Data Analytics Report

Matthew Campbell

26892855

01/10/2025

Introduction

As a new data analyst, the task is to examine the available datasets with no prior handover. The goal is to carry out basic analysis to understand the data, highlight any quality issues, and link the findings to quality assurance ideas. The data comes from four CSV files: customers, products, products from head office, and sales. This report follows a structure: Data Inspection, Descriptive statistics, data visualization, and interpretation.

Data Inspection

The customer dataset has 5000 rows and includes gender, age, income, and city. The product datasets contain pricing and markup information across different categories. products_Headoffice contains 360 rows and products_data contains 60 rows. The sales dataset has 100 000 rows covering orders from 2022 and 2023. Initial checks showed no missing values or duplicate records across the datasets, which means the data quality is acceptable. This suggests that data collection processes are fairly consistent and thorough, which is positive from a QA perspective.

Descriptive Statistics

Customer data

The customer age distribution shows most customers are between 30 and 60 years old. Income distribution shows that most customers are in middle to upper income brackets. From a QA perspective, these differences in demographics highlight that service quality must be consistent across groups with different needs and expectations.

Products data

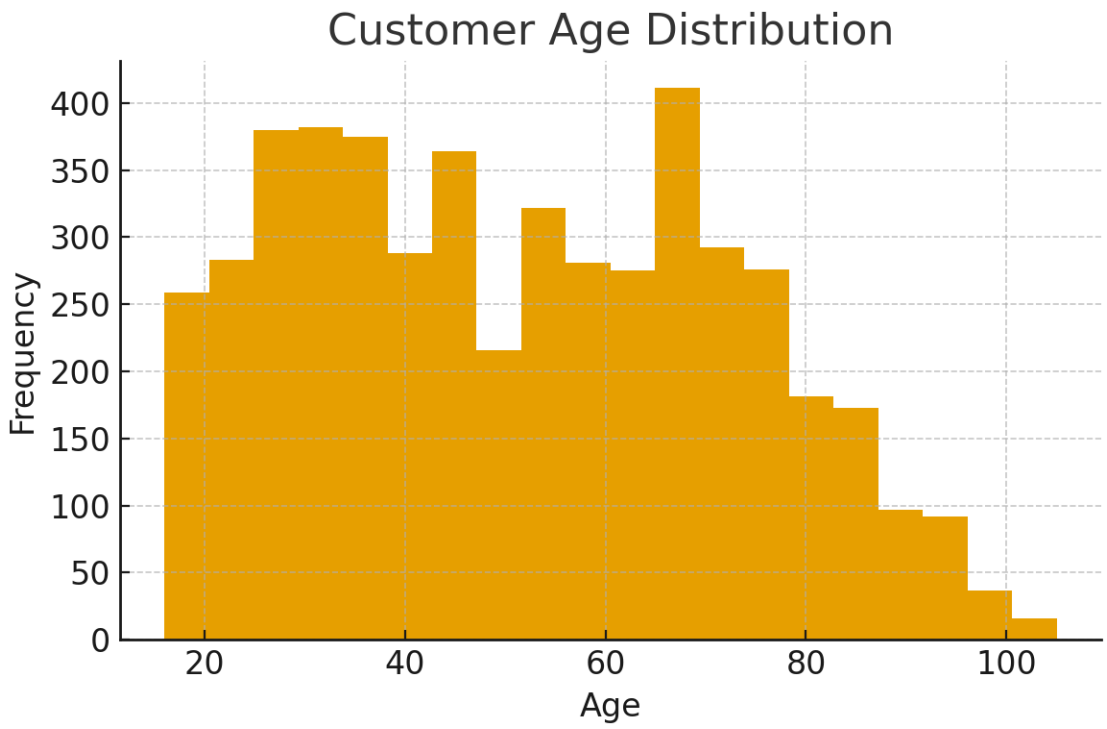
The average selling price of products from products_data is 4493.6, with values ranging from 350.4 to 19725.2. Markup percentages all fall between 10% and 30% with an average of 20.46%. The average selling price of products from products_Headoffice is 797.2 with values ranging from 290.5 to 22420.1. Markup percentages all fall between 10% and 30% with an average of 20.39%. QA concerns may arise when higher-priced products do not deliver an acceptable level of quality. Consistent quality across products is vital.

