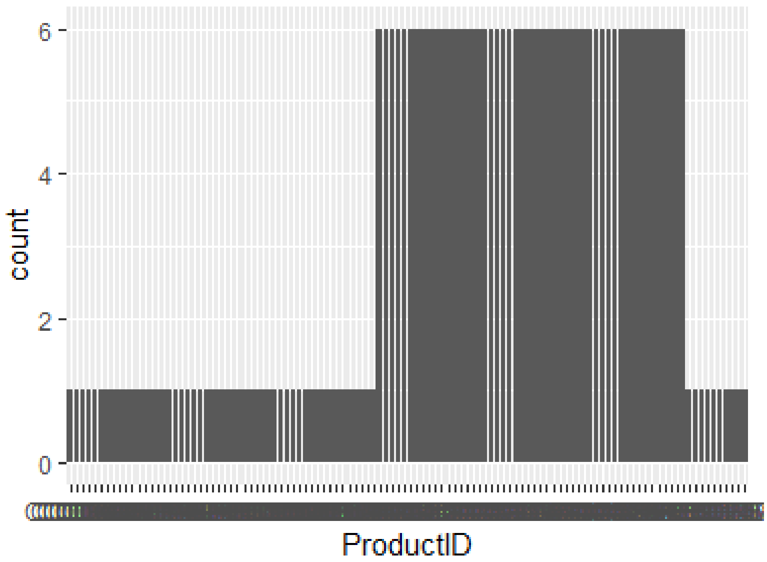
Summary table

	Variable class	# unique values	Missing observations	Any problems?
ProductID	character	110	0.00 %	×
Category	character	6	0.00 %	
Description	character	60	0.00 %	×
SellingPrice	numeric	359	0.00 %	
Markup	numeric	331	0.00 %	

Variable list

ProductID

Feature	Result
Variable type	character
Number of missing obs.	0 (0 %)
Number of unique values	110
Mode	“NA011”



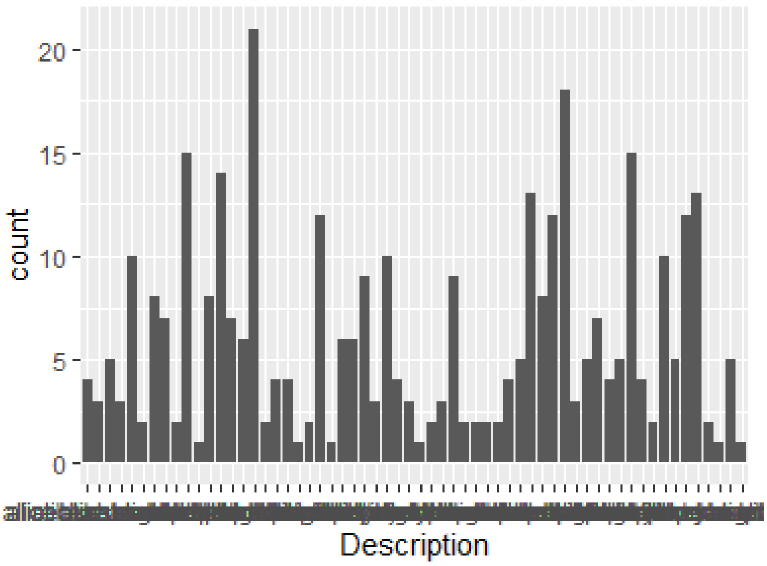
- Note that the following levels have at most five observations: "CLO001", "CLO002", "CLO003", "CLO004", "CLO005", …, "SOF006", "SOF007", "SOF008", "SOF009", "SOF010" (50 values omitted).

Category

Feature	Result
Variable type	character
Number of missing obs.	0 (0 %)
Number of unique values	6
Mode	“Cloud Subscription”

Description

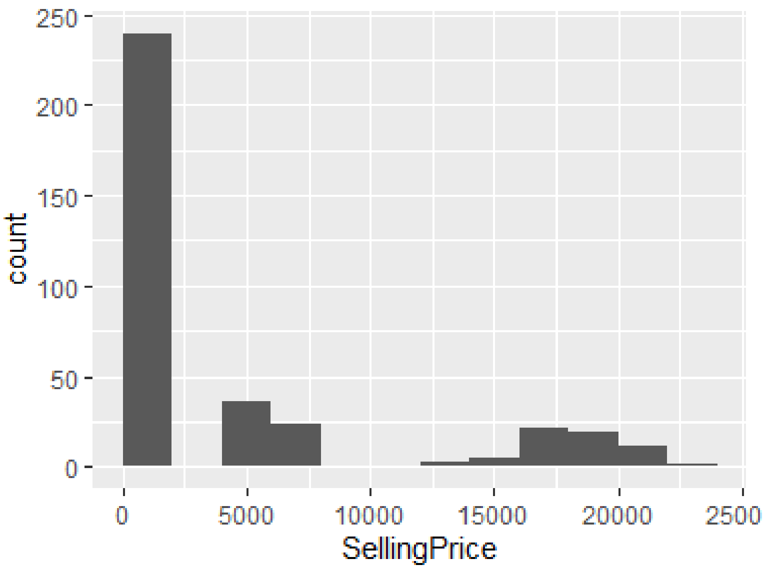
Feature	Result
Variable type	character
Number of missing obs.	0 (0 %)
Number of unique values	60
Mode	“black silk”



- Note that the following levels have at most five observations: "aliceblue bright", "aliceblue marble", "aliceblue matt", "aliceblue sandpaper", "aliceblue wood", …, "cornflowerblue matt", "cornflowerblue wood", "cyan sandpaper", "cyan silk", "cyan wood" (26 values omitted).

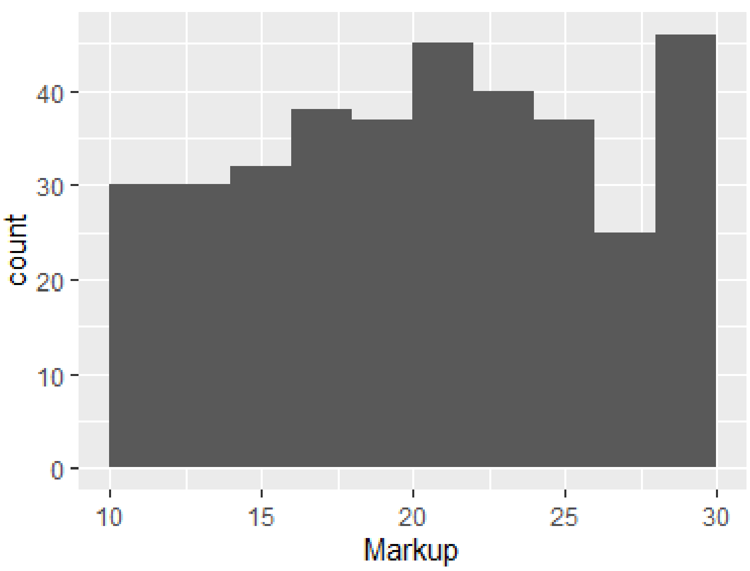
SellingPrice

Feature	Result
Variable type	numeric
Number of missing obs.	0 (0 %)
Number of unique values	359
Median	797.22
1st and 3rd quartiles	495.94; 5843.33
Min. and max.	290.52; 22420.14



Markup

Feature	Result
Variable type	numeric
Number of missing obs.	0 (0 %)
Number of unique values	331
Median	20.58
1st and 3rd quartiles	15.84; 24.84
Min. and max.	10.06; 30



Data report overview – Customer Data

The dataset examined has the following dimensions:

Feature	Result
Number of observations	5000
Number of variables	5

Summary table

	Variable class	# unique values	Missing observations	Any problems?
CustomerID	character	5000	0.00 %	×
Gender	character	3	0.00 %	
Age	numeric	90	0.00 %	
Income	numeric	28	0.00 %	
City	character	7	0.00 %	

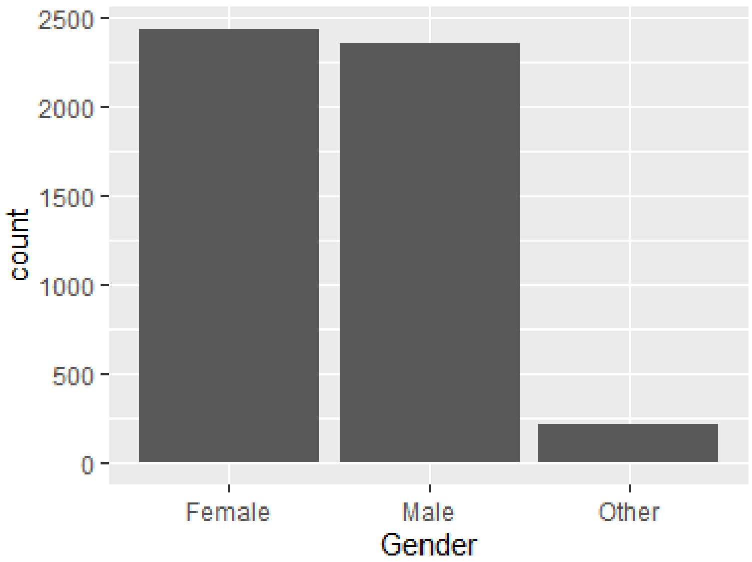
Variable list

CustomerID

- The variable is a key (distinct values for each observation).

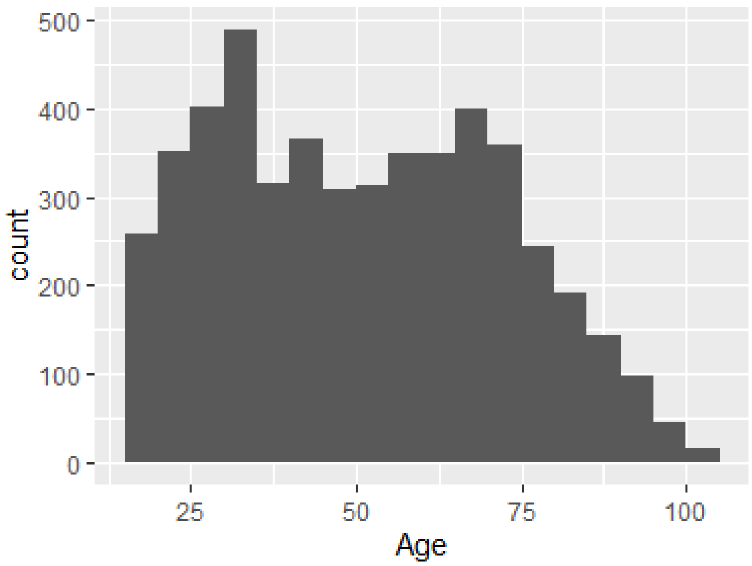
Gender

Feature	Result
Variable type	character
Number of missing obs.	0 (0 %)
Number of unique values	3
Mode	“Female”



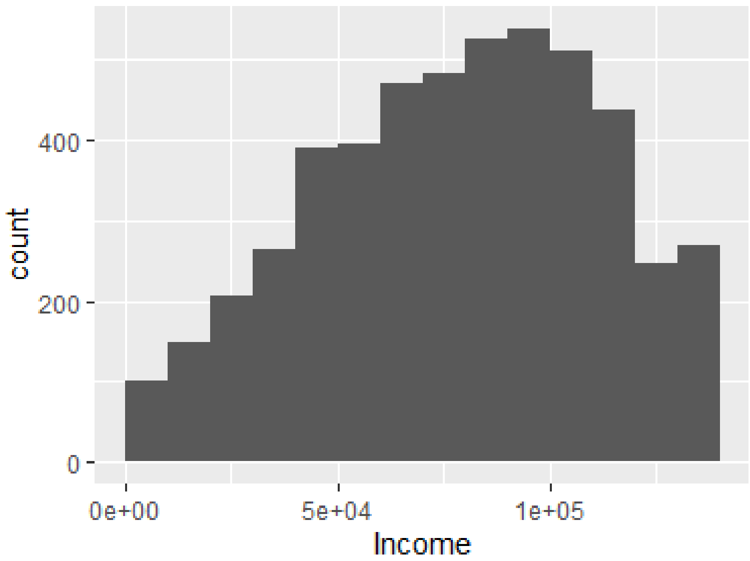
Age

Feature	Result
Variable type	numeric
Number of missing obs.	0 (0 %)
Number of unique values	90
Median	51
1st and 3rd quartiles	33; 68
Min. and max.	16; 105



