ECSA Project

Quality Assurance 344

NR Bruwer (25875914@sun.ac.za)

24 October 2025

Table of Contents

Table of Figures

List of Tables

Introduction

This project is part of the ECSA 2025 Quality Assurance and Data Analysis module and demonstrates how statistical and analytical techniques can be applied to real business data. The aim was to explore how data can provide insights into operational performance, identify trends, and support decision-making. The work began with descriptive and exploratory analyses of sales, customer, and product data to understand the company's overall performance. It then progressed to more advanced methods, including Statistical Process Control (SPC) charts, process capability assessments, and ANOVA tests, to evaluate consistency and variability in delivery times and product performance. Additional sections addressed data validation, the risks of Type I and Type II errors, and optimisation models to improve staffing levels and profitability. Each step in the project highlights how combining quality assurance and data analytics can guide practical improvements in business operations.

Part 1: Descriptive Statistics

1.1. Average Income by City

The average customer income by city was calculated to identify cities where people have more money to spend.

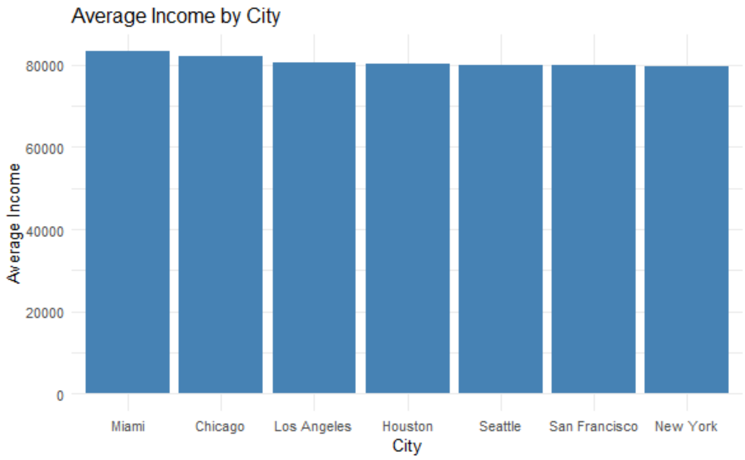


Figure 1: Average Income by City

Figure 1 shows the average income for each city. Miami and Chicago have the highest average incomes, while New York and San Francisco have lower average incomes. The focus of marketing should be on the cities with the most money to spend – Miami and Chicago.

1.2. Correlation Between Age and Income

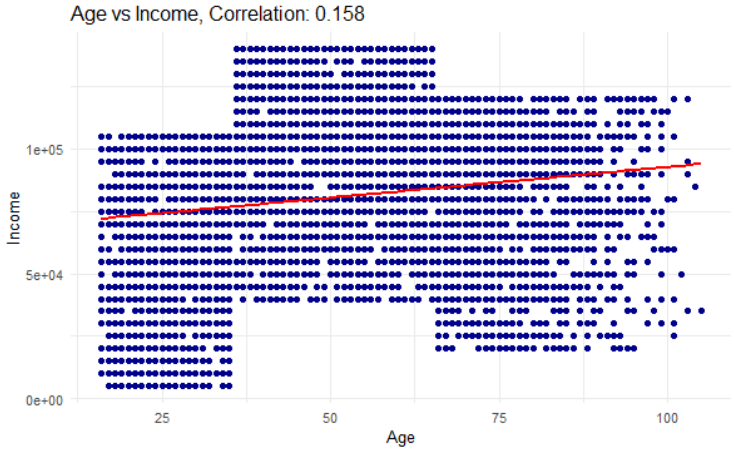


Figure 2: Age vs Income

Figure 2 shows a scatter plot of Age vs. Income. There is a small positive correlation (r ≈ 0.158) between the data, which means that there is a trend, but only slightly. As a person's age increases, their income also increases, but it varies a lot. Thus, age is not a strong predictor of income in this dataset.

1.3. High-Income Customers by City

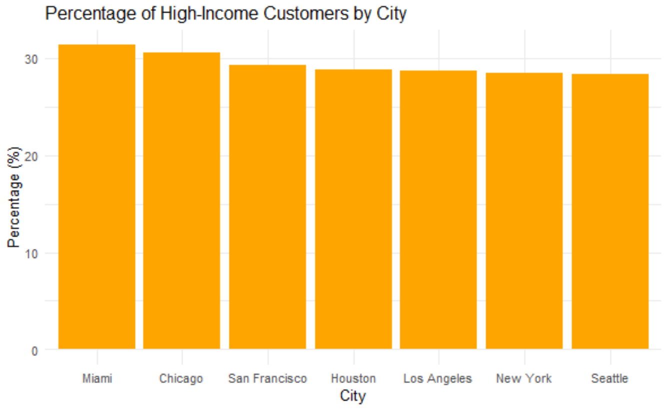


Figure 3: Percentage of High-Income Customers by City

In figure 3, high-income customers were defined as those that earn more than R80 000 (the top 25%). Miami has the most high-income customers, followed by Chicago. Thus, sales campaigns that target high-income individuals should focus on these cities.

1.4. Income Differences by Gender

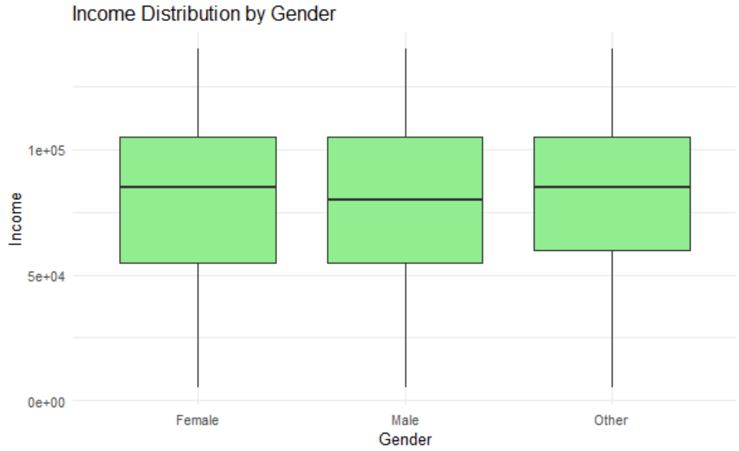
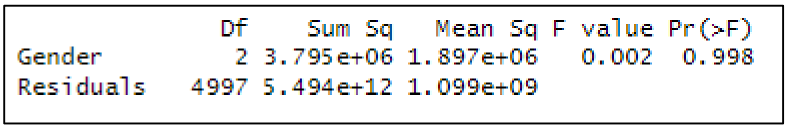


Figure 4: Income Distribution by Gender



```
              Df    Sum Sq   Mean Sq F value Pr(>F)
Gender         2 3.795e+06 1.897e+06   0.002  0.998
Residuals   4997 5.494e+12 1.099e+09
```

Table 1: ANOVA Results for Income Differences by Gender

Figure 4 shows boxplots of income by gender. The median income (bold line in the box) is slightly lower for males, but very similar across genders. There is no real difference between genders when it comes to earnings.