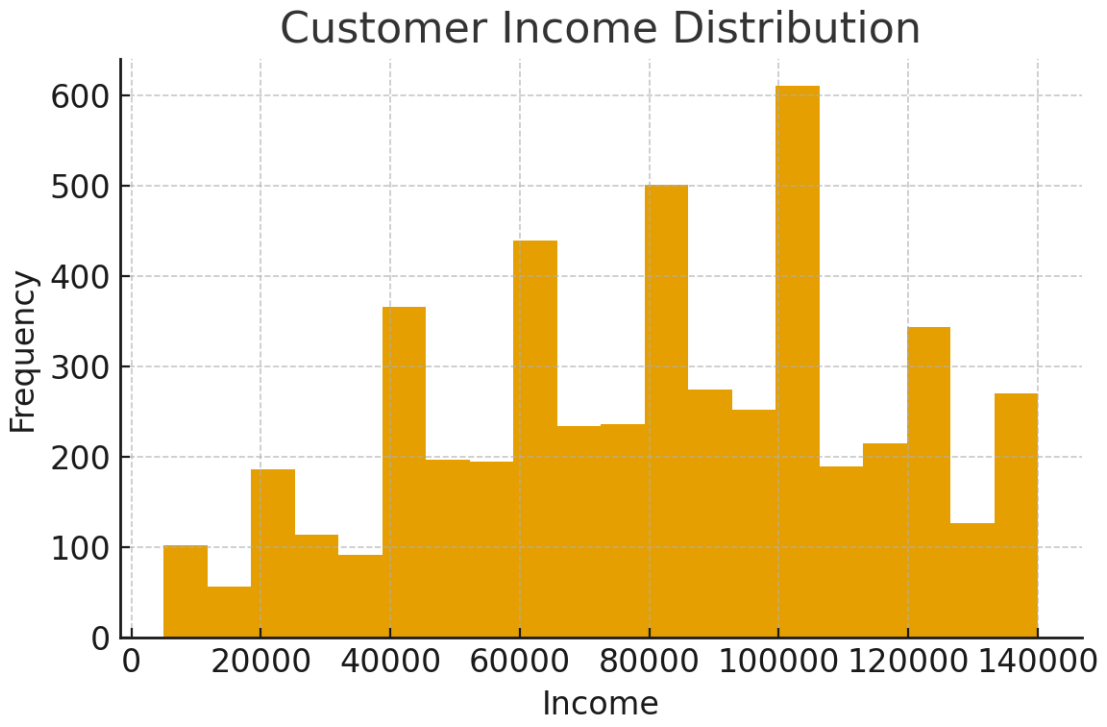
Sales Data

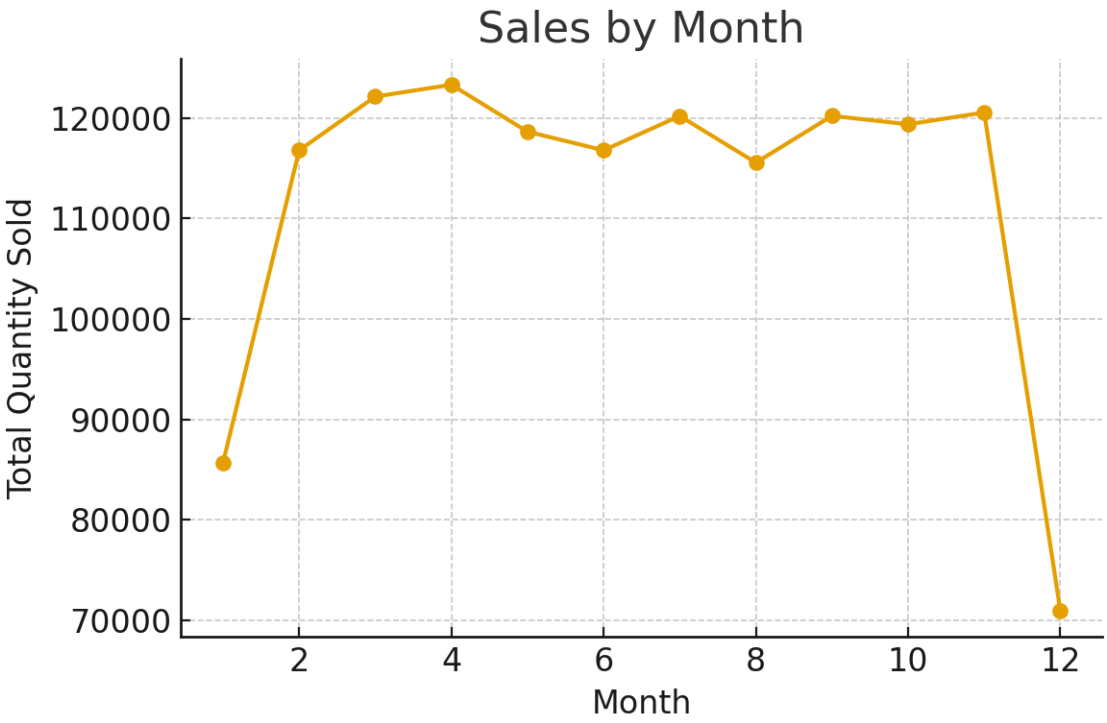
Sales volumes peaked in March and April, with a drop in December. Total sales also dropped from 2022 to 2023. This drop in sales may have been due to several reasons, including customer dissatisfaction which may be linked to quality issues. The average quantity of products ordered is 13.5. the average time of day, day of the month and month that products are ordered are 12:56pm, 15th and June respectively.

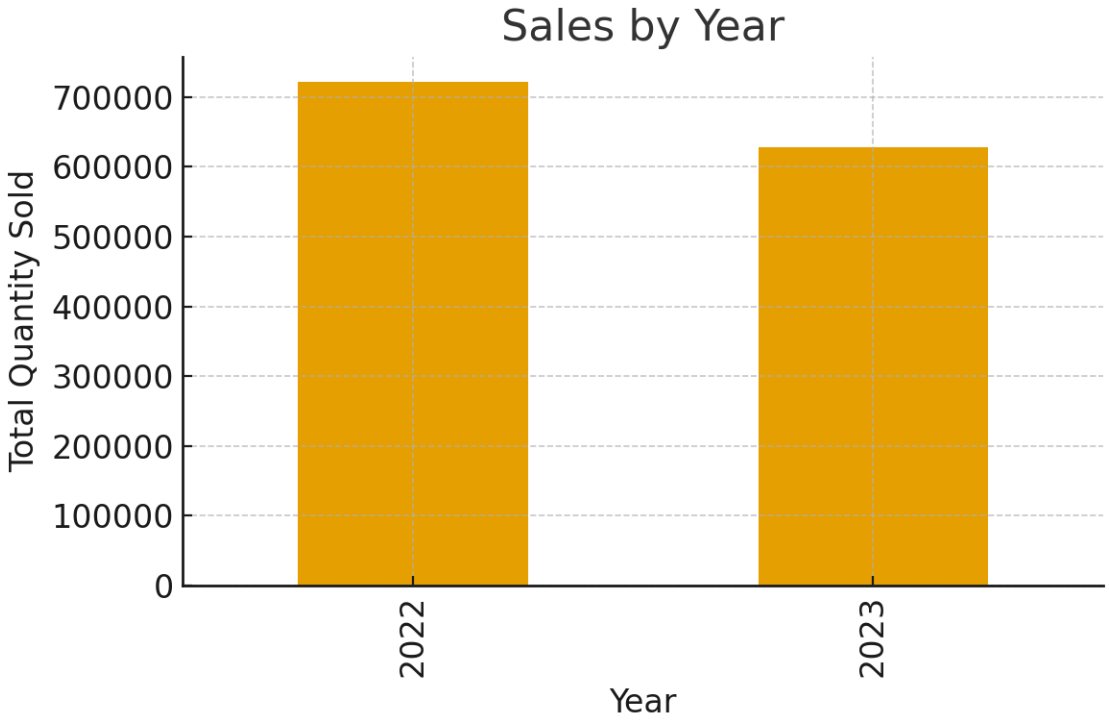
The average delivery time is 17.5 hours. Delivery times show large variation, with some orders taking less than an hour and others taking over 38 hours. In QA terms, this represents a process with high variability. Statistical Process Control (SPC) methods could help to monitor and reduce this inconsistency, improving reliability for customers.

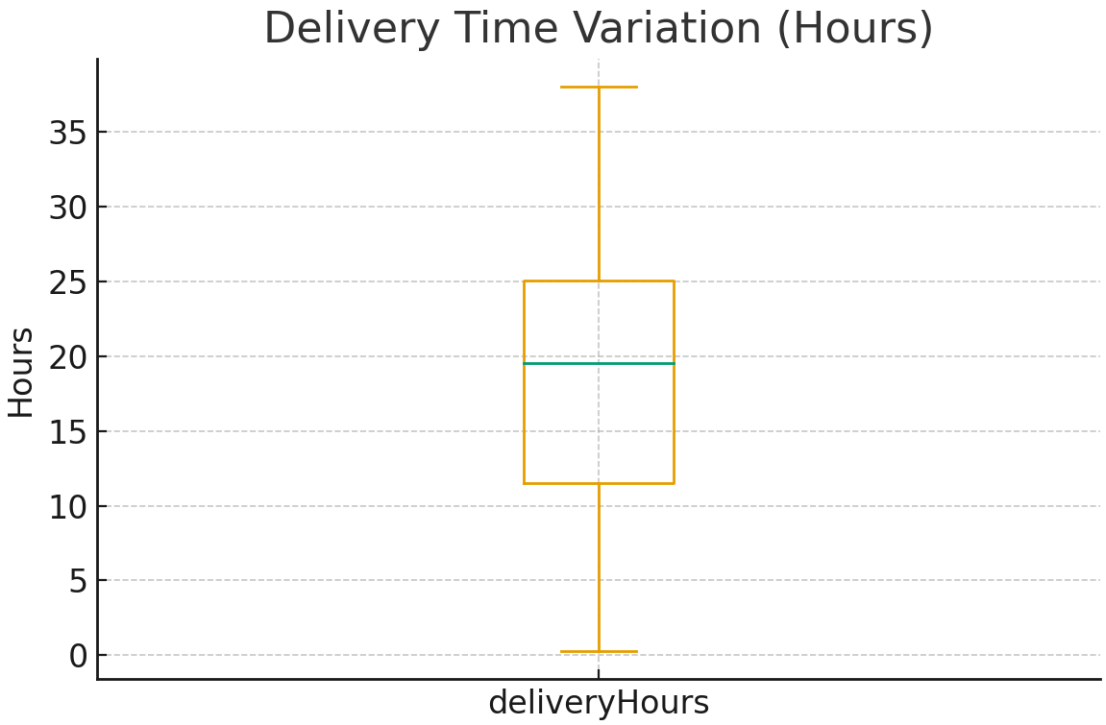
Data Visualization






Sales by Month
Total Quantity Sold
120000
110000
100000
90000
80000
70000
2
4
6
8
10
12
Month