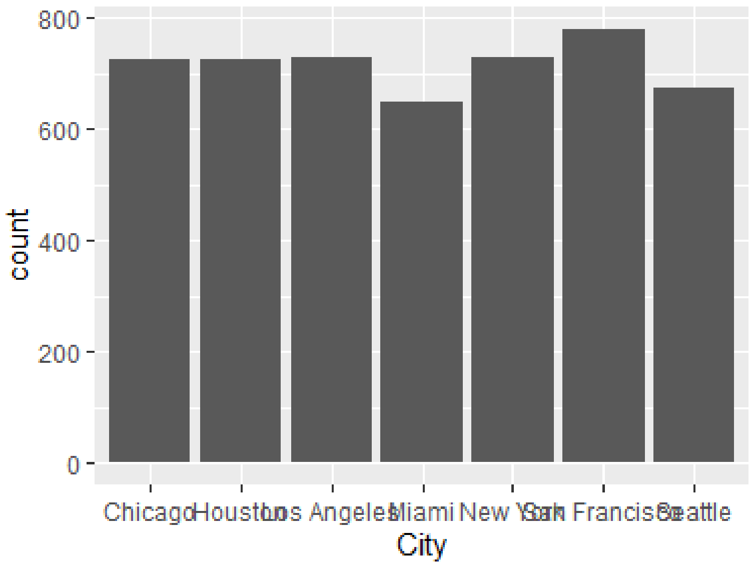
Income

Feature	Result
Variable type	numeric
Number of missing obs.	0 (0 %)
Number of unique values	28
Median	85000
1st and 3rd quartiles	55000; 105000
Min. and max.	5000; 140000



City

Feature	Result
Variable type	character
Number of missing obs.	0 (0 %)
Number of unique values	7
Mode	“San Francisco”



800
600
400
200
0
count
Chicago
Houston
Los Angeles
Miami
New York
San Francisco
Seattle
City

Data report overview – Product Data

The dataset examined has the following dimensions:

Feature	Result
Number of observations	60
Number of variables	5

Summary table

	Variable class	# unique values	Missing observations	Any problems?
ProductID	character	60	0.00 %	×
Category	character	6	0.00 %	
Description	character	35	0.00 %	×
SellingPrice	numeric	60	0.00 %	
Markup	numeric	60	0.00 %	

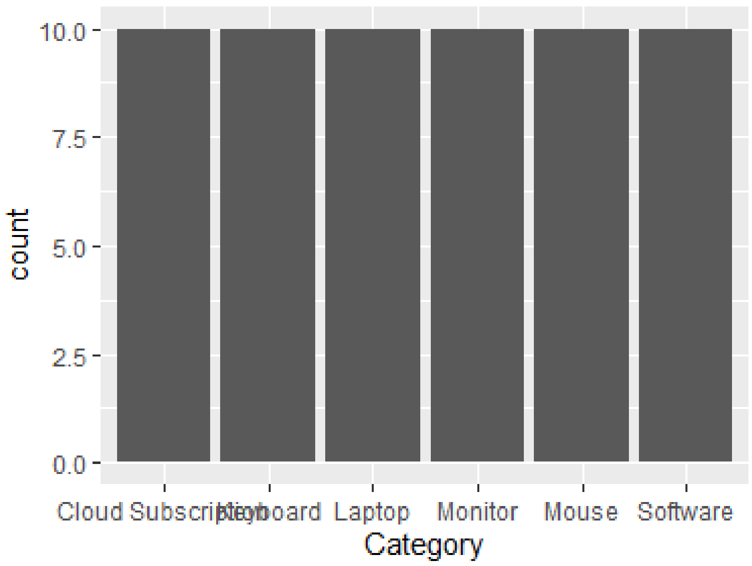
Variable list

ProductID

- The variable is a key (distinct values for each observation).

Category

Feature	Result
Variable type	character
Number of missing obs.	0 (0 %)
Number of unique values	6
Mode	“Cloud Subscription”

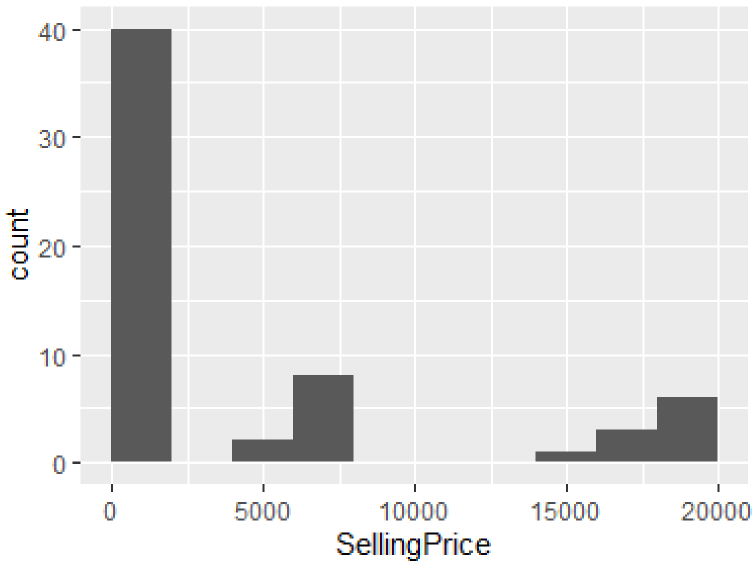


Description

Feature	Result
Variable type	character
Number of missing obs.	0 (0 %)
Number of unique values	35
Mode	“chocolate silk”

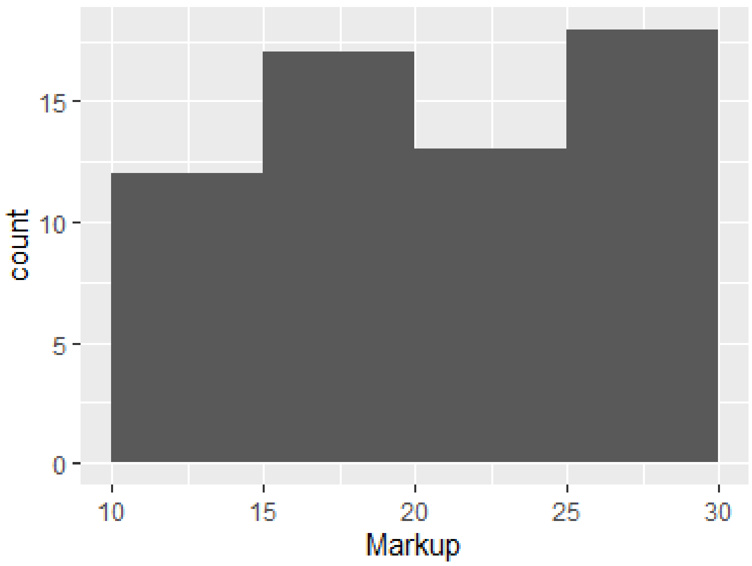
SellingPrice

Feature	Result
Variable type	numeric
Number of missing obs.	0 (0 %)
Number of unique values	60
Median	794.18
1st and 3rd quartiles	512.18; 6416.66
Min. and max.	350.45; 19725.18



Markup

Feature	Result
Variable type	numeric
Number of missing obs.	0 (0 %)
Number of unique values	60
Median	20.34
1st and 3rd quartiles	16.14; 25.71
Min. and max.	10.13; 29.84



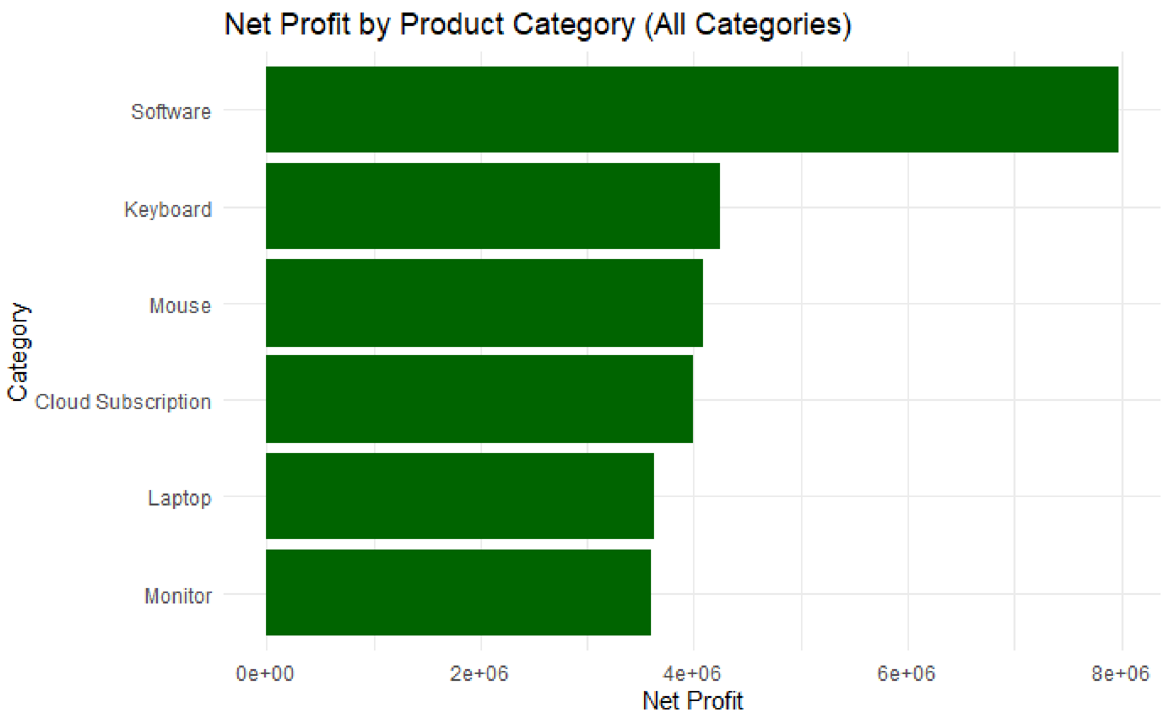
Data Filtering and Sub setting

To ensure consistency in product information, discrepancies between the head office product data and the regional product data were resolved by prioritizing the head office data. Specifically, for each product, if there were conflicting values for the Category or Description fields, the values from the head office dataset were selected. This was implemented in R using the **left_join()** function to combine the datasets, followed by the **coalesce()** function from the **dplyr** library, which chooses the first non-missing value from the specified columns, thereby ensuring that the head office values take precedence whenever available.

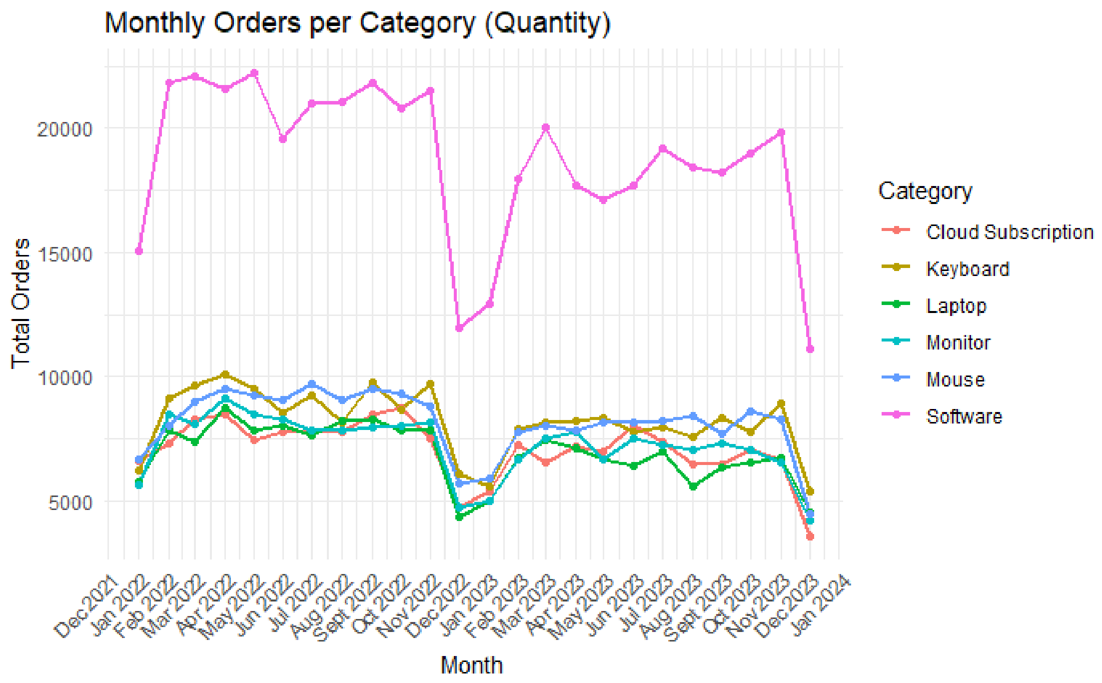
Additionally, the Description column was deemed to have limited analytical value because it primarily contained material or color information, which does not contribute meaningfully to the sales analysis. To simplify the dataset and focus on relevant features, this column was removed using the **select()** function from the **dplyr** library, with negative column selection (-Description) to exclude it. This approach reduced unnecessary complexity in the dataset while retaining all critical variables for subsequent analysis.