As seen in Table 1, the Gender factor has 2 degrees of freedom, which is expected, since we are comparing three groups (Male, Female, Other). The residuals have 4997 degrees of freedom, which represents the total number of observations minus the number of groups. The sum of squares for Gender is very small relative to the residual sum of squares. This tells us that the variation in income explained by gender is very small in comparison to the variation within the groups. Gender explains almost none of the variation in income (F-value is 0.002; p-value is

0.998). there is no statistically significant difference in income among the gender groups. So, targeting by gender is not a smart move when trying to reach people with more spending power.

1.5. Customer Distribution Across Age Brackets

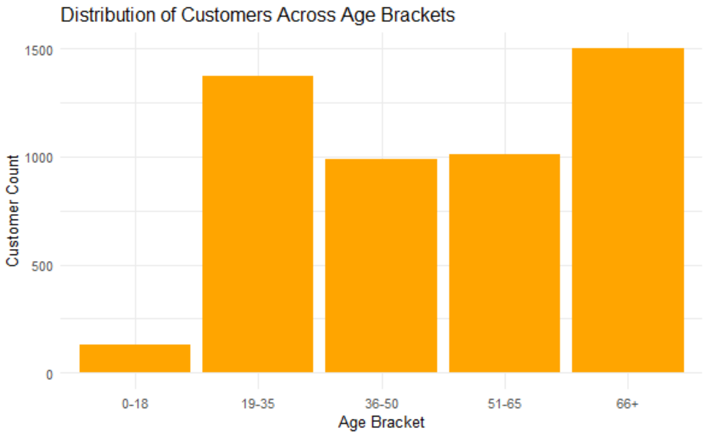


Figure 5: Distribution of Customers Across Age Brackets

The distribution of customers across age brackets is shown in Figure 5. Most customers are over 66 years old, so the customer base is older adults. Products and services should be designed according to this group's needs and preferences.

Part 2: Basic Data Analysis

2.1. Distribution of Order Quantities

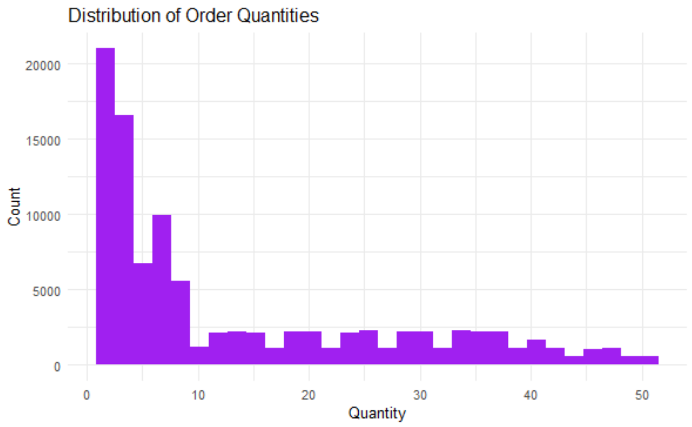


Figure 6: Distribution of Order Quantities

The histogram in Figure 6 shows that order quantities follow a right-skewed, exponential distribution. This means that most of the orders are relatively small, while some orders are medium and large. The large orders are still very important, even if they occur less, because they still make up a significant share of the total sales volume.

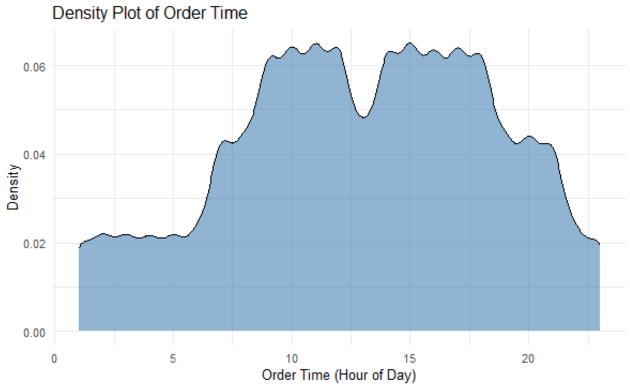


Figure 7: Density Plot of Order Time

As shown in Figure 7, there are two peaks in the density plot of order times. This tells us that the customers tend to put in their orders either late morning or early afternoon.

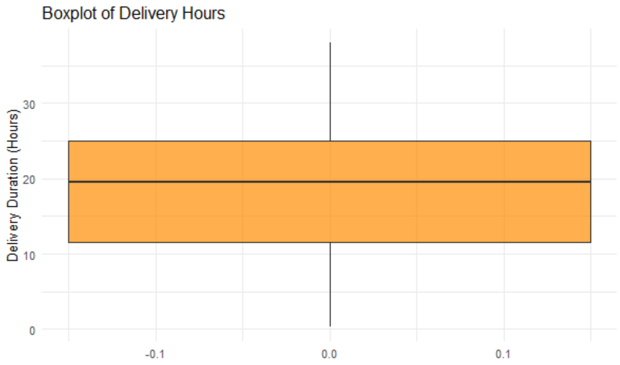


Figure 8: Boxplot of Delivery Hours

The boxplot in Figure 8 summarizes the delivery time data. The thick horizontal line within the box represents the median delivery time (approximately 20 hours). This tells us that half of all the deliveries are completed in under 20 hours, while the other half takes longer.

The box itself covers the middle 50% of delivery times, which ranges from about 11 hours (Q1) to 25 hours (Q3). Most deliveries fall in this range, which shows medium variation in how long deliveries take to be completed.

The whiskers of the boxplot extend from roughly 0 to 38 hours, which suggests that while most deliveries take place within a typical range, a few of them take shorter or longer. Though, no extreme outliers were noted.

The delivery process is consistent in general, with some variability that can be due to occasional traffic or operational related problems.