Sales by Year
Total Quantity Sold
700000
600000
500000
400000
300000
200000
100000
0
2022
2023
Year


Delivery Time Variation (Hours)
Hours
35
30
25
20
15
10
5
0
deliveryHours

Interpretation

1. Customer demographics vary widely, meaning QA must ensure consistency for all groups.

2. Product pricing and markup differences require QA checks to ensure fairness and reliability.

3. Sales declined between 2022 and 2023, possibly linked to quality issues or process inefficiency.

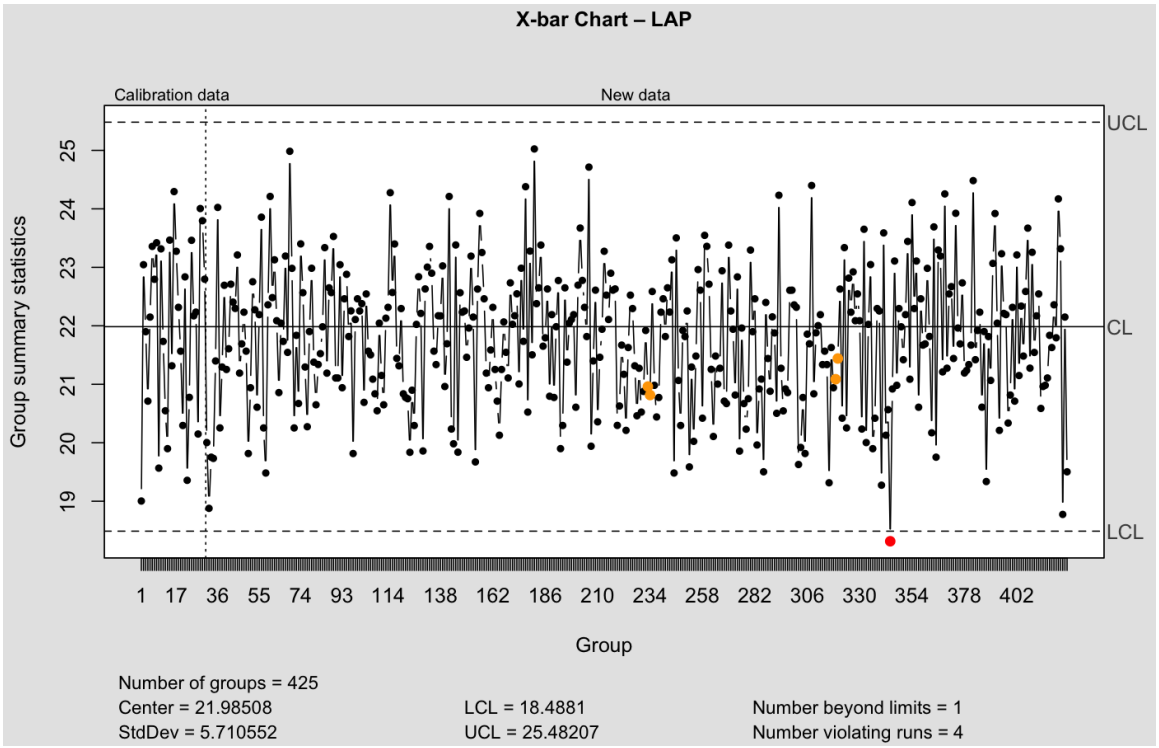
4. Delivery time variation is a major QA concern that needs monitoring and control.