Data Visualisation

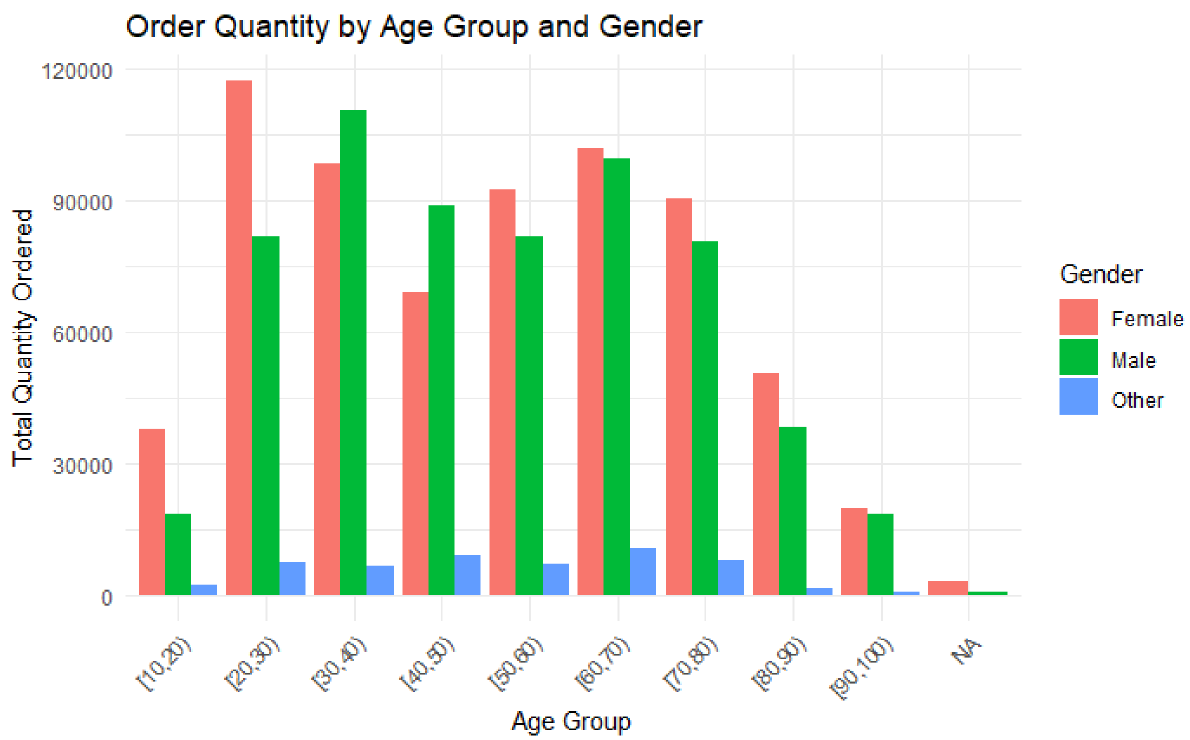
The company has a clear leader in category sales by net profit, as software accounts for more than double the second place (Keyboards)



There is also a much larger amount of Software being purchased than other categories, as seen in the graph below. The hardware sold by the company only offers a marginal profit when compared to software. The order are uniformly distributed over the 2 years in the dataset.

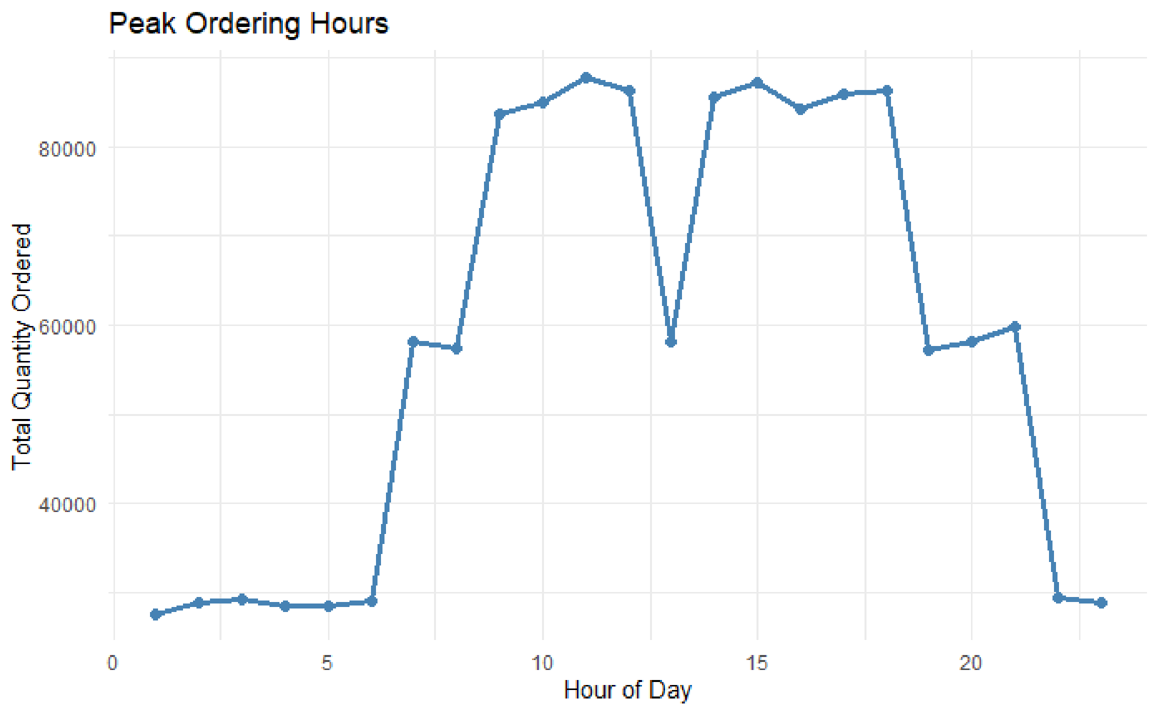


There is a uniform distribution in order quantities by age group and gender, with a slight skewness to the right. This can be attributed to a large number of young individuals who have technology needs.

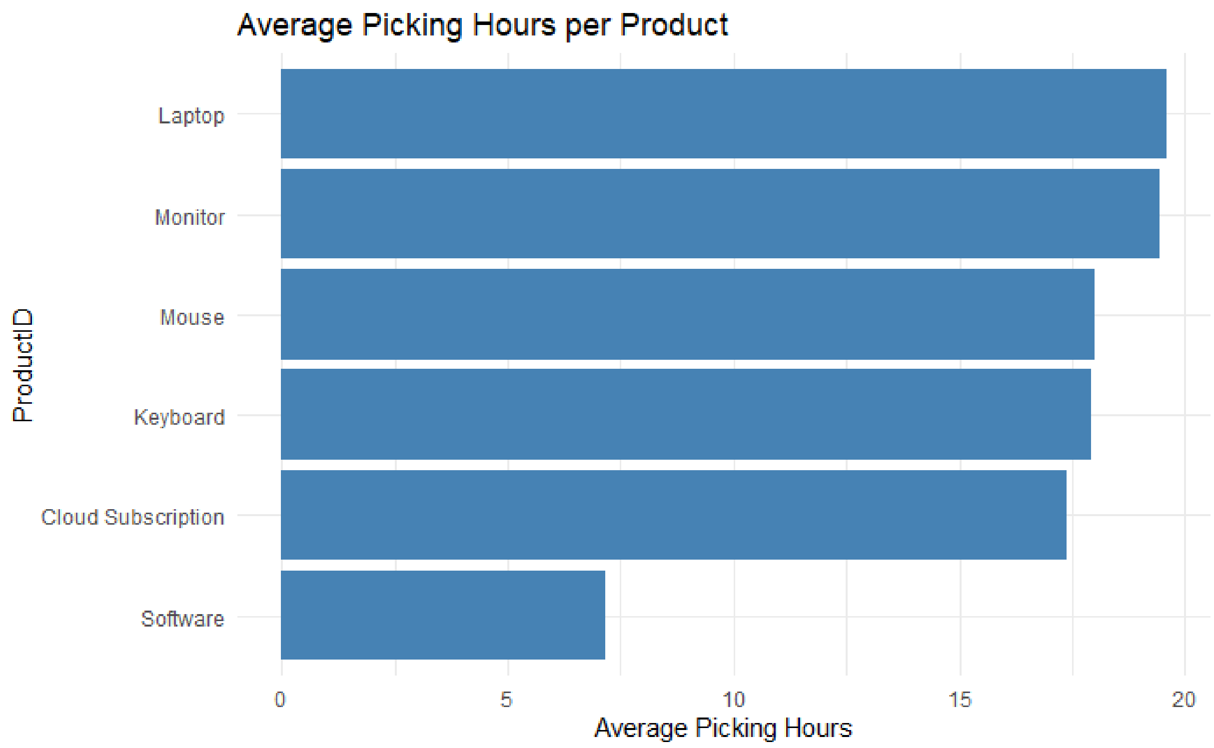


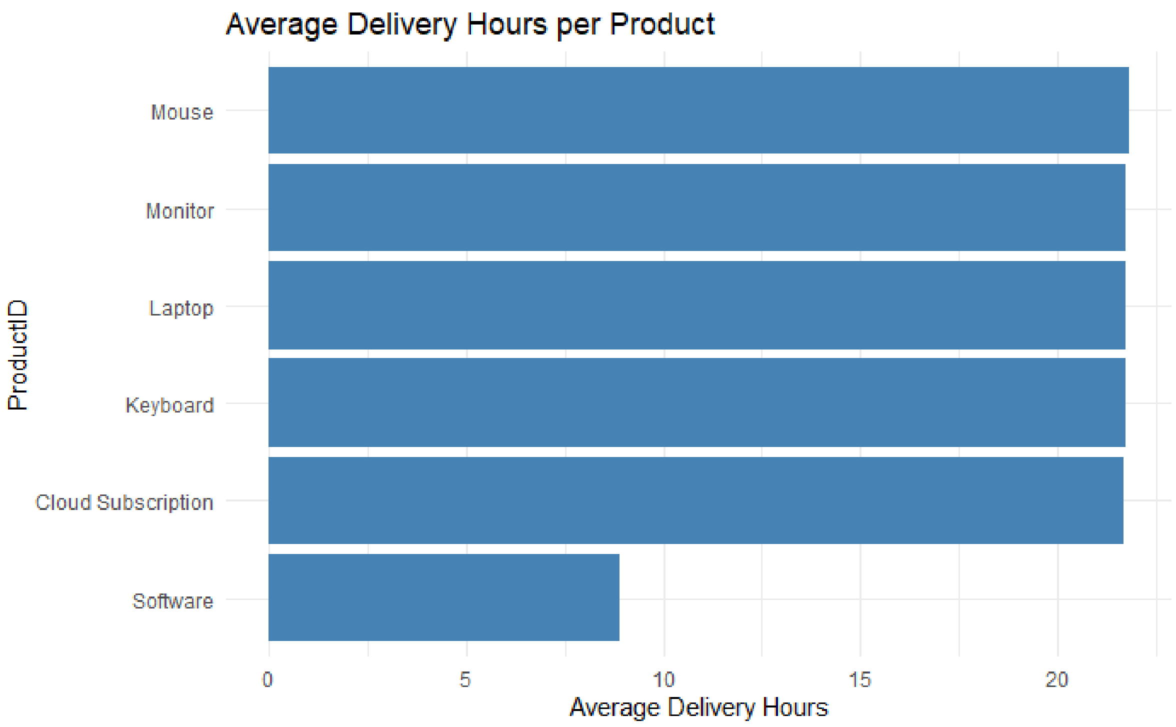
The 7 cities that are served by the company all attract roughly the same number of customers. The largest of the 7 cities is Los Angeles, with almost 50000 more units delivered than in the least popular city, Miami.

There is a clear peak in ordering hours during the afternoons and late morning, with a dip in lunch hour. There is little order activity during late at night or early mornings. The data is normally distributed.