2.2. Monthly Sales Trend

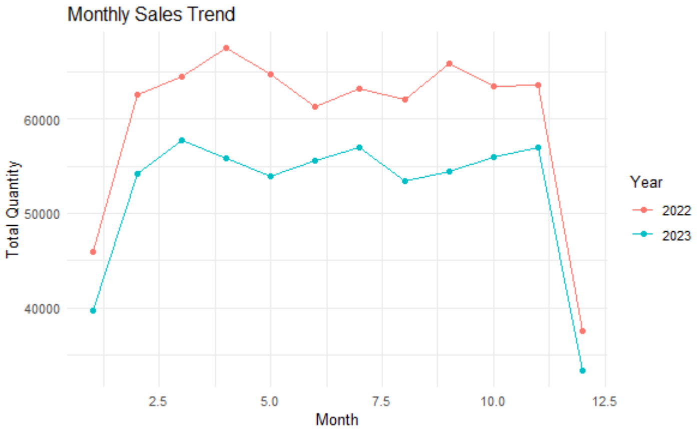
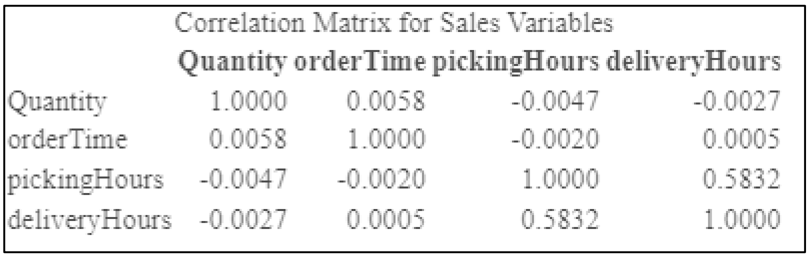


Figure 9: Monthly Sales Trend

Figure 9 shows the monthly sales totals for 2022 and 2023. Sales peak in February-March, July, and October-November, which may be due to seasonal patterns, promotions, or holidays. To be prepared for these high-demand periods, planning of inventory levels and staffing should be done early on. There is a noticeable decline in sales towards the end of the year (November-December). This can be because of slower customer activity or company closures due to it being a popular holiday period, or incomplete data for the final months of the year.

2.3. Correlation Between Delivery and Picking Times



Correlation Matrix for Sales Variables

	Quantity	orderTime	pickingHours	deliveryHours
Quantity	1.0000	0.0058	-0.0047	-0.0027
orderTime	0.0058	1.0000	-0.0020	0.0005
pickingHours	-0.0047	-0.0020	1.0000	0.5832
deliveryHours	-0.0027	0.0005	0.5832	1.0000

Table 2: Correlation Matrix for Sales Variables

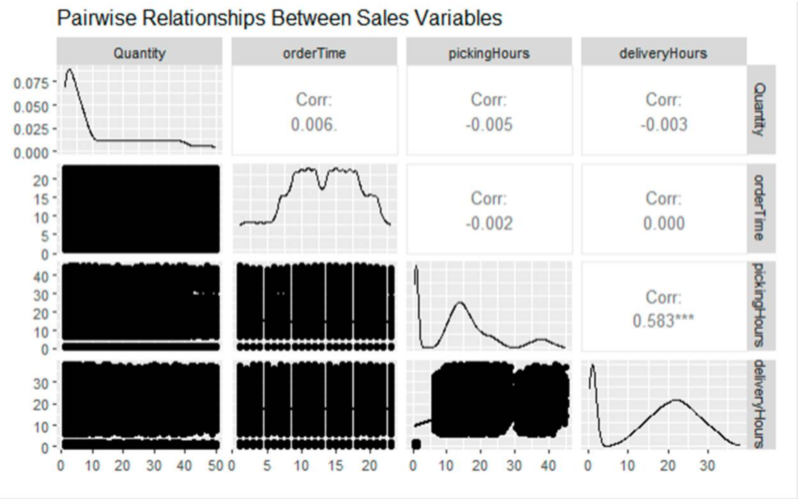
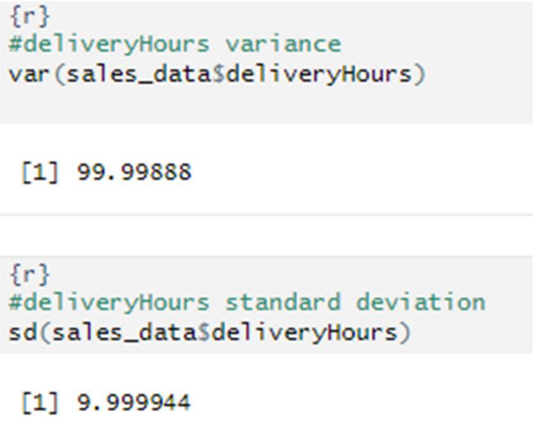


Figure 10: Pairwise Relationship Between Sales Variables

The correlation matrix (Table 2) and scatterplot matrix (Figure 10) show that most variables do not have strong linear relationships with each other, except for picking hours and delivery hours. The correlations between order quantity and other variables, such as order time (0.0058), picking hours (-0.0047), and delivery hours (-0.0027) are all very close to zero. This tells us that the number of items ordered does not really influence when an order is placed, how long it takes to pick, or how long it takes to deliver. The correlations between order time and picking hours (-0.0020), and between order time and delivery hours (0.0005) show almost no relationship at all, which means that operational times stay consistent no matter when orders are placed. However, a correlation of 0.5832 indicates a moderate positive relationship between picking hours and delivery hours. This means that orders that take longer to pick also tend to take longer to deliver. This relationship can be seen in Figure 10.

2.4. Variability of Delivery Hours

deliveryHours has the following variance and standard deviation:



```
{r}
#deliveryHours variance
var(sales_data$deliveryHours)
```

```
 [1] 99.99888
```

```
{r}
#deliveryHours standard deviation
sd(sales_data$deliveryHours)
```

```
 [1] 9.999944
```

Figure 11: Variance and SD of deliveryHours

The variance of delivery hours is approximately 99.999, which means delivery times vary quite a bit around the average. This indicates that not all deliveries are completed within a consistent timeframe. This variability can be measured by the standard deviation, which is the square root of the variance. The standard deviation of deliveryHours is approximately 10 hours. This means that on average the delivery times differ by almost 10 hours from the mean.

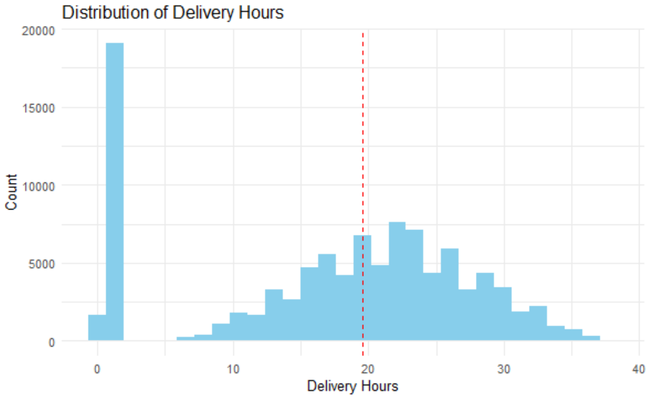


Figure 12: Distribution of Delivery Hours

In Figure 12, the histogram of delivery hours includes a dotted red line marking the median delivery time. This helps visualize where most deliveries are centered within the overall distribution.