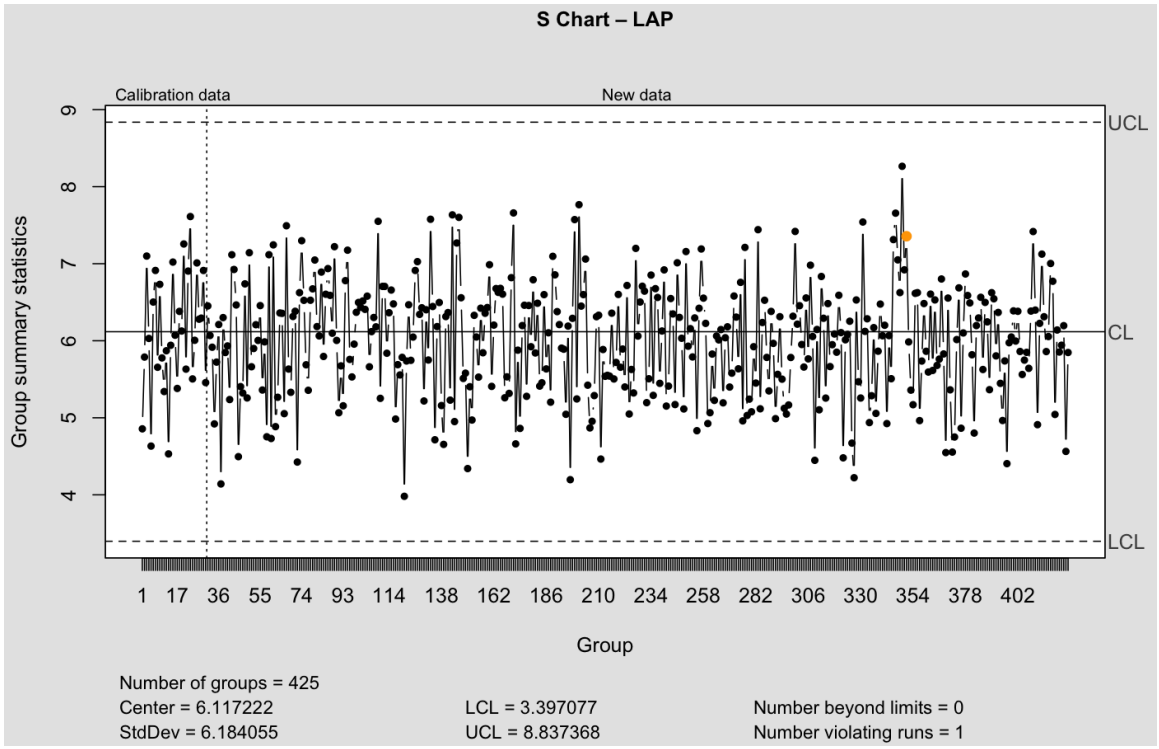
SPC Analysis

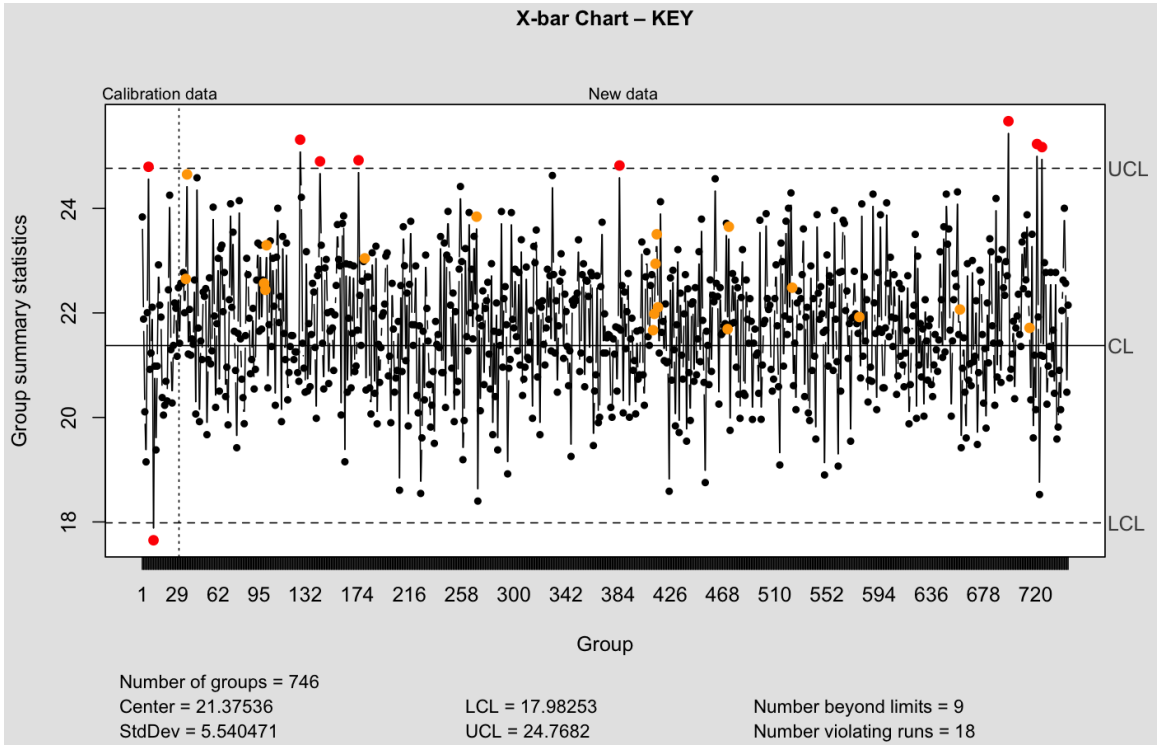
This section applies SPC to evaluate delivery times from 2026 to 2027. Using samples of 24 deliveries per product type. X-bar and S-charts were constructed to monitor process means and variation over time.

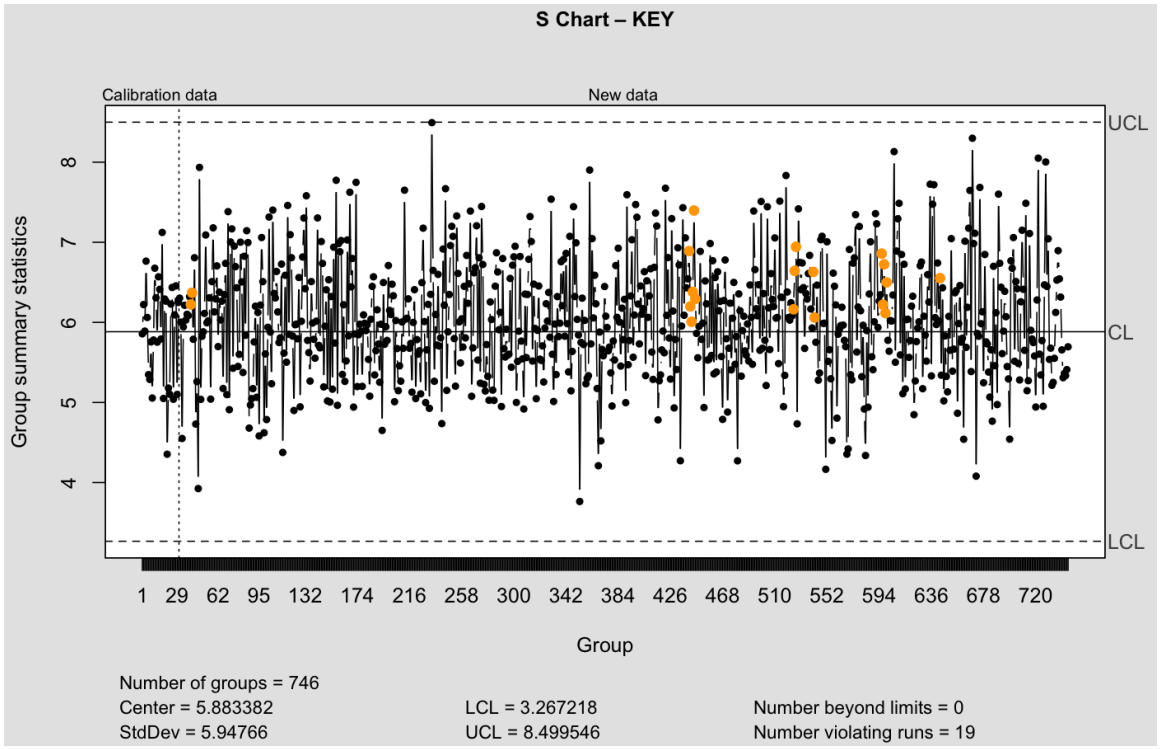
X-bar and S-Charts

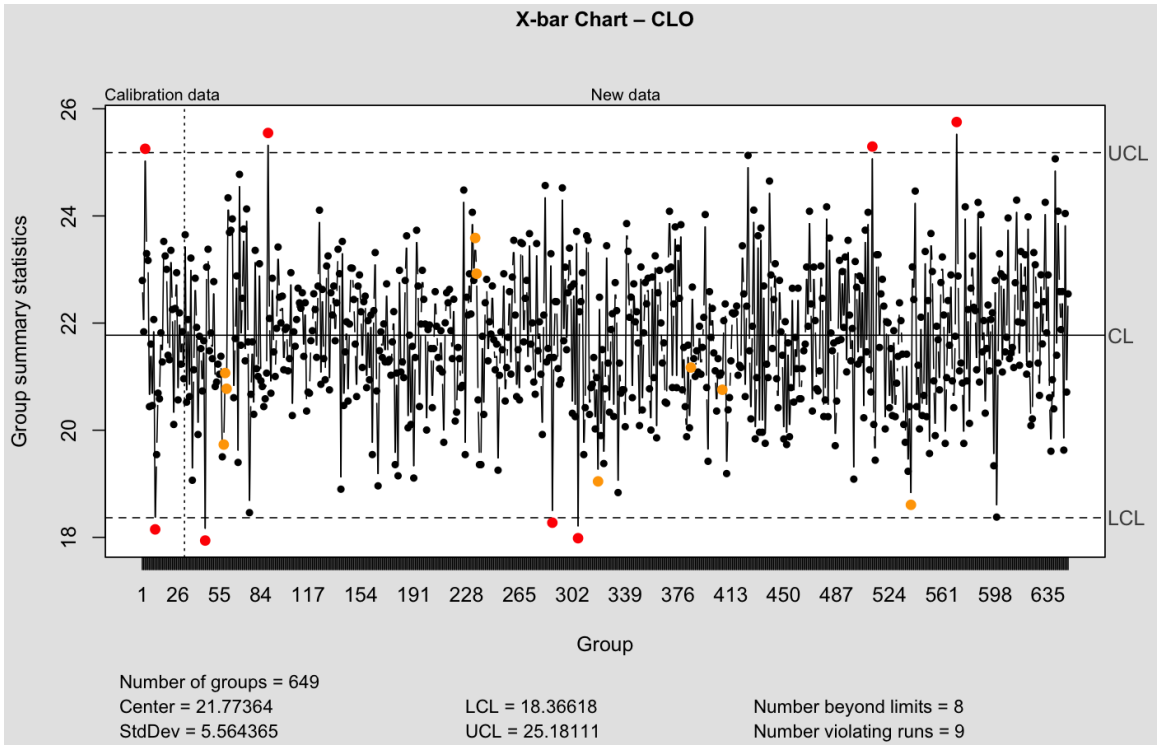
The following are X-bar and S-charts for the product types. Each chart displays center lines, +-1σ, +-2σ, and +-3σ control limits, using delivery time (hours) as the process metric.