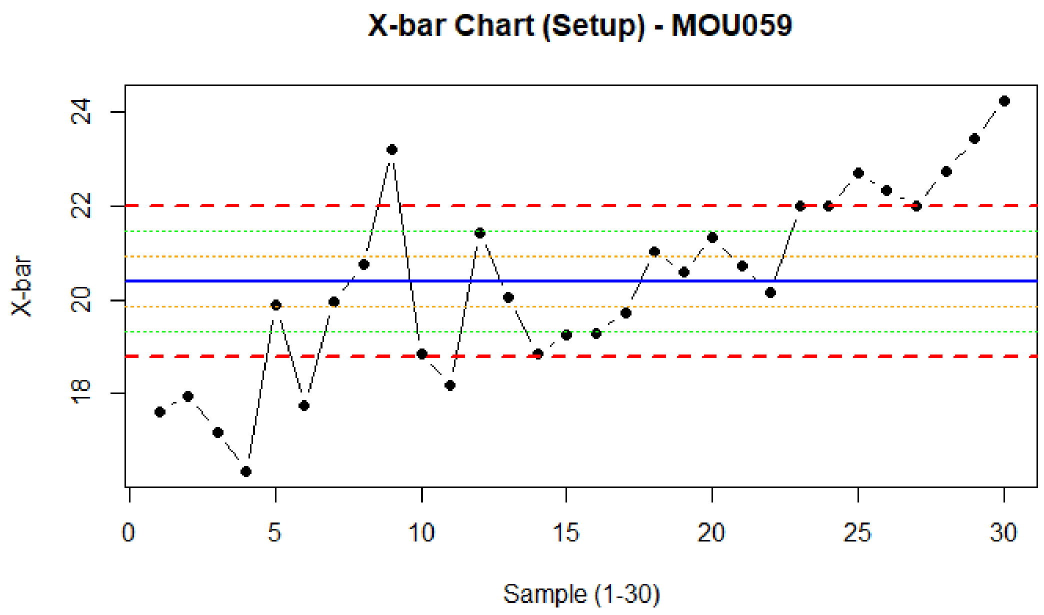
The laptops and monitors take the longest to pick from the warehouse, possibly due to their large size. Keyboards and monitors do not as long as they are smaller, and software and subscriptions take the least amount of time.

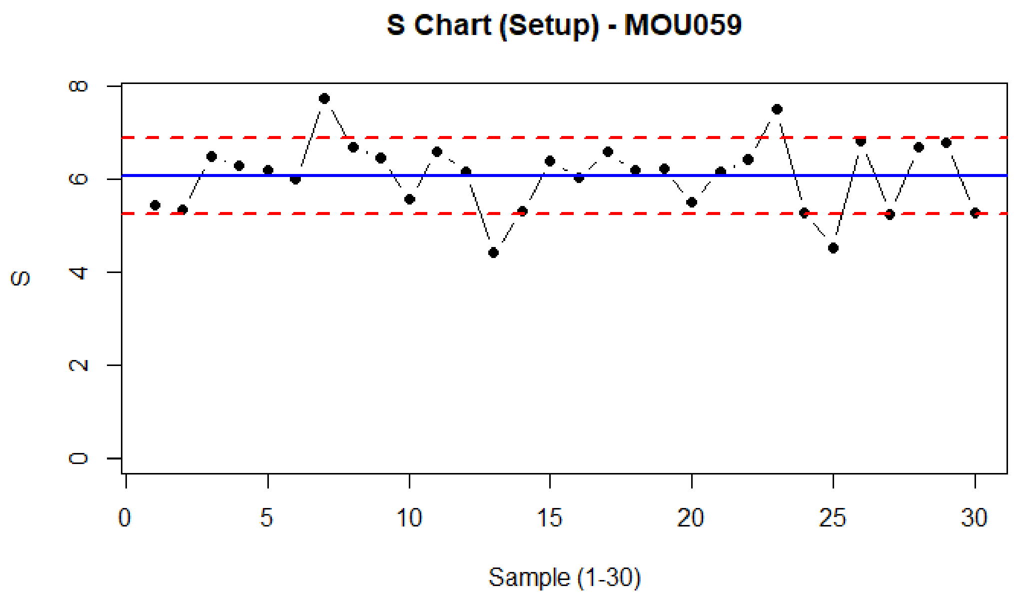



Average Delivery Hours per Product
ProductID
Mouse
Monitor
Laptop
Keyboard
Cloud Subscription
Software
0
5
10
15
20
Average Delivery Hours

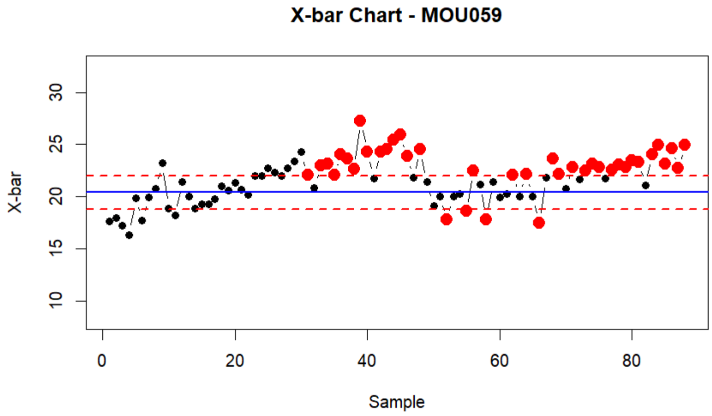
Control Charts

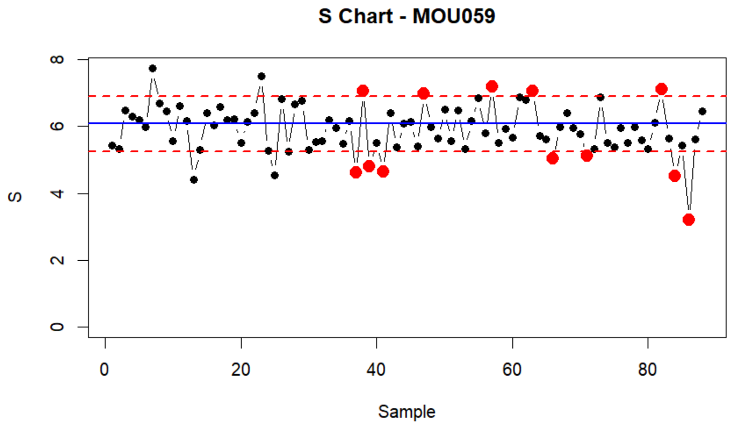
The x and s control charts were set up in RStudio. Below is an extract of the first 2 graphs generated (ProductID: MOU059), with the centre lines, outer control limits, and the 1- and 2-sigma control limits.





The control limits were then applied to the next samples in the dataset. Below is an extract of the first 2 graphs generated, with the points in red being outside of the 1 sigma control line.





Inspecting out of control processes

Samples that show process control issues

All of the processes had at least one sample with either the variance or mean out of control. All products therefore exhibited some degree of process instability. This suggests that the sources of variation are not fully controlled or that special causes may be influencing the process. To improve stability, the process should be reviewed for factors affecting consistency, such as operator practices, material quality, or equipment calibration.

	ProductID	**Rules Flagged**	**RuleA_Samples**	**RuleB_Samples**	**RuleC_Samples**
1	MOU059	A, B, C	31, 33, 34, ..., 86, 87, 88 (Total: 39)	53–54 (Length=2)	33, 34, 35, ..., 86, 87, 88 (Total: 33)
2	KEY049	A, B, C	32, 33, 34, ..., 70, 71, 73 (Total: 25)	59–60 (Length=2)	31, 32, 33, ..., 71, 72, 73 (Total: 21)
3	SOF009	A, B, C	31, 32, 33, ..., 80, 82, 83 (Total: 32)	63–67 (Length=5)	31, 32, 33, ..., 81, 82, 83 (Total: 27)
4	CLO019	A, B, C	32, 33, 34, ..., 61, 62, 63 (Total: 17)	41–41 (Length=1)	56, 57, 58, ..., 61, 62, 63 (Total: 8)
5	KEY045	A, B, C	31, 33, 34, ..., 71, 72, 73 (Total: 25)	52–52 (Length=1)	31, 32, 33, ..., 71, 72, 73 (Total: 19)
6	SOF010	A, B, C	31, 34, 37, ..., 85, 86, 87 (Total: 31)	55–60 (Length=6)	36, 37, 38, ..., 85, 86, 87 (Total: 26)
7	KEY046	A, B, C	31, 32, 34, ..., 74, 75, 76 (Total: 29)	33–33 (Length=1)	37, 38, 39, ..., 74, 75, 76 (Total: 20)
8	CLO012	A, B, C	31, 32, 33, ..., 61, 62, 64 (Total: 16)	50–52 (Length=3)	31, 32, 33, ..., 62, 63, 64 (Total: 8)
9	KEY047	A, B, C	31, 32, 34, ..., 71, 72, 73 (Total: 21)	54–56 (Length=3)	31, 32, 33, ..., 71, 72, 73 (Total: 19)

The table above is an extract of the first 9 entries. It is clear that most of the products flag all of the rules, indicating a need for process control.

Processes capable of meeting VOC

To determine if a process is capable of meeting the Voice of the Customer, the Process Capability (C_p) and Process Capability Index (C_{pk}) are calculated as follows:

$$C_p = \frac{USL - LSL}{6\sigma}$$

$$C_{pk} = \min\left(\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right)$$

Using only the first 1000 deliveries of each product type, all 10 of the products offered under the ‘Software’ category. The table below shows the calculated Process Capability Indices, as well as the VOC indicator. The average C_p and C_{pk} values for non-software products are 0.87 and 0.57 respectively. This indicates that the processes are not capable in general, and that the spread exceeds the specification range. When comparing to the average software C_p and C_{pk} values, 17 and 1.18 respectively, it is clear that these processes are extremely capable, with tight, well-centred control. The products that meet VOC has a C_{pk} > 1.

The high C_p value for software indicates that there is an extremely small standard deviation in delivery times. This suggests that software delivery is done as instant digital transactions. While real, this distorts average capability comparisons.

When comparing hardware C_{pl} and C_{pu}, it becomes clear that the mean is shifted towards the upper limit ($C_{pl} > C_{pu}$), indicating a risk that the process will exceed the required delivery time.

Finding	**Interpretation**	**Action**
Only **Software (SOF)** products meet VOC (Cpk > 1.1).	Indicates automated, low-variability process.	Maintain current controls; monitor monthly.
All other product categories have **Cpk ≈ 0.55–0.60**.	Poor centring and high variation.	Implement SPC feedback loops and operator training.
Cp consistently > Cpk.	Mean shift detected.	Recentre processes by adjusting control parameters.
Cpl > Cpu across dataset.	Output skewed toward high end of tolerance (delays).	Investigate upstream causes (e.g., bottlenecks).