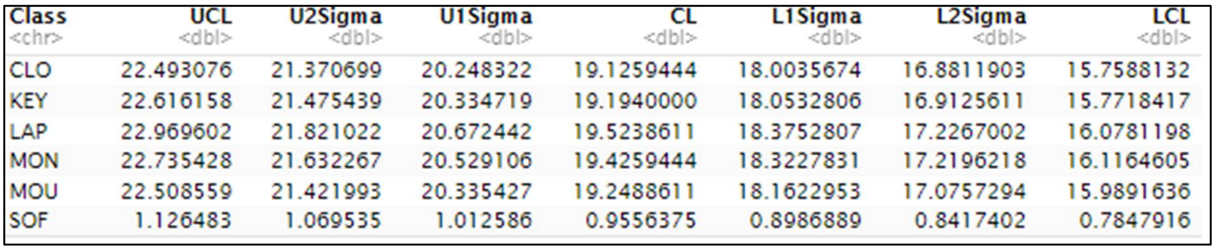
Part 3: Statistical Process Control (SPC)

3.1. Introduction

Statistical Process Control (SPC) uses statistical methods to monitor and control processes (ASQ, 2022). The sales delivery data were ordered chronologically and grouped into samples of 24 deliveries per product. The first 30 samples for each product were used to establish control limits for X-bar and S charts. These initial SPC values are critical for monitoring the process in real time (samples 31 onwards).

3.2. Control Charts for the First 30 Samples

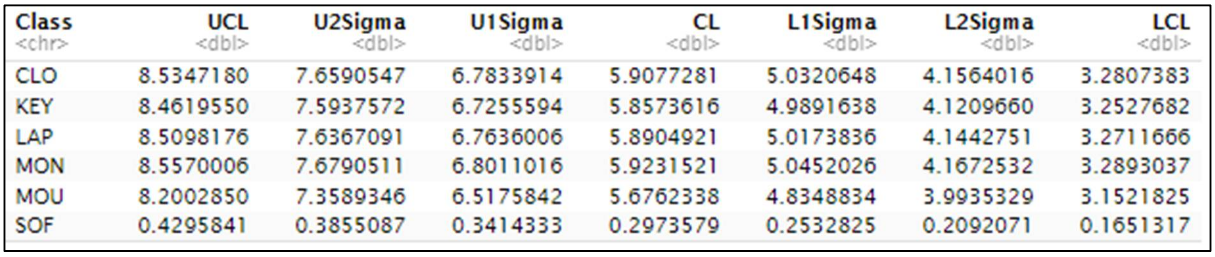
The X-chart tracks the average of the process over time. It plots the means of samples taken at regular intervals and is mostly used to monitor changes in the process variables.



Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
CLO	22.493076	21.370699	20.248322	19.1259444	18.0035674	16.8811903	15.7588132
KEY	22.616158	21.475439	20.334719	19.1940000	18.0532806	16.9125611	15.7718417
LAP	22.969602	21.821022	20.672442	19.5238611	18.3752807	17.2267002	16.0781198
MON	22.735428	21.632267	20.529106	19.4259444	18.3227831	17.2196218	16.1164605
MOU	22.508559	21.421993	20.335427	19.2488611	18.1622953	17.0757294	15.9891636
SOF	1.126483	1.069535	1.012586	0.9556375	0.8986889	0.8417402	0.7847916

Table 3: X Chart for 30 Samples

The S-chart is a type of control chart used to track the variability or standard deviation of a process, especially when sample sizes are bigger than five. Each point on the chart corresponds to the standard deviation of one sample.



Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
CLO	8.5347180	7.6590547	6.7833914	5.9077281	5.0320648	4.1564016	3.2807383
KEY	8.4619550	7.5937572	6.7255594	5.8573616	4.9891638	4.1209660	3.2527682
LAP	8.5098176	7.6367091	6.7636006	5.8904921	5.0173836	4.1442751	3.2711666
MON	8.5570006	7.6790511	6.8011016	5.9231521	5.0452026	4.1672532	3.2893037
MOU	8.2002850	7.3589346	6.5175842	5.6762338	4.8348834	3.9935329	3.1521825
SOF	0.4295841	0.3855087	0.3414333	0.2973579	0.2532825	0.2092071	0.1651317