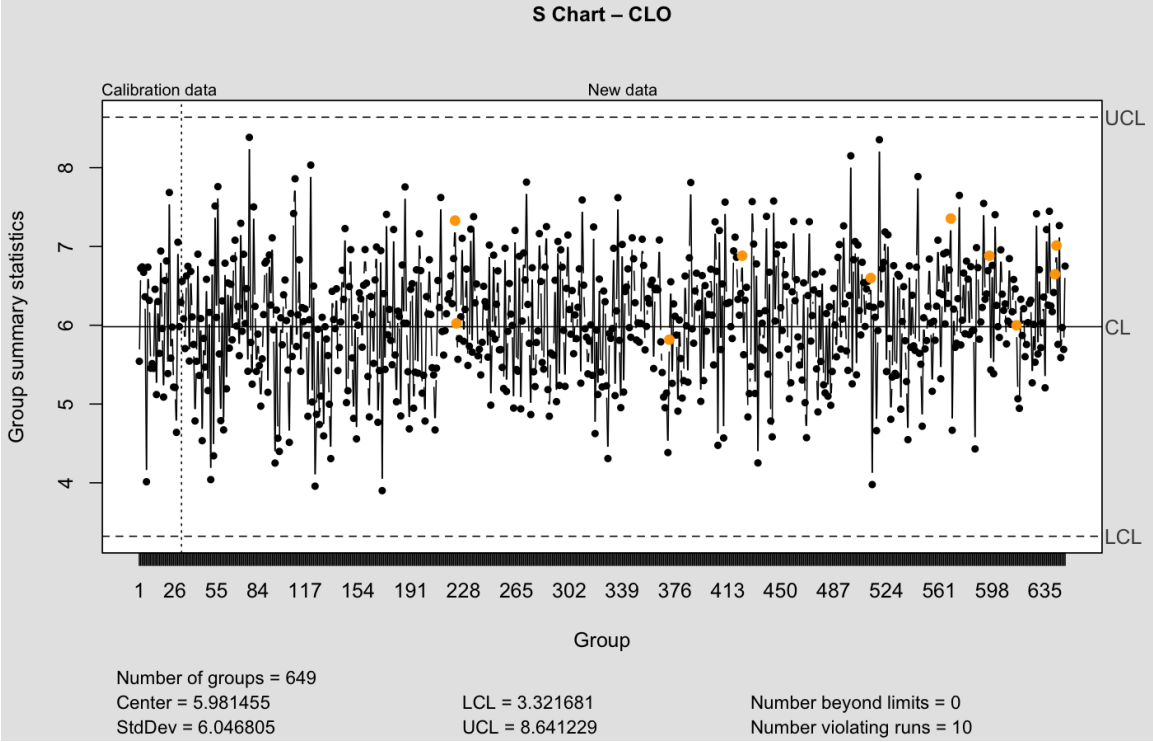


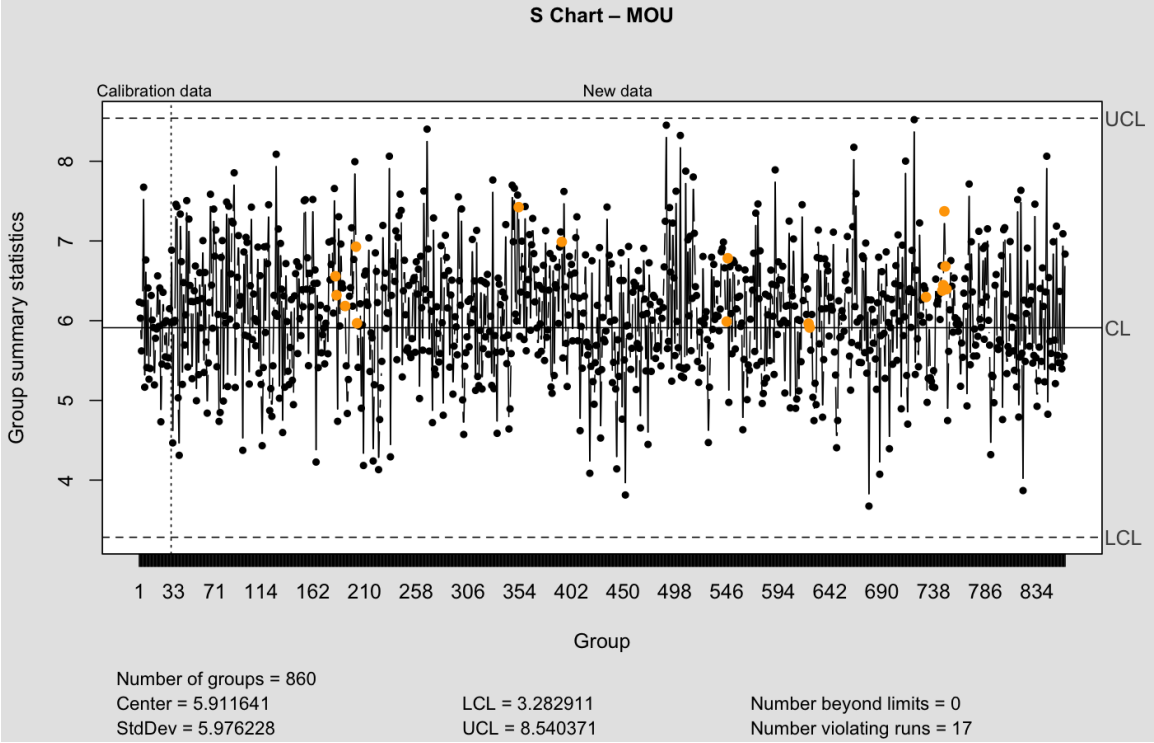

S Chart – LAP
Calibration data
New data
UCL
CL
LCL
Group summary statistics
4
5
6
7
8
9
1 17 36 55 74 93 114 138 162 186 210 234 258 282 306 330 354 378 402
Group
Number of groups = 425
Center = 6.117222
StdDev = 6.184055
LCL = 3.397077
UCL = 8.837368
Number beyond limits = 0
Number violating runs = 1