	ProductID	**Cp**	**Cpu**	**Cpl**	**Cpk**	**Meets_VOC**
1	MOU059	0.844	0.570	1.118	0.570	FALSE
2	KEY049	0.845	0.529	1.161	0.529	FALSE
3	SOF009	17.486	33.785	1.187	1.187	TRUE
4	CLO019	0.869	0.568	1.170	0.568	FALSE

5	KEY045	0.847	0.538	1.156	0.538	FALSE
6	SOF010	17.990	34.777	1.202	1.202	TRUE
7	KEY046	0.896	0.572	1.219	0.572	FALSE
8	CLO012	0.865	0.557	1.172	0.557	FALSE
9	KEY047	0.875	0.574	1.176	0.574	FALSE
10	CLO020	0.895	0.621	1.169	0.621	FALSE
11	KEY043	0.880	0.567	1.192	0.567	FALSE
12	MOU058	0.884	0.587	1.182	0.587	FALSE
13	KEY042	0.866	0.566	1.166	0.566	FALSE
14	SOF007	17.516	33.843	1.189	1.189	TRUE
15	CLO011	0.850	0.570	1.130	0.570	FALSE
16	LAP030	0.869	0.553	1.185	0.553	FALSE
17	SOF001	17.201	33.253	1.150	1.150	TRUE

18	SOF002	17.303	33.455	1.151	1.151	TRUE
19	MOU051	0.887	0.568	1.205	0.568	FALSE
20	LAP028	0.856	0.544	1.168	0.544	FALSE
21	SOF005	17.295	33.425	1.166	1.166	TRUE
22	MON037	0.901	0.593	1.209	0.593	FALSE
23	MOU057	0.884	0.585	1.183	0.585	FALSE
24	CLO017	0.878	0.580	1.176	0.580	FALSE
25	KEY048	0.889	0.559	1.218	0.559	FALSE
26	MON032	0.891	0.604	1.177	0.604	FALSE
27	MOU060	0.873	0.560	1.185	0.560	FALSE
28	MON031	0.887	0.573	1.201	0.573	FALSE
29	MON033	0.843	0.566	1.120	0.566	FALSE
30	MOU054	0.857	0.567	1.148	0.567	FALSE

31	LAP023	0.913	0.584	1.241	0.584	FALSE
32	KEY044	0.880	0.575	1.185	0.575	FALSE
33	MON035	0.875	0.575	1.176	0.575	FALSE
34	CLO016	0.856	0.560	1.152	0.560	FALSE
35	MOU052	0.899	0.575	1.222	0.575	FALSE
36	SOF003	18.050	34.893	1.206	1.206	TRUE
37	LAP022	0.917	0.590	1.245	0.590	FALSE
38	LAP025	0.879	0.563	1.195	0.563	FALSE
39	LAP029	0.883	0.569	1.197	0.569	FALSE
40	MOU055	0.888	0.588	1.187	0.588	FALSE
41	SOF004	17.527	33.882	1.172	1.172	TRUE
42	MOU056	0.872	0.560	1.184	0.560	FALSE
43	CLO018	0.846	0.573	1.120	0.573	FALSE

44	SOF006	17.666	34.162	1.170	1.170	TRUE
45	MON040	0.862	0.575	1.149	0.575	FALSE
46	MON036	0.886	0.580	1.191	0.580	FALSE
47	KEY050	0.850	0.539	1.162	0.539	FALSE
48	MON034	0.869	0.582	1.156	0.582	FALSE
49	MON039	0.878	0.599	1.157	0.599	FALSE
50	CLO015	0.886	0.579	1.193	0.579	FALSE
51	MON038	0.875	0.574	1.176	0.574	FALSE
52	CLO014	0.878	0.585	1.170	0.585	FALSE
53	LAP024	0.878	0.557	1.200	0.557	FALSE
54	SOF008	18.237	35.247	1.226	1.226	TRUE
55	LAP021	0.868	0.577	1.158	0.577	FALSE
56	MOU053	0.865	0.547	1.182	0.547	FALSE

57	CLO013	0.859	0.565	1.152	0.565	FALSE
58	LAP027	0.887	0.573	1.202	0.573	FALSE
59	KEY041	0.880	0.579	1.181	0.579	FALSE
60	LAP026	0.881	0.574	1.188	0.574	FALSE

The table above shows the calculated Process Capability Indices, as well as if the product meets VOC.

Risk, Data Correction and optimising for Maximum Profit

Chance of a type I error: A = 0.00135, B = 0.003261; C = 0.00000027

$$\alpha_A = P(Z > 3) = 1 - \Phi(3) \approx 0.00135$$

$$\alpha_C = [P(Z > 2)]^4 = \left(1 - \Phi(2)\right)^4 \approx 0.00000027$$

Chance of a type II error = 0.8412. This means there is an 84.12% chance of failing to detect the process shift with the next sample.

$$H_0 : \mu_0 = 25.05$$

$$H_a : \mu_1 = 25.028, \sigma_{\bar{x}} = 0.017$$

$$\text{Limits: } LCL = 25.011, UCL = 25.089$$

$$\beta = P(\, LCL < \bar{X} < UCL \mid H_a \text{ is true}\,)$$

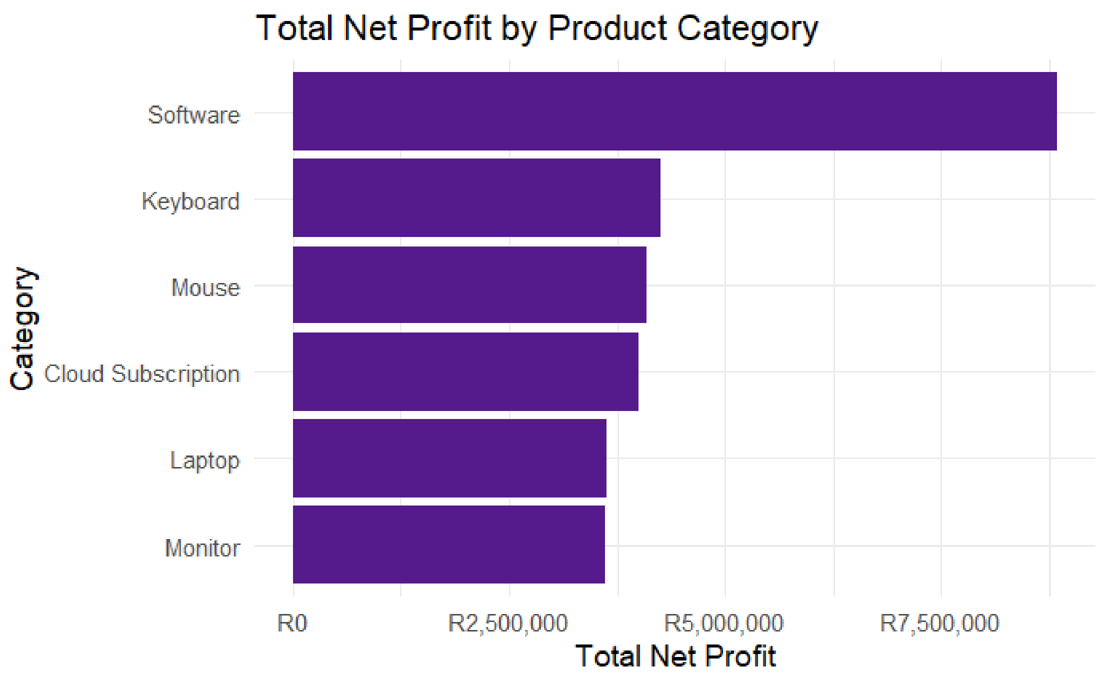
$$\beta = P\left(\frac{LCL - \mu_1}{\sigma_{\bar{x}}} < Z < \frac{UCL - \mu_1}{\sigma_{\bar{x}}}\right)$$

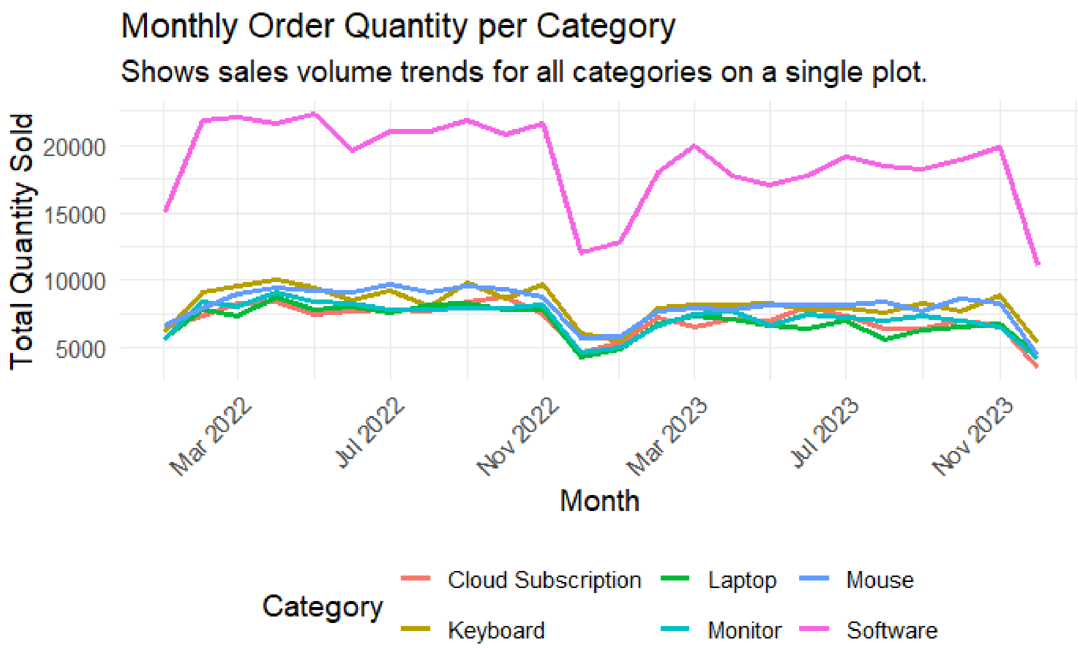
$$\beta = P\left(\frac{25.011 - 25.028}{0.017} < Z < \frac{25.089 - 25.028}{0.017}\right)$$

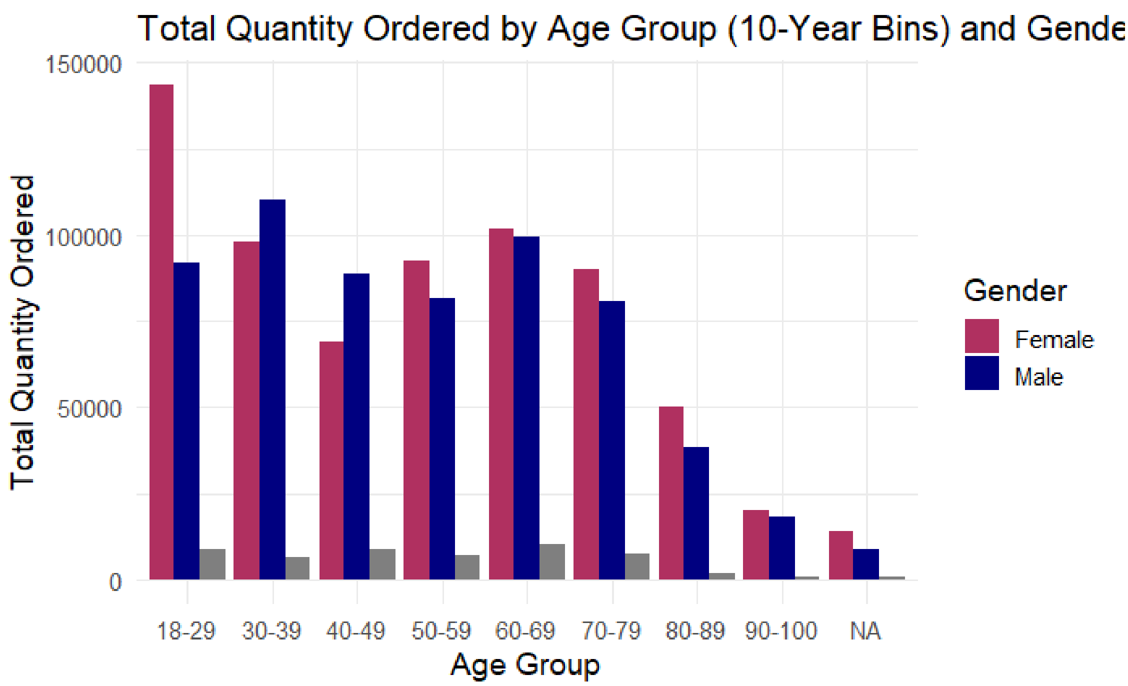
$$\beta = P(-1.0 < Z < 3.588) \approx 0.8412$$