Table 4: S Chart for 30 Samples

CLO

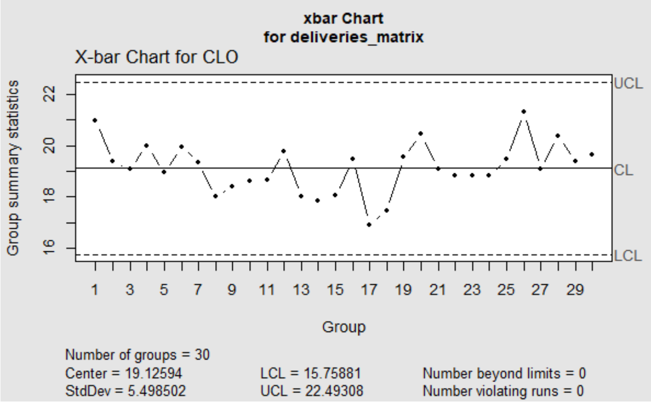


Figure 13: X Chart for CLO

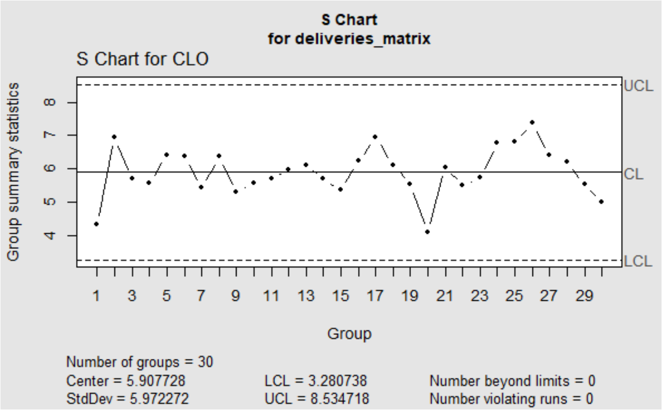


Figure 14: S Chart for CLO

KEY

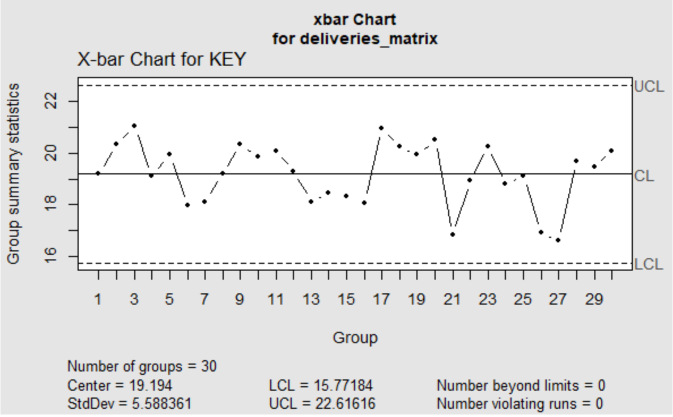


Figure 15: X Chart for KEY

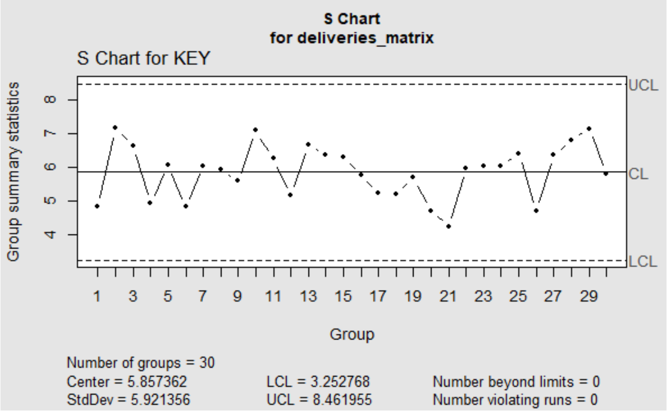


Figure 16: S Chart for KEY

LAP

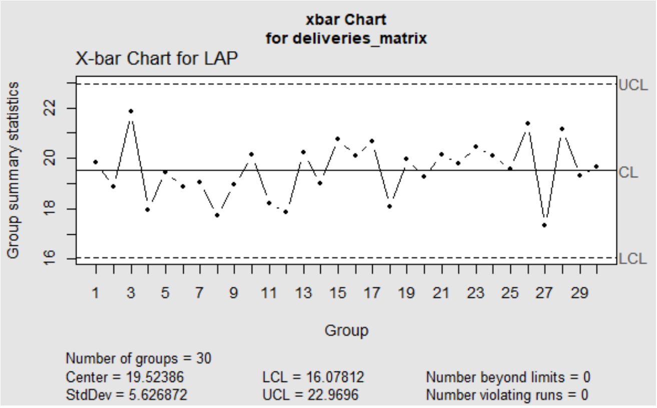


Figure 17: X Chart for LAP

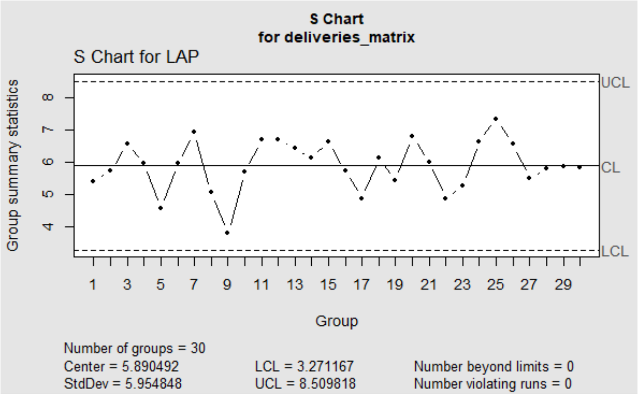


Figure 18: S Chart for LAP

MON

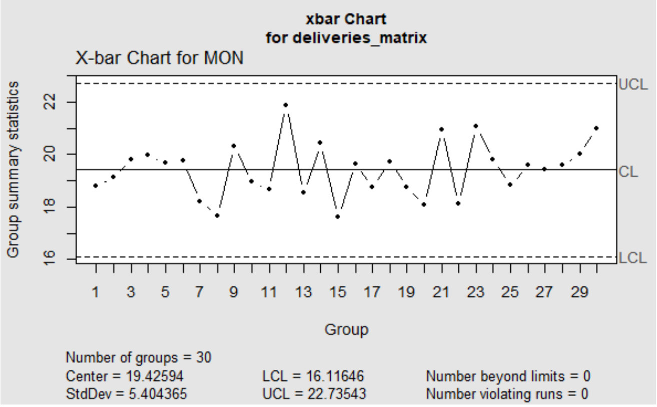
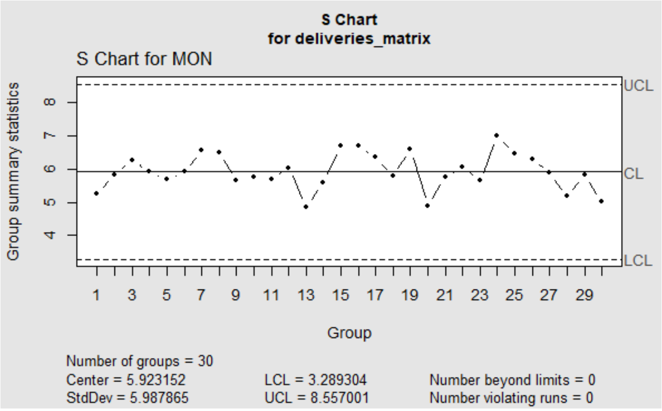


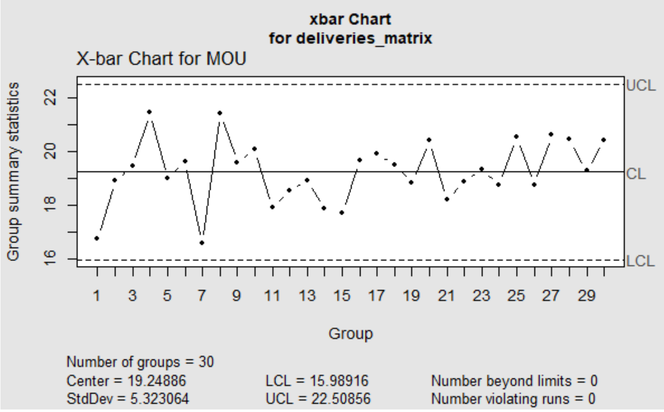
Figure 19: X Chart for MON



```
Number of groups = 30
Center = 5.923152        LCL = 3.289304        Number beyond limits = 0
StdDev = 5.987865        UCL = 8.557001        Number violating runs = 0
```

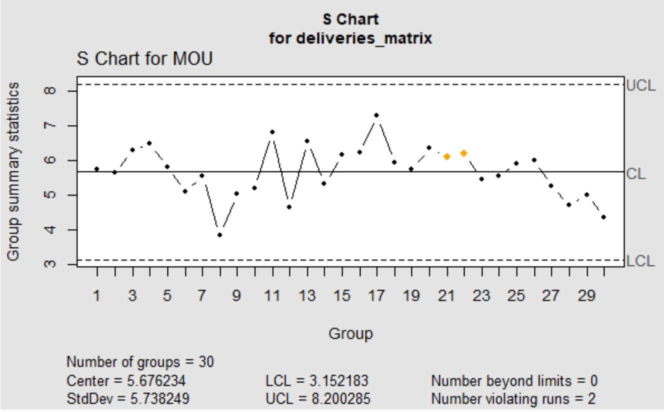
Figure 20: S Chart for MON

MOU



```
Number of groups = 30
Center = 19.24886        LCL = 15.98916        Number beyond limits = 0
StdDev = 5.323064        UCL = 22.50856        Number violating runs = 0
```