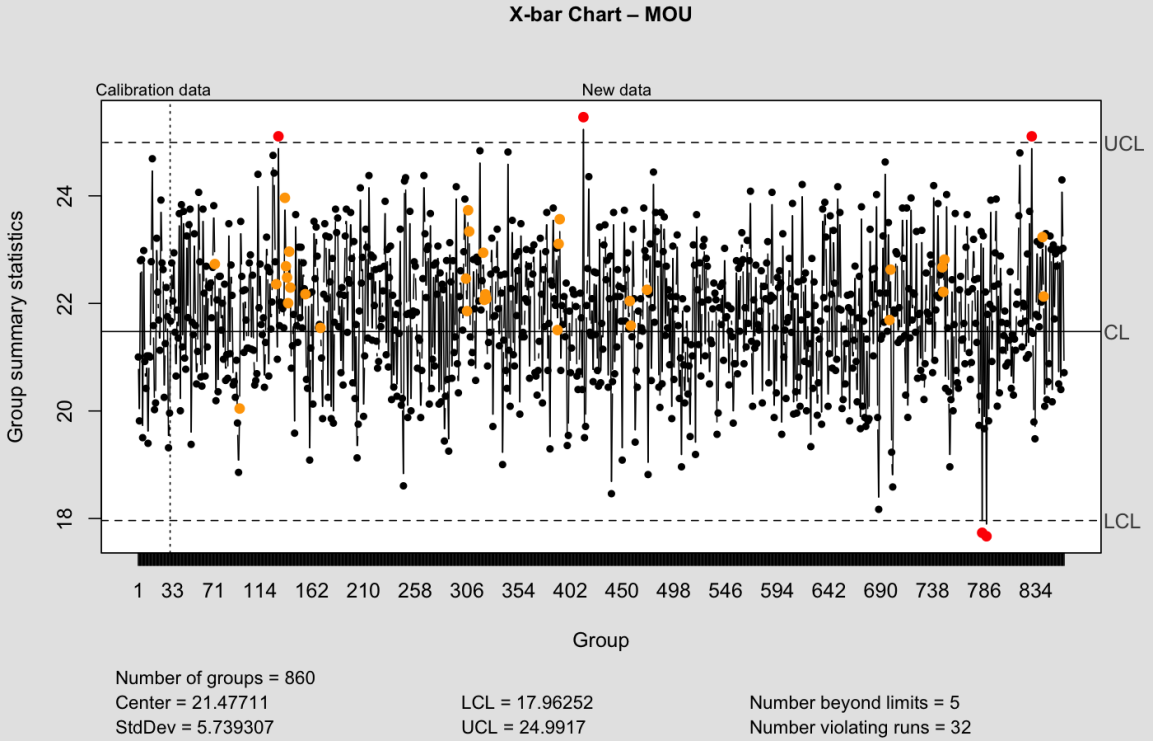

X-bar Chart – KEY
Calibration data
New data
UCL
CL
LCL
Group summary statistics
18
20
22
24
1 29 62 95 132 174 216 258 300 342 384 426 468 510 552 594 636 678 720
Group
Number of groups = 746
Center = 21.37536
StdDev = 5.540471
LCL = 17.98253
UCL = 24.7682
Number beyond limits = 9
Number violating runs = 18

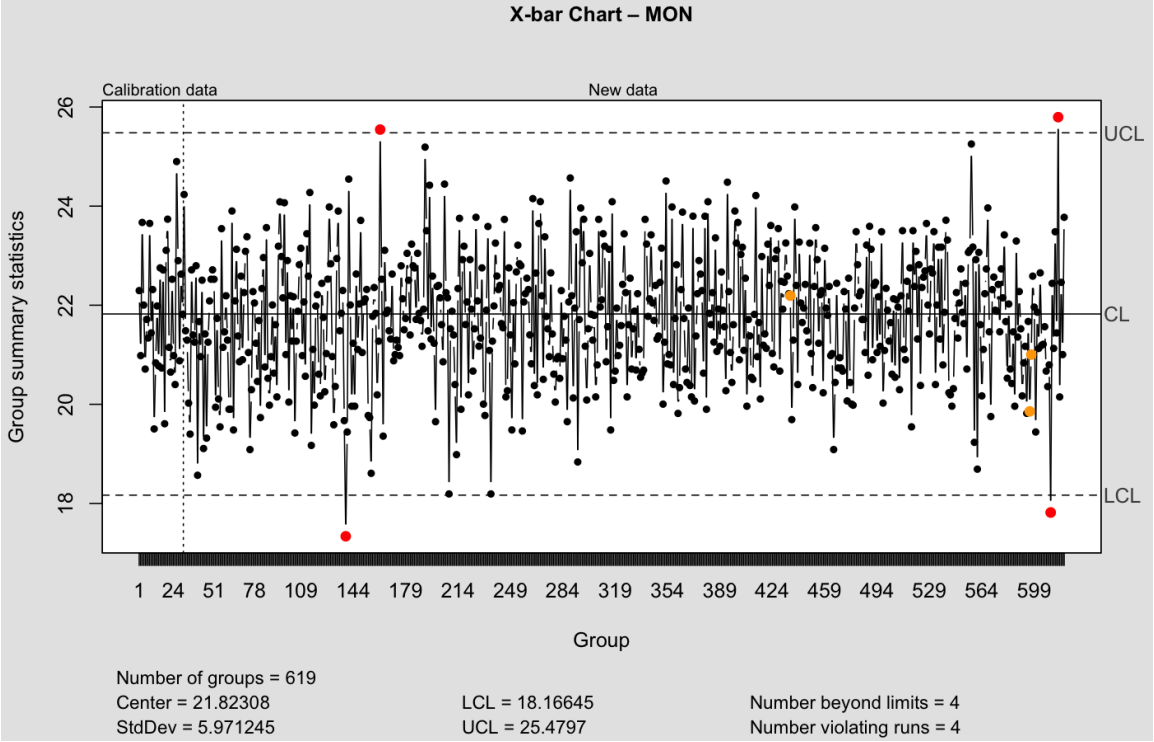

S Chart – KEY
Calibration data
New data
UCL
CL
LCL
Group summary statistics
4
5
6
7
8
1 29 62 95 132 174 216 258 300 342 384 426 468 510 552 594 636 678 720
Group
Number of groups = 746
Center = 5.883382
StdDev = 5.94766
LCL = 3.267218
UCL = 8.499546
Number beyond limits = 0
Number violating runs = 19