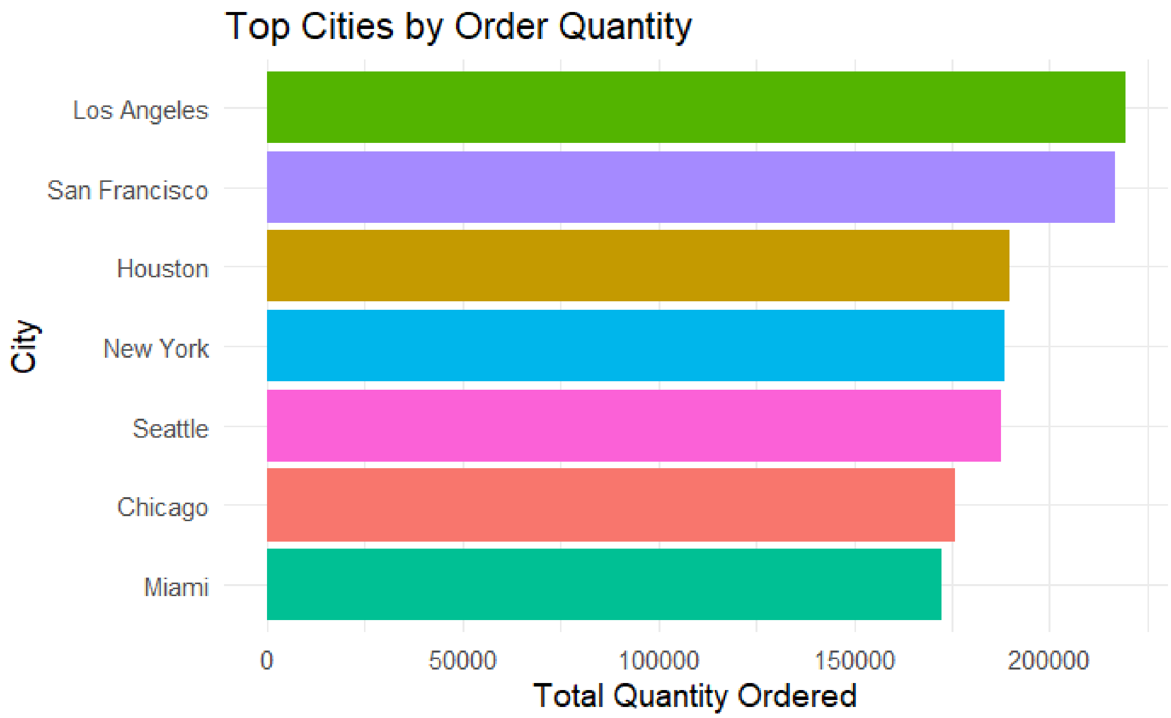
Secondary data analysis

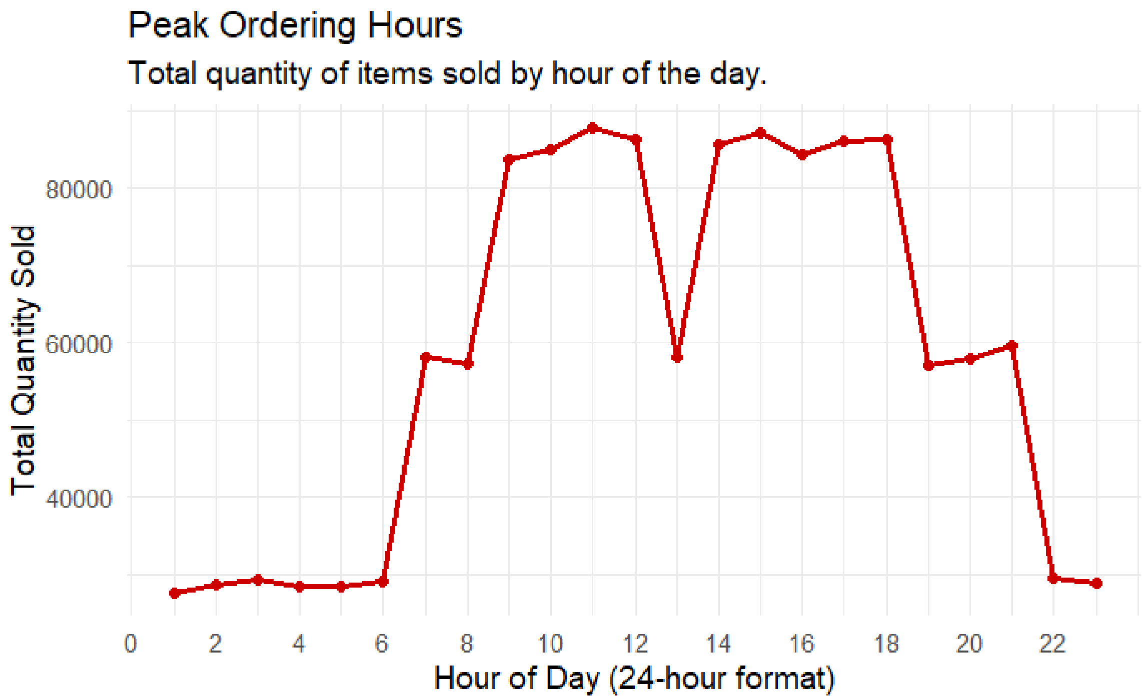
A second data analysis is required after correcting the erroneous data. The following data analysis was done and

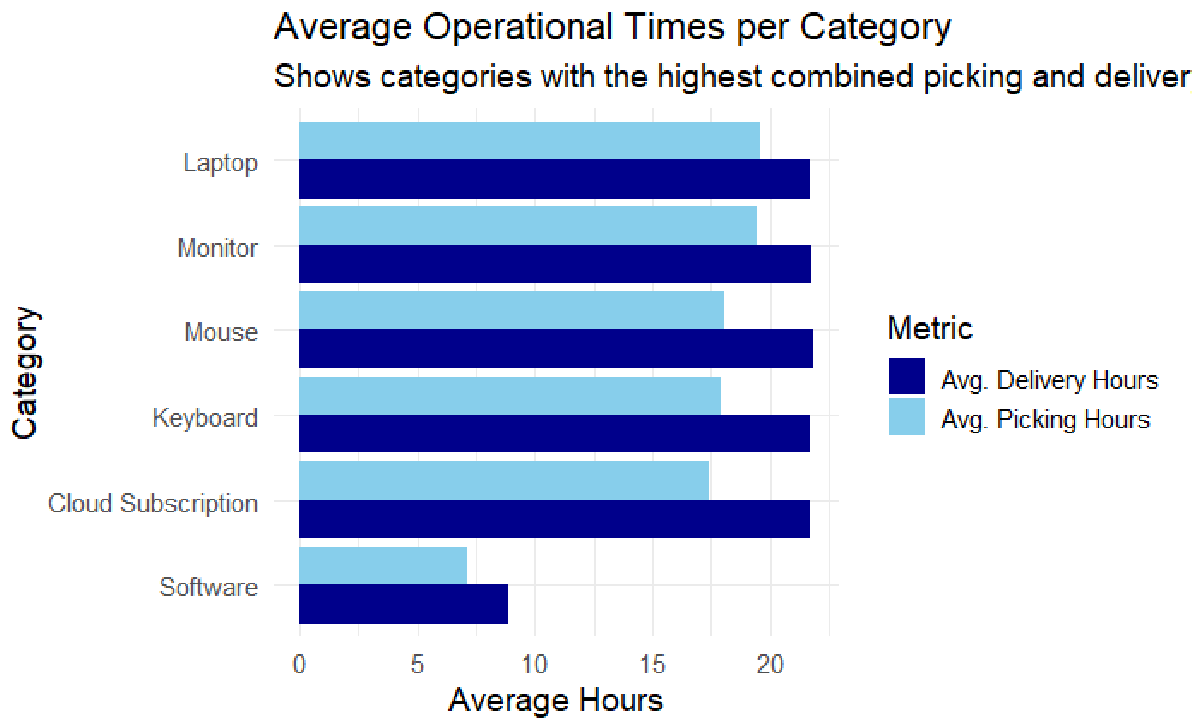

Total Net Profit by Product Category
Software
Keyboard
Mouse
Cloud Subscription
Laptop
Monitor
Category
R0
R2,500,000
R5,000,000
R7,500,000
Total Net Profit


Monthly Order Quantity per Category
Shows sales volume trends for all categories on a single plot.
Total Quantity Sold
20000
15000
10000
5000
Mar 2022
Jul 2022
Nov 2022
Mar 2023
Jul 2023
Nov 2023
Month
Category
Cloud Subscription
Laptop
Mouse
Keyboard
Monitor
Software


Total Quantity Ordered by Age Group (10-Year Bins) and Gende
150000
100000
50000
0
Total Quantity Ordered
18-29
30-39
40-49
50-59
60-69
70-79
80-89
90-100
NA
Age Group
Gender
Female
Male

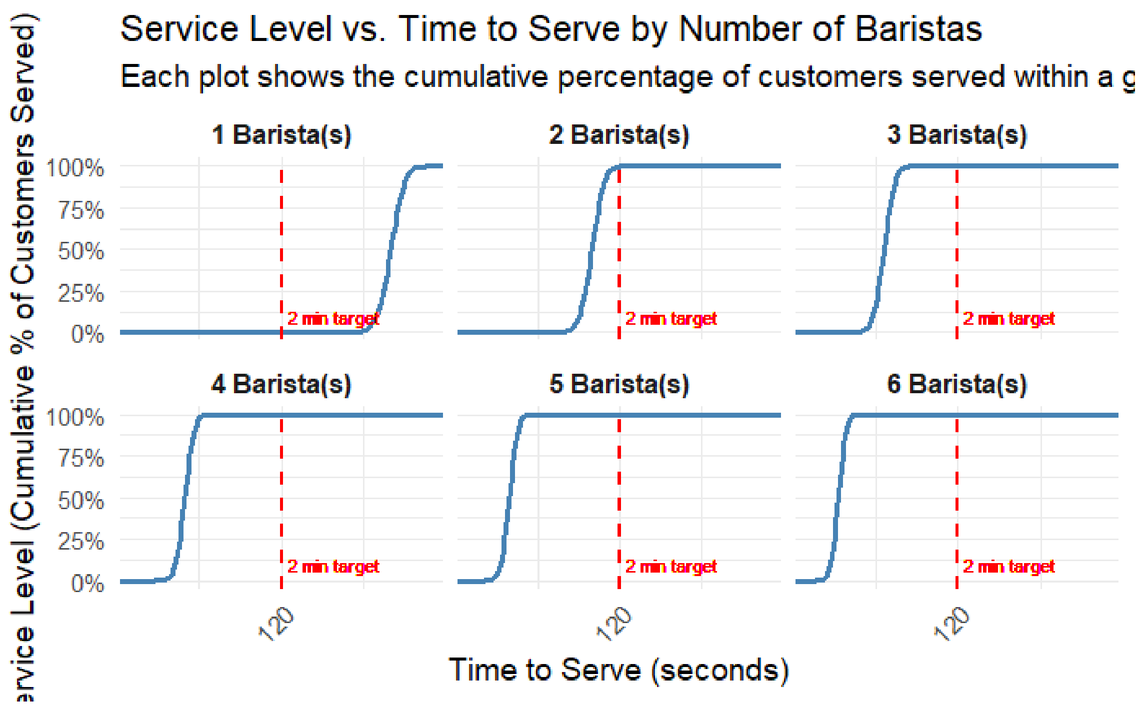

Top Cities by Order Quantity
Los Angeles
San Francisco
Houston
New York
Seattle
Chicago
Miami
City
0
50000
100000
150000
200000
Total Quantity Ordered


Peak Ordering Hours
Total quantity of items sold by hour of the day.
Total Quantity Sold
80000
60000
40000
0
2
4
6
8
10
12
14
16
18
20
22
Hour of Day (24-hour format)


Average Operational Times per Category
Shows categories with the highest combined picking and deliver
Category
Laptop
Monitor
Mouse
Keyboard
Cloud Subscription
Software
0
5
10
15
20
Average Hours
Metric
Avg. Delivery Hours
Avg. Picking Hours

Product Type	Sales value
Laptop	R8147226355
Monitor	R 4627084556
Cloud	R 592292890
Keyboard	R 514493466
Software	R 398810913
Mouse	R 358537041

When optimising the number of baristas to be hired, it will be useful to compare the service level against the number of baristas. We set an arbitrary required serve time of less than 2 minutes. An analysis of the first dataset provided the following graphs:



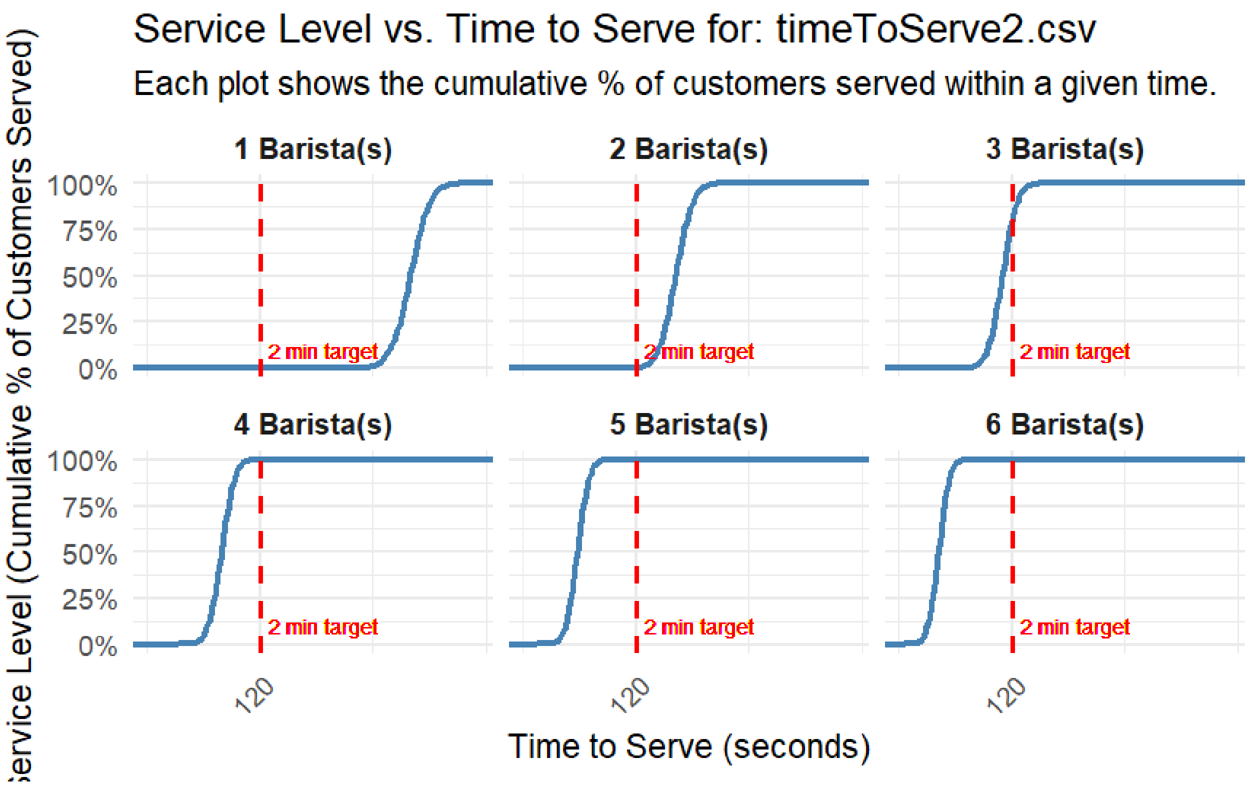
From the graph it is clear that having more than 3 baristas will ensure that we do not run into service level problems. The profits of each barista case were also modelled, and the following was obtained:

Number of Baristas	Avg Service Time (sec)	Potential Customers per Day	Estimated Daily Revenue (R)	Daily Personnel Cost (R)	Estimated Daily Profit (R)
1	200.16	143.89	4316.64	1000	3316.64
2	100.17	287.51	8625.25	2000	6625.25
3	66.61	432.36	12970.69	3000	9970.69
4	49.98	576.23	17286.78	4000	13286.78

5	39.96	720.69	21620.63	5000	16620.63
6	33.36	863.42	25902.66	6000	19902.66

From the first dataset the profit of the coffee shop is maximised when there are 6 baristas employed.

The second data set was analysed, and the following graphs were obtained, with the same 2-minute cut off for good service.



Number of Baristas	Avg Service Time (sec)	Potential Customers per Day	Estimated Daily Revenue (R)	Daily Personnel Cost (R)	Estimated Daily Profit (R)
1	200.17	143.88	4316.35	1000	3316.35
2	141.51	203.51	6105.38	2000	4105.38
3	115.44	249.48	7484.35	3000	4484.35
4	100.02	287.96	8638.68	4000	4638.68
5	89.44	322.02	9660.54	5000	4660.54
6	81.64	352.76	10582.69	6000	4582.69

The second dataset shows indicates a maximum profit at 5 baristas, which is a good balance between the amount of people that can be served in a day and personnel costs.

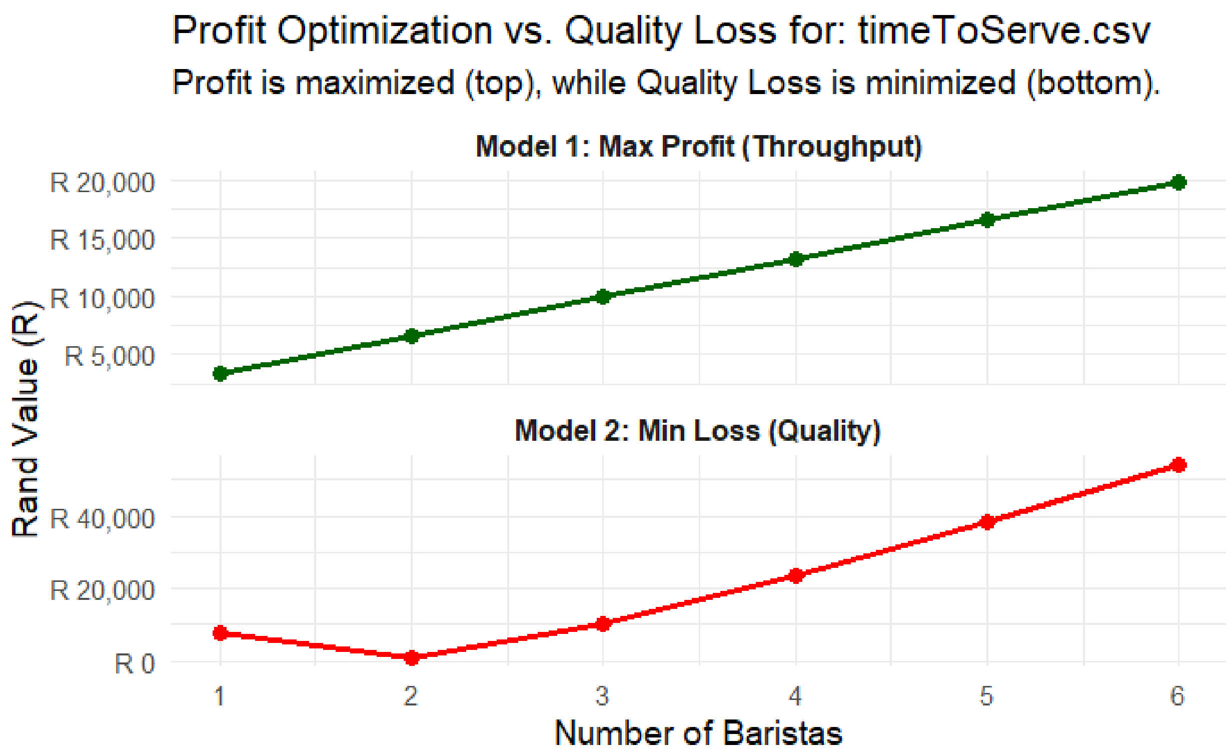
While the initial profit model defines "loss" as simple opportunity cost (where faster service always yields less loss), the Taguchi Quality Loss Function (QLF) offers a customer-centric alternative. Taguchi's philosophy posits that any deviation from the ideal target (T)—in this case, 120 seconds—results in a "loss to society" (like customer dissatisfaction), even if it's within specification. This loss is quadratic, meaning it increases exponentially the further the average service time (x) moves from the target. The general formula is:

$$L(x) = k(x - T)^2$$

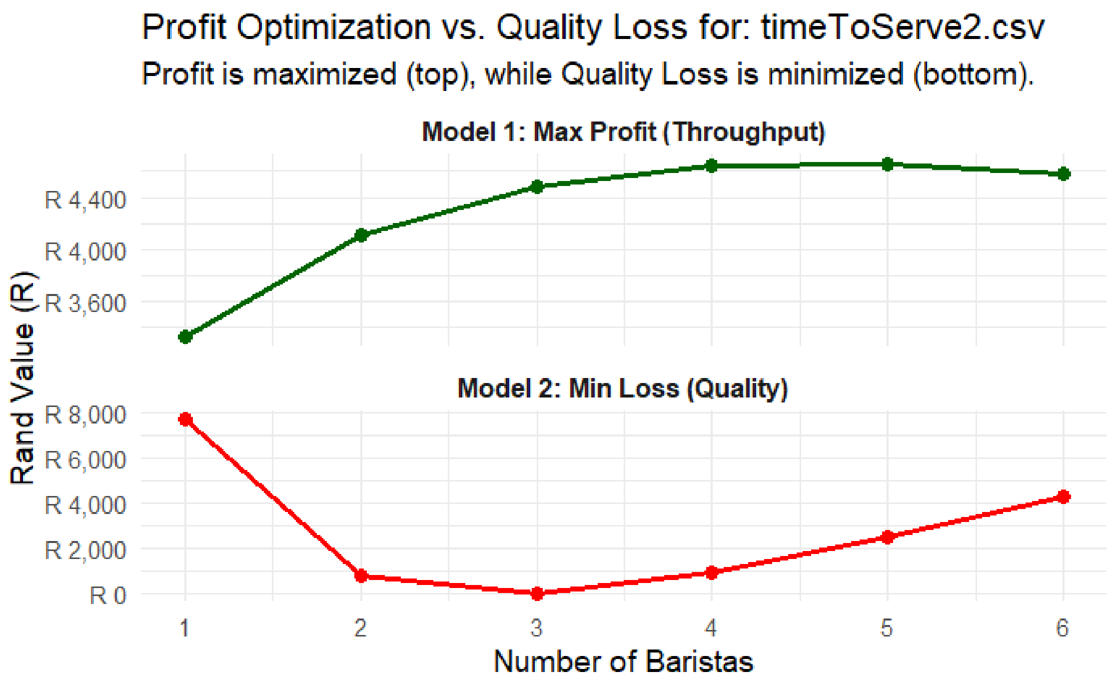
The loss coefficient k is calibrated by defining a cost of failure (A_0) at a specified tolerance limit (Δ_0), using the formula:

$$k = A_0/\Delta_0^2$$

The following plots visually compare the 'best' decision from the profit-centric model (which optimizes for maximum throughput) against the 'best' decision from the Taguchi model (which optimizes for minimum quality loss and consistency).



Although the profit of the company increases with the number of baristas for the first dataset, the loss in quality is minimized when there are 2 baristas assigned.



The second dataset indicates that 3 baristas lead to the lowest loss in quality, whilst maximum profit is obtained at 5 baristas.

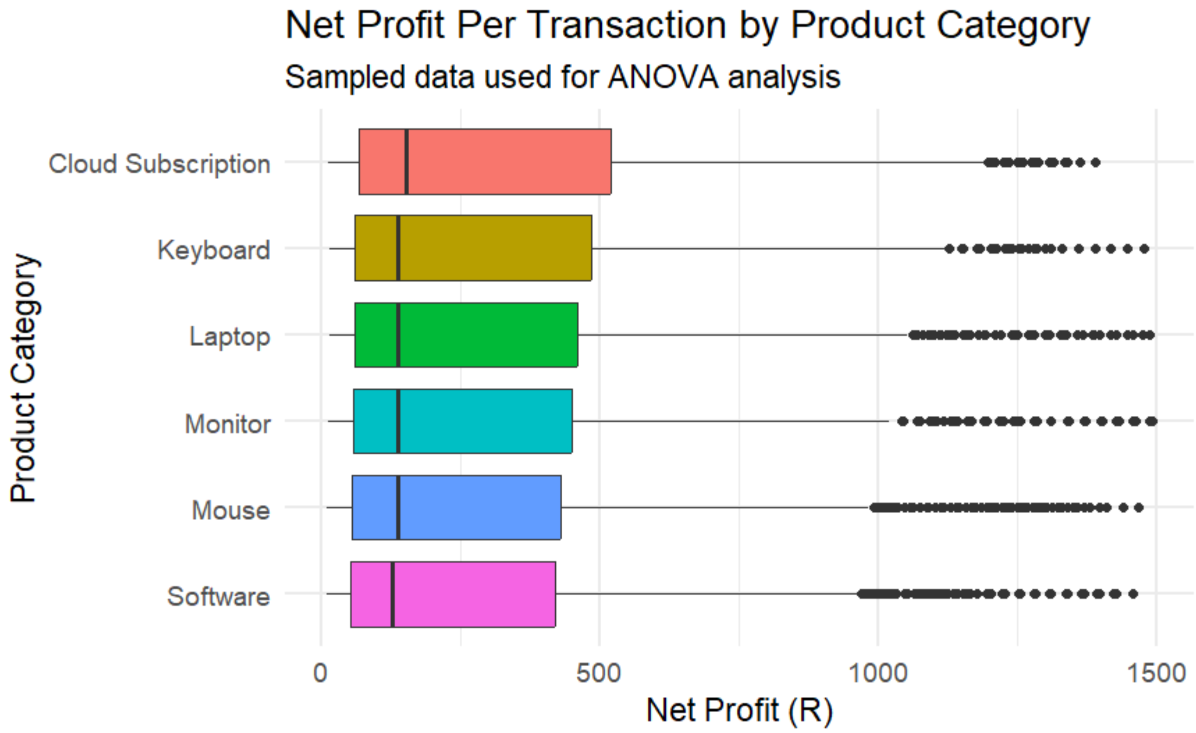
Design of Experiments/ANOVA

Analysis 1: Net Profit by Product Category

Hypothesis:

- H_0 : The mean net profit per transaction is the same for all product categories.
- H_a: At least one product category has a different mean net profit.