Figure 21: X Chart for MOU



```
Number of groups = 30
Center = 5.676234        LCL = 3.152183        Number beyond limits = 0
StdDev = 5.738249        UCL = 8.200285        Number violating runs = 2
```

Figure 22: S Chart for MOU

SOF

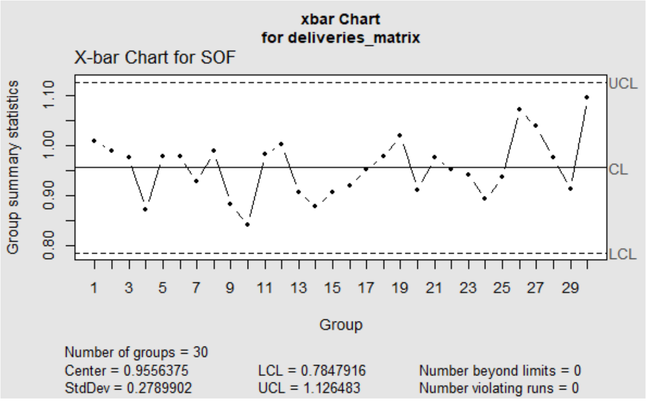


Figure 23: X Chart for SOF

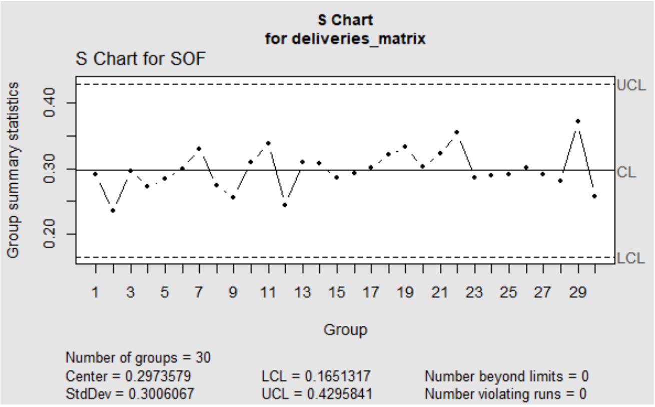


Figure 24: S Chart for SOF

Evaluation of Control Charts

As can be seen in the S Charts above (Figures 14, 16, 18, 20, 22, and 24) the graphs show that the process variabilities are under control, as all points fall within the control limits. These charts haves some variation as some points are close to the centreline while others are closer to the limits. The X Charts above (Figures 13, 15, 17, 19, 21, and 23) also display some variation, and process variabilities that are under control. Even though there are variabilities in all of these figures, the charts still remain in control, which indicates that the processes are stable and not exceeding any operational limits.

3.2. Control Charts for the Rest of the Data

For the control charts for the rest of the data, the different dots and lines mean the following:

- Red dots: samples which are out of control.
- Black dots: samples which are in control.
- Yellow dots: points that are within control limits but are part of a warning zone.
- Dashed lines: the calculated upper and lower control limits (UCL, LCL).
- Solid line: centreline.

CLO

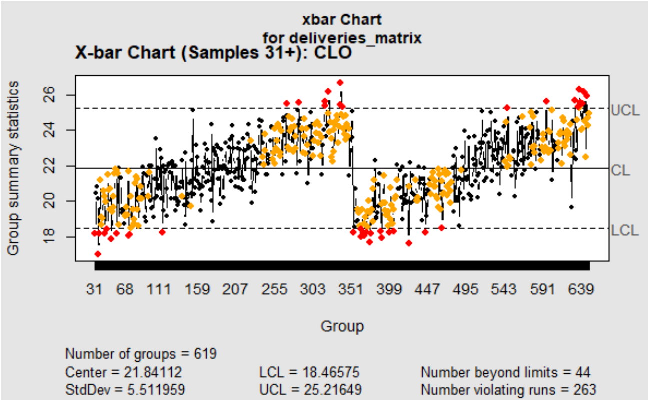


Figure 25: X Chart for CLO (samples 31+)

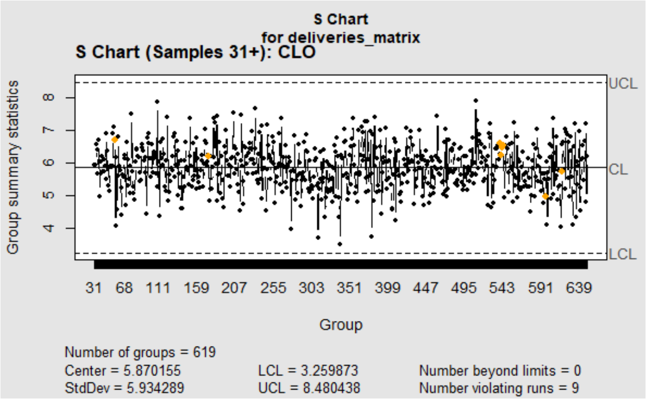


Figure 26: S Chart for CLO (samples 31+)

KEY

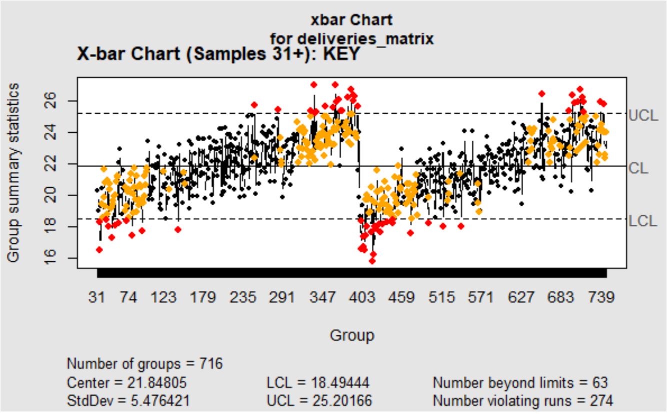


Figure 27: X Chart for KEY (samples 31+)

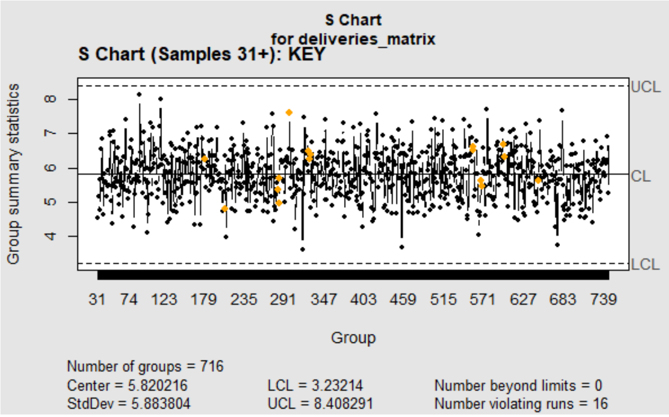


Figure 28: S Chart for KEY (samples 31+)