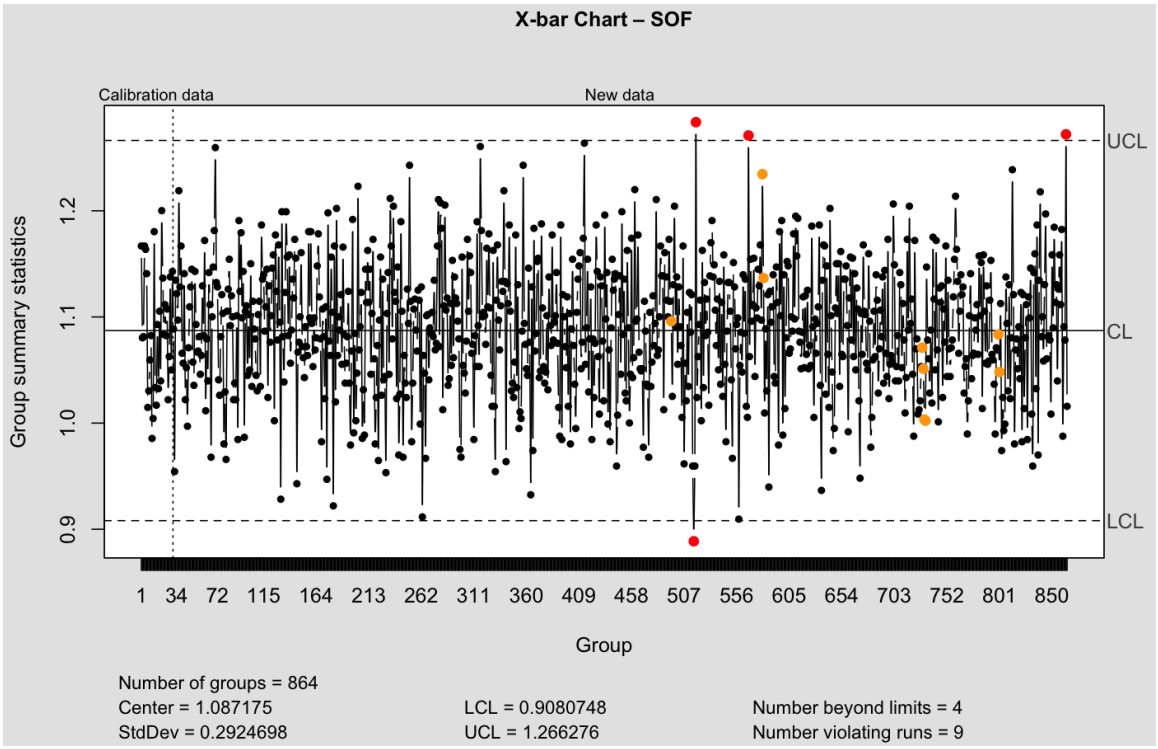

X-bar Chart – CLO
Calibration data
New data
UCL
CL
LCL
Group summary statistics
18
20
22
24
26
1 26 55 84 117 154 191 228 265 302 339 376 413 450 487 524 561 598 635
Group
Number of groups = 649
Center = 21.77364
StdDev = 5.564365
LCL = 18.36618
UCL = 25.18111
Number beyond limits = 8
Number violating runs = 9

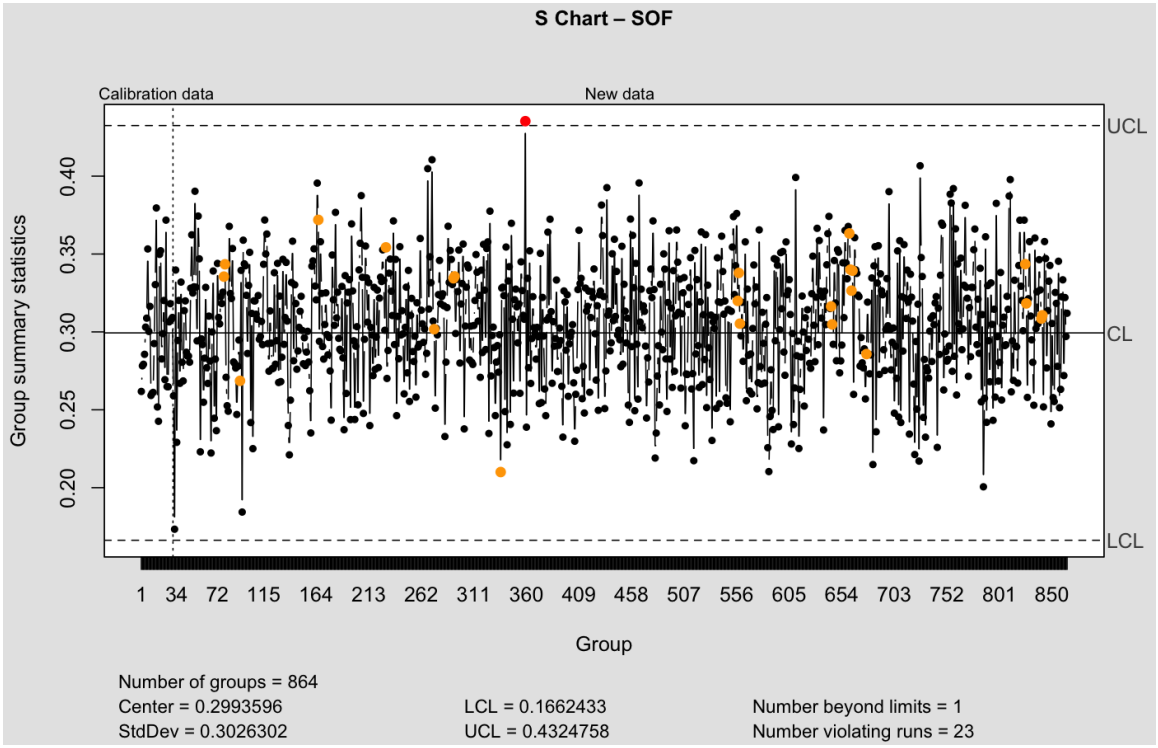

S Chart – CLO
Calibration data
New data
UCL
CL
LCL
Group summary statistics
4
5
6
7
8
1 26 55 84 117 154 191 228 265 302 339 376 413 450 487 524 561 598 635
Group
Number of groups = 649
Center = 5.981455
StdDev = 6.046805
LCL = 3.321681
UCL = 8.641229
Number beyond limits = 0
Number violating runs = 10