The boxplot below displays the distribution of net profit for a balances sample of transactions per category.



Visually, the median net profits over all categories are similar. Each category also shows a considerable number of high-profit outliers, and the interquartile ranges have a large overlap. A statistical test is required to confirm if the differences are indeed significant.

Source	**SS**	**Df**	**MS**	**F-value (fo)**	**P-value**
Treatment	17,292,491	5	3,458,498	35.25	0
Error	7,388,770,145	75,318	98,101	---	---
Total	**7,406,062,636**	**75,323**	---	---	---

The ANOVA test provided a statistically significant result. The F-value of 35.25 indicates a large variation between the categories relative to within each category. The p value is close to 0, far below the 0.05 significance level.

Because the p-value is below the 0.05 significance level, we reject the null hypothesis. There is strong statistical evidence to conclude that the mean net profit per transaction is not the same across all product categories.

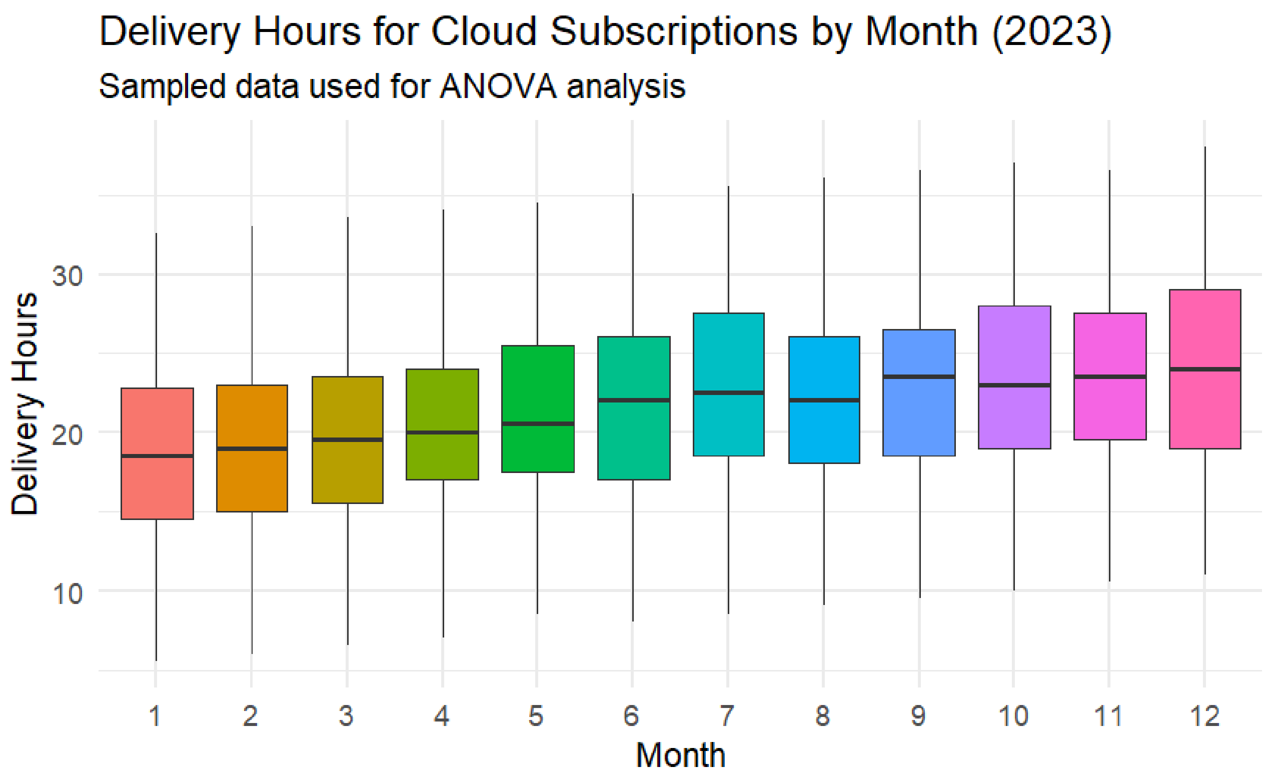
Fisher's LSD test, with a value of 7.75 confirms this. For example, the difference between "Cloud Subscription" and "Keyboard" (24.82) is greater than 7.75, making it a significant difference. In contrast, the difference between "Monitor" and "Mouse" (0.0018) is not statistically significant. This indicates that profitability is highly dependent on the specific product category.

Analysis 2: Delivery Hours for Cloud Subscriptions (Monthly, 2023)

Hypothesis:

- H_0: The mean delivery hours for cloud subscriptions are the same for all 12 months.
- H_a: At least one month has a different mean for delivery hours.

The boxplot below displays the distribution of delivery hours for cloud subscriptions:



Visually, there seems to be an upwards trend in delivery hours as the year progresses. The change in median delivery hours from January to December is on the order of 6 hours. This suggests a degradation of service, and an ANOVA test is required to confirm the visual observation.

Source	SS	Df	MS	F-value (fo)	P-value
Treatment	10,032.55	11	912.05	25.55	0
Error	119,960.77	3,360	35.70	---	---
Total	**129,993.33**	**3,371**	---	---	---

The ANOVA test confirms this visual observation. The F-value of 25.55 is high, and the p-value was 0, which is less than the 0.05 significance level.

Because the p-value is below the 0.05 significance level, we reject the null hypothesis. There is strong statistical evidence to conclude that the mean delivery hours for Cloud Subscriptions were not the same across all months in 2023.

Fisher's LSD test (LSD value = 0.988) provides further detail. While the small difference between adjacent months (e.g., Jan vs. Feb, a difference of 0.56) is not statistically significant, the cumulative change across the year is. The difference between January (Month 1) and December (Month 12) is 5.41 hours, which is far greater than the 0.988 LSD value. This confirms that the degradation in service performance over the year is statistically significant and a real issue to be investigated.

Reliability of Service

Estimate number of days with reliable service

Reliable days in sample (>= 15 people): 366

Proportion of reliable days:

$$P_{\text{reliable}} = \frac{\text{Days with } k \geq 15}{\text{Total Sample Days}} = \frac{96 + 270}{397} \approx 0.9219$$

Estimated reliable days per year:

$$D_{\text{reliable}} = P_{\text{reliable}} \times 365 \approx 336.50 \text{ days}$$

Optimisation of personnel cost:

Model Parameters and Assumptions

The optimisation model is built on the following parameters provided in the case study:

- **Problem Threshold:** A "problem day" occurs if fewer than 15 people are on duty.
- **Cost of Unreliability:** Each problem day results in an average loss of R20,000.
- **Cost of Personnel:** Each assigned person costs R25,000 per month.

For the model, the following calculations and assumptions were made:

1. **Annual Personnel Cost**: The monthly cost was annualized:

   $$C_{\text{person}} = \text{R}25{,}000 \times 12 = \text{R}300{,}000 \text{ per person, per year}$$

2. **Baseline Assigned Staff:** Based on the data provided in Part 7.1, the maximum staff ever present was 16. We will assume the current baseline policy is to have **n = 16** personnel assigned.

3. **Staff Attendance Probability (p):** The probability (p) of any single assigned person showing up for duty is estimated from the empirical data. The expected (mean) number of staff on duty from the 397-day sample is:

   - $E[k] = \frac{\sum(k_i \times \text{days}_i)}{N}$
   - $E[k] = \frac{(12\times1)+(13\times5)+(14\times25)+(15\times96)+(16\times270)}{397} \approx 15.58$
   - Using the binomial mean formula $E[k] = n * p$, we can solve for p:
   - $p = \frac{E[k]}{n_{\text{base}}} = \frac{15.58}{16} \approx 0.9738$

Total Cost Optimisation Model

A function was built to calculate the total annual cost for any given number of assigned staff (n).

1. **Annual Personnel Cost**:

$$C_{\text{Personnel}}(n) = n \times \text{R}300{,}000$$

2. **Annual Unreliability Cost :**

   - The probability of a problem day P(problem) is the probability of fewer than 15 staff (k < 15) showing up, given n are assigned.
   - $P(problem|n) = P(k \leq 14) = \sum_{i=0}^{14} \binom{n}{i} p^i (1-p)^{n-i}$
   - This is calculated using the Binomial Cumulative Distribution Function (CDF): pbinom(14, n, p=0.9738).
   - $C_{\text{Unreliability}}(n) = P(\text{ problem } | \ n\ ) \times 365 \times \text{R}20{,}000$

3. **Total Annual Cost :**

$$C_{\text{Total}}(n) = C_{\text{Personnel}}(n) + C_{\text{Unreliability}}(n)$$

Results

The cost was simulated for a range of employees on duty (12 to 25). The results are presented in the table below.

Assigned_Staff_n <int>	**Total_Annual_Cost** <dbl>	**Personnel_Cost** <dbl>	**Unreliability_Cost** <dbl>	**P_Problem_Day** <dbl>
12	10900000	3600000	7.300000e+06	1.000000e+00
13	11200000	3900000	7.300000e+06	1.000000e+00
14	11500000	4200000	7.300000e+06	1.000000e+00
15	6881108	4500000	2.381108e+06	3.261792e-01
16	5264506	4800000	4.645062e+05	6.363098e-02

17	5166220	5100000	6.621982e+04	9.071208e-03
18	5407593	5400000	7.592970e+03	1.040133e-03
19	5700740	5700000	7.399418e+02	1.013619e-04
20	6000063	6000000	6.348579e+01	8.696684e-06
21	6300005	6300000	4.913564e+00	6.730909e-07
22	6600000	6600000	3.491348e-01	4.782668e-08
23	6900000	6900000	2.307848e-02	3.161435e-09
24	7200000	7200000	1.433787e-03	1.964092e-10
25	7500000	7500000	8.440171e-05	1.156188e-11

Recommendation:

The binomial simulation shows that annual costs are minimized when 17 staff are assigned on any given day. This will account for savings of R98 286 annually.

Conclusion

The ECSA GA4 analysis provided a comprehensive evaluation of process performance and quality across multiple datasets. The descriptive statistics provided clear dominance of the software products by profitability and mentioned even distribution of the customers among demographic and geographic segments. The SPC and capability studies of processes showed that software-related processes were well-capable (Cpk > 1.1), whereas most hardware categories were badly capable and out-of-control, which indicated special causes of variation. The outcomes clearly indicate that regular monitoring of processes and countermeasures such as equipment calibration, operator training, and improved standardisation are essential.

The optimisation models demonstrated profitability and reliability in service operations can significantly be attained with data-driven decision-making. At the barista study, profit was maximized with five to six employees, whereas the model for reliability under binomial simulation showed allocating 17 employees reduces annual costs while maintaining regular service provision. ANOVAs statistically tested statistically significant differences in mean profit between product categories and identified declining performance of delivery hours over months, regarding the need for constant quality improvement.

As a whole, the project demonstrated the application of statistical thinking, computational efficiency, and engineering judgment to solve complex quality assurance problems. Results show how process control and data analytics tools may be applied to improve product consistency, improve service reliability, and help with evidence-based management decisions. It is recommended that SPC, capability analysis, and optimisation modelling be employed routinely to maintain high process performance and customer expectation agreement.

References

- Montgomery, D.C. (2012) *Introduction to statistical quality control*. 7th edn. Hoboken, NJ: John Wiley & Sons.
- Grolemund, G. and Wickham, H. (2011) 'Dates and times made easy with lubridate', *Journal of Statistical Software*, 40(3), pp. 1–25. doi: 10.18637/jss.v040.i03.
- Petersen, A.H. and Ekstrøm, C.T. (2019) 'dataMaid: Your assistant for documenting supervised data quality screening in R', *Journal of Statistical Software*, 90(6), pp. 1–38. doi: 10.18637/jss.v090.i06.
- R Core Team (2023) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. and Yutani, H. (2019) 'Welcome to the tidyverse', *Journal of Open Source Software*, 4(43), p. 1686. doi: 10.21105/joss.01686.
- Xie, Y. (2015) *Dynamic documents with R and knitr*. 2nd edn. Boca Raton, Florida: Chapman and Hall/CRC.