LAP

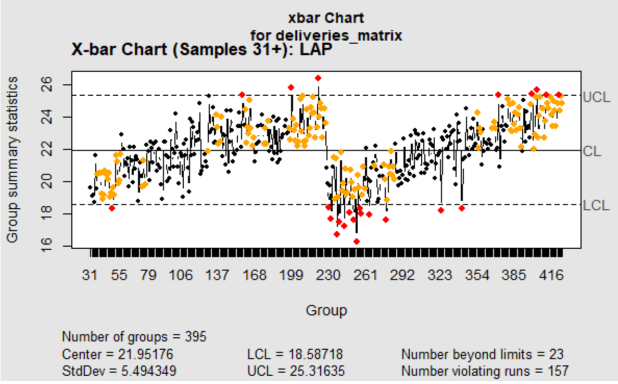


Figure 29: X Chart for LAP (samples 31+)

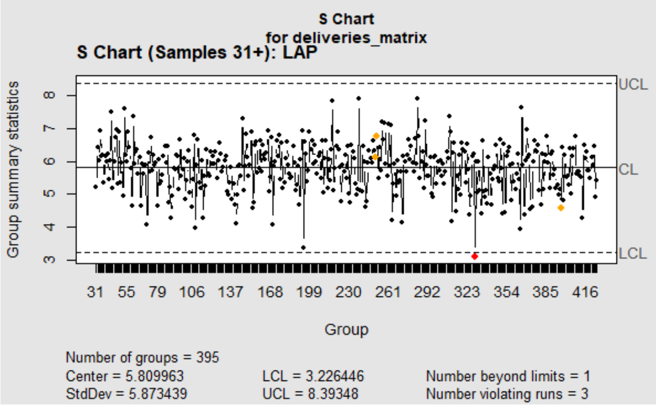


Figure 30: S Chart for LAP (samples 31+)

MON

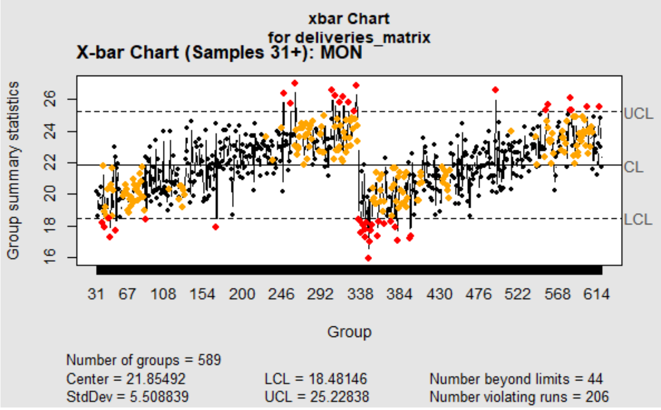


Figure 31: X Chart for MON (samples 31+)

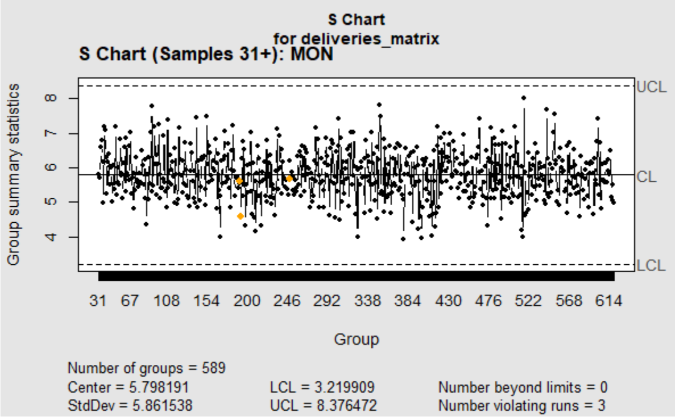


Figure 32: S Chart for MON (samples 31+)

MOU

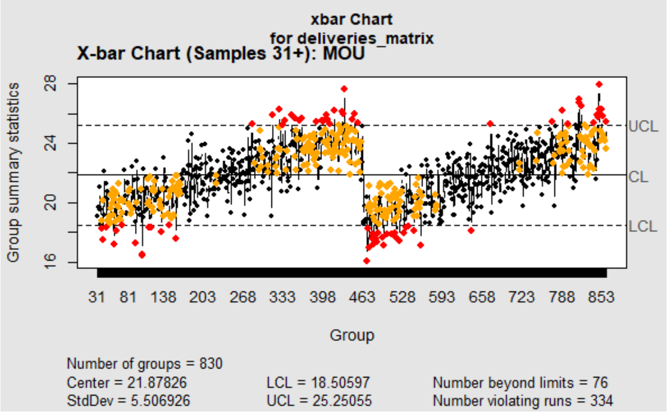


Figure 33: X Chart for MOU (samples 31+)

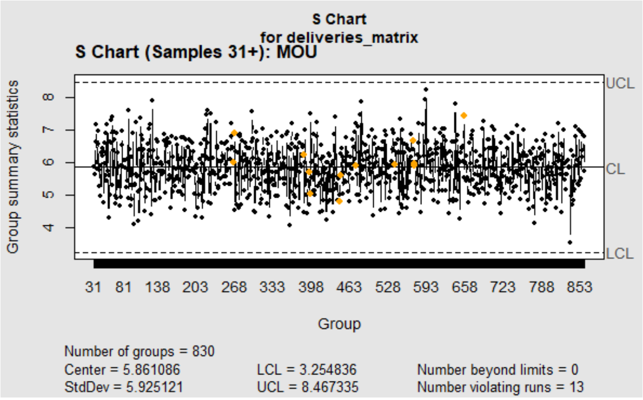


Figure 34: S Chart for MOU (samples 31+)

SOF

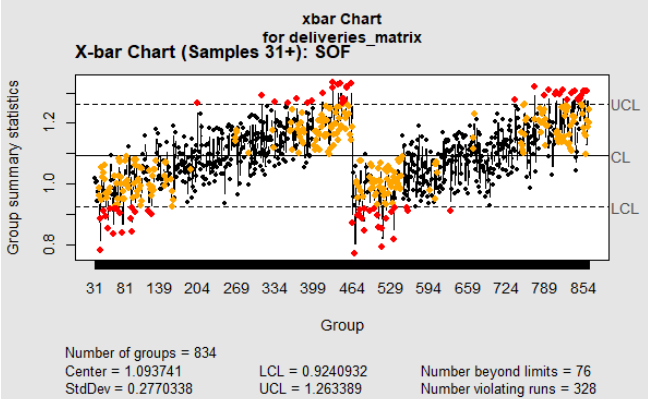


Figure 35: X Chart for SOF (samples 31+)

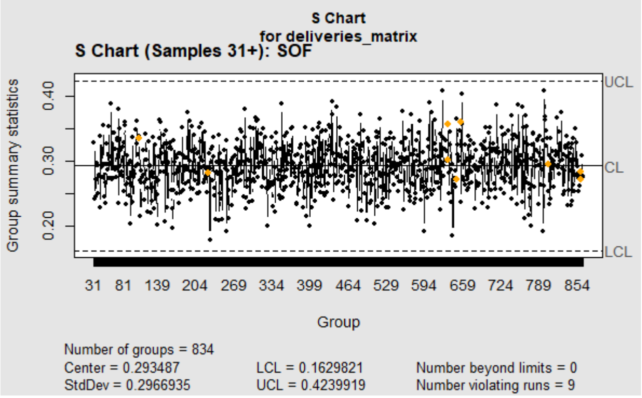


Figure 36: S Chart for SOF (samples 31+)

Evaluation of Control Charts

After looking at the X Charts for the rest of the data, it is clear that they all have similar patterns in their data: high numbers of sample means beyond control limits (as high as 76), and high numbers of samples violating runs (as high as 334). The X Charts clearly contain step-changes and long stretches that are consistently above or below the centreline. This indicates that the process means for the different products are not stable. There have been multiple shifts in average delivery time over the monitoring period.

After looking at the S Charts for the rest of the data, it is clear that they also have similar patterns in their data: only small numbers of violating runs (biggest is 16), and almost no values beyond the UCL or LCL (only LAP has 1 run beyond limits – the rest have none). This indicates that the spread of individual observations within each sample is relatively consistent through time.

3.3. R-Charts for every product type

For the R-Charts for the data, the different dots and lines mean the following:

- Red dots: samples which are out of control.
- Black dots: samples which are in control.
- Yellow dots: points that are within control limits but are part of a warning zone.
- Red dashed line: the calculated upper and lower control limits (UCL, LCL).
- Blue dashed line: centreline (R-bar).

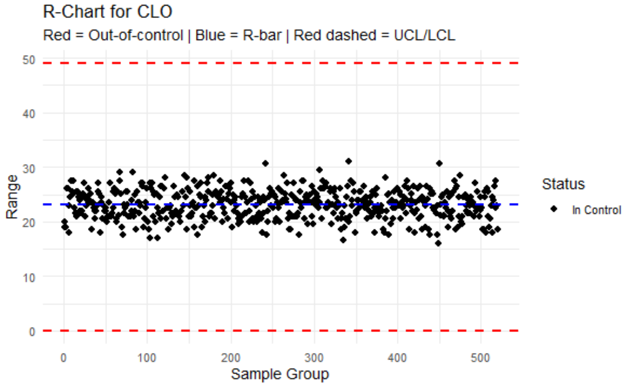


Figure 37: R Chart for CLO

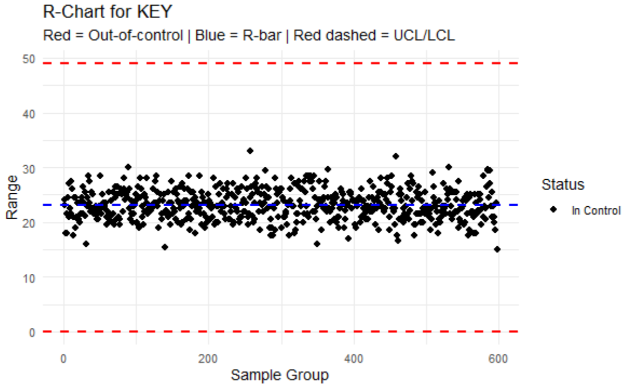


Figure 38: R Chart for KEY

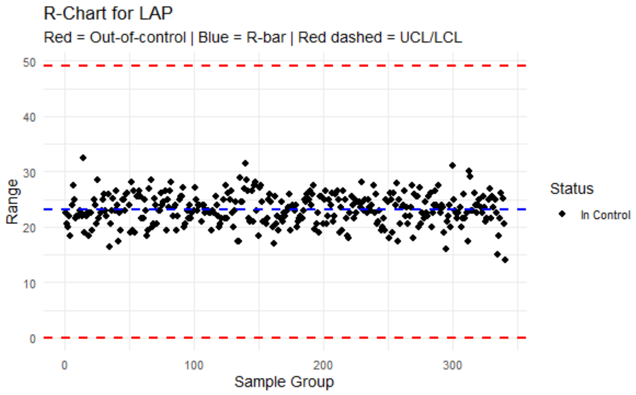


Figure 39: R Chart for LAP

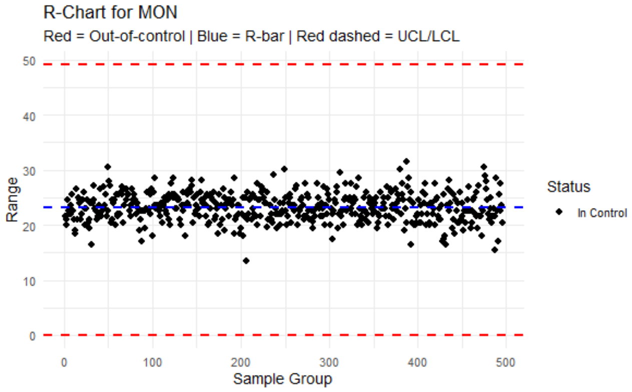


Figure 40: R Chart for MON

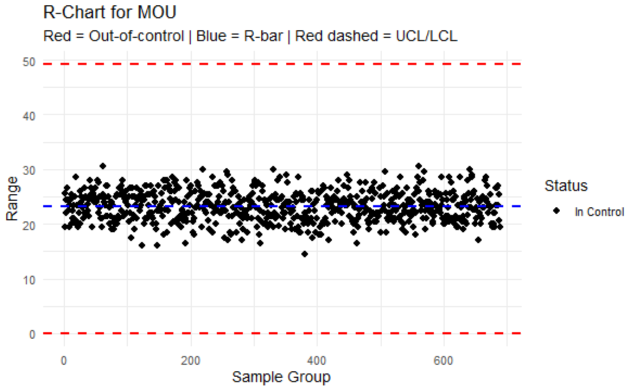


Figure 41: R Chart for MOU

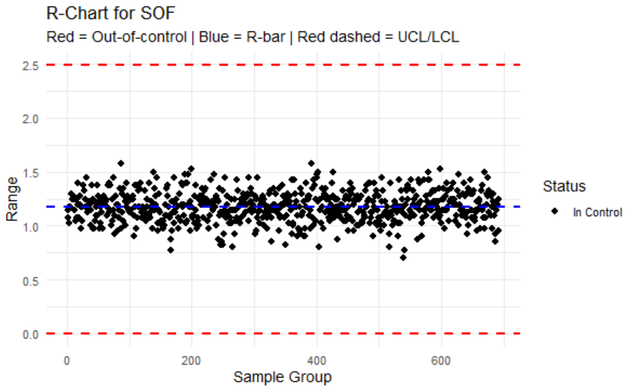