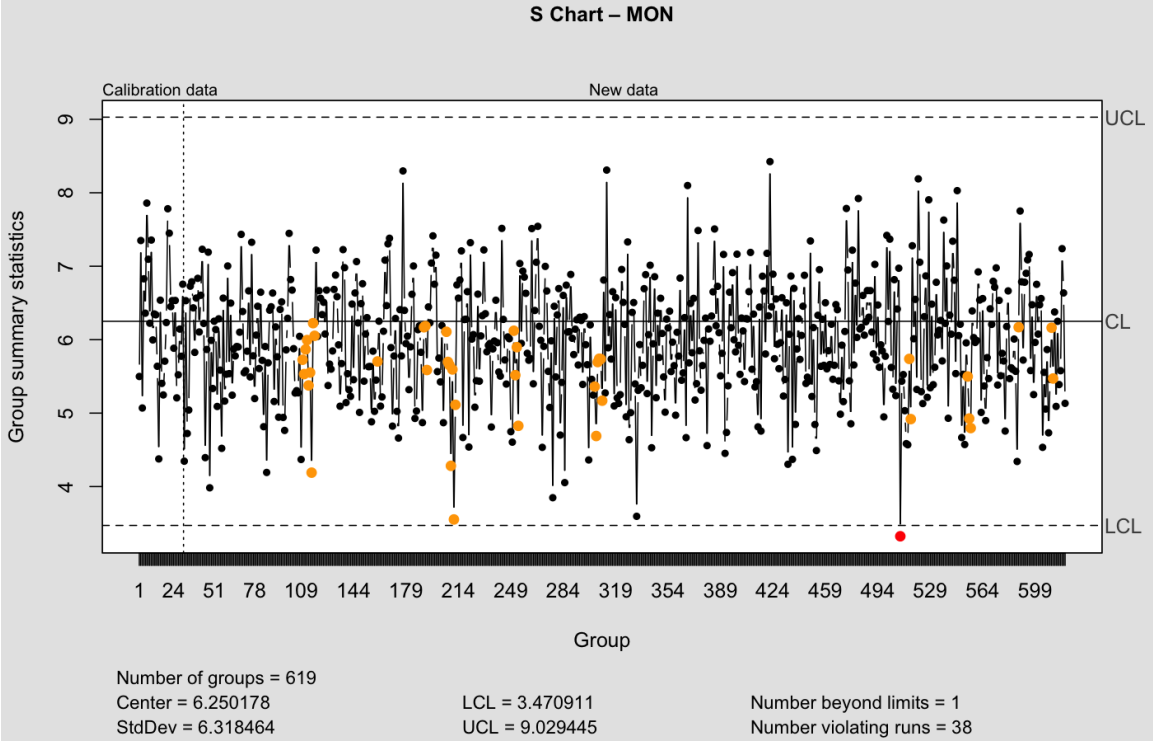

S Chart – MOU
Calibration data
New data
UCL
CL
LCL
Group summary statistics
4
5
6
7
8
1 33 71 114 162 210 258 306 354 402 450 498 546 594 642 690 738 786 834
Group
Number of groups = 860
Center = 5.911641
StdDev = 5.976228
LCL = 3.282911
UCL = 8.540371
Number beyond limits = 0
Number violating runs = 17


X-bar Chart – MOU
Calibration data
New data
UCL
CL
LCL
Group summary statistics
18
20
22
24
1 33 71 114 162 210 258 306 354 402 450 498 546 594 642 690 738 786 834
Group
Number of groups = 860
Center = 21.47711
StdDev = 5.739307
LCL = 17.96252
UCL = 24.9917
Number beyond limits = 5
Number violating runs = 32


X-bar Chart – MON
Calibration data
New data
UCL
CL
LCL
Group summary statistics
18
20
22
24
26
1 24 51 78 109 144 179 214 249 284 319 354 389 424 459 494 529 564 599
Group
Number of groups = 619
Center = 21.82308
StdDev = 5.971245
LCL = 18.16645
UCL = 25.4797
Number beyond limits = 4
Number violating runs = 4


X-bar Chart – SOF
Calibration data
New data
UCL
CL
LCL
Group summary statistics
0.9
1.0
1.1
1.2
1 34 72 115 164 213 262 311 360 409 458 507 556 605 654 703 752 801 850
Group
Number of groups = 864
Center = 1.087175
StdDev = 0.2924698
LCL = 0.9080748
UCL = 1.266276
Number beyond limits = 4
Number violating runs = 9


S Chart – SOF
Calibration data
New data
UCL
CL
LCL
Group summary statistics
0.20
0.25
0.30
0.35
0.40
1 34 72 115 164 213 262 311 360 409 458 507 556 605 654 703 752 801 850
Group
Number of groups = 864
Center = 0.2993596
StdDev = 0.3026302
LCL = 0.1662433
UCL = 0.4324758
Number beyond limits = 1
Number violating runs = 23


S Chart – MON
Calibration data
New data
UCL
CL
LCL
Group summary statistics
4
5
6
7
8
9
1 24 51 78 109 144 179 214 249 284 319 354 389 424 459 494 529 564 599
Group
Number of groups = 619
Center = 6.250178
StdDev = 6.318464
LCL = 3.470911
UCL = 9.029445
Number beyond limits = 1
Number violating runs = 38

Capability Indices

Capability indices were calculated for each product type based on the first 1000 deliveries, with LSL = 0h and USL = 32h. The results are summarized below.

Product Type	Cp	Cpu	Cpl	Cpk
LAP	0.87	0.56	1.18	0.56
KEY	0.89	0.58	1.20	0.58
CLO	0.88	0.56	1.19	0.56
MOU	0.88	0.58	1.19	0.58
MON	0.85	0.55	1.15	0.55
SOF	17.93	34.64	1.22	1.22

Product Types with Cpk ≥ 1.0 are capable of meeting the VOC. SOF is capable. LAP, KEY, CLO, MOU and MON are below capability and require tighter control.

Process Control Issues

Rule A — One S-sample outside the +3σ control limits

Product Type	Total Samples Outside Limit	First 3 Samples	Last 3 Samples
CLO011	18	2, 8, 15	60, 61, 63
CLO012	19	68, 69, 71	119, 120, 128
CLO013	18	131, 132, 134	179, 182, 183
CLO014	17	191, 197, 198	249, 251, 252
CLO015	14	261, 266, 267	290, 297, 307

Rule B — Longest consecutive samples within ±1σ (good control)

Product Type	Longest Stable Run (Samples)
CLO011	64
CLO012	63
CLO013	61
CLO014	63
CLO015	65

Rule C — Four consecutive X-bar samples outside +2σ

No samples were found that meet this criterion.

Interpretation

Delivery performance remains within acceptable variation limits overall, but several product types exhibit capability issues (Cpk < 1.0) and occasional deviations beyond ±2σ and ±3σ. SPC monitoring is recommended to detect deviations early. Prioritize process adjustments for LAP, KEY, CLO, MOU and MON to reduce spread and improve centering.

Probability of Errors

Type 1:

A: 0.001349898

B: 0.0078125

C: 2.678772e-07

Type 2:

Beta = 0.8411783

Power = 0.1588217

Question 5: Optimize Profit and Reliability

% clients with reliable service (≤5 min)	100.0%
Estimated demand per day	547.95 customers/day
Optimal number of baristas (profit max, c≥2)	2
Expected customers served per day	547.95
Expected daily profit	R 14438.36

Question 6

The hypotheses are formulated as follows:

- H_0 (null hypothesis): There is no significant difference in average delivery time between 2022 and 2023.
- H_1 (alternative hypothesis): There is a significant difference in average delivery time between 2022 and 2023.

Anova showed the p-value to be $0.021 < 0.05$. This indicates a statistically significant difference in mean delivery times between 2022 and 2023. The boxplot revealed slightly higher median delivery times and a wider spread in 2023. This implies that the process became less stable year by year.

Question 7

7.1

If we experience problems when there are less than 15 people at work, we will experience problems for 31 out of 397 days. Thus, for 366 days we will experience reliable service. For 92% of 397 days, we will experience reliable service.

7.2

P of reliability = 0.92

expected number of unreliable days in a year: $0.08 \times 365 = 29.2$

$\therefore$ loss = $20\,000\,(29.2) = R584\,000$.

with 1 employee:

$\frac{6}{397} \times 365 = 5.51$ days

$\therefore$ loss = $20000\,(5.51) + 300000 = R410200$

net gain = 584000 - 410200

= R 173800 / year

with 2 employees:

$\frac{1}{397} \times 365 = 0.92$ days

$\therefore$ loss = $20000\,(0.92) + 600000 = R618400$

net gain = 584000 - 618400

= -R34400 / year

$\therefore$ The optimal number of employees to add is 1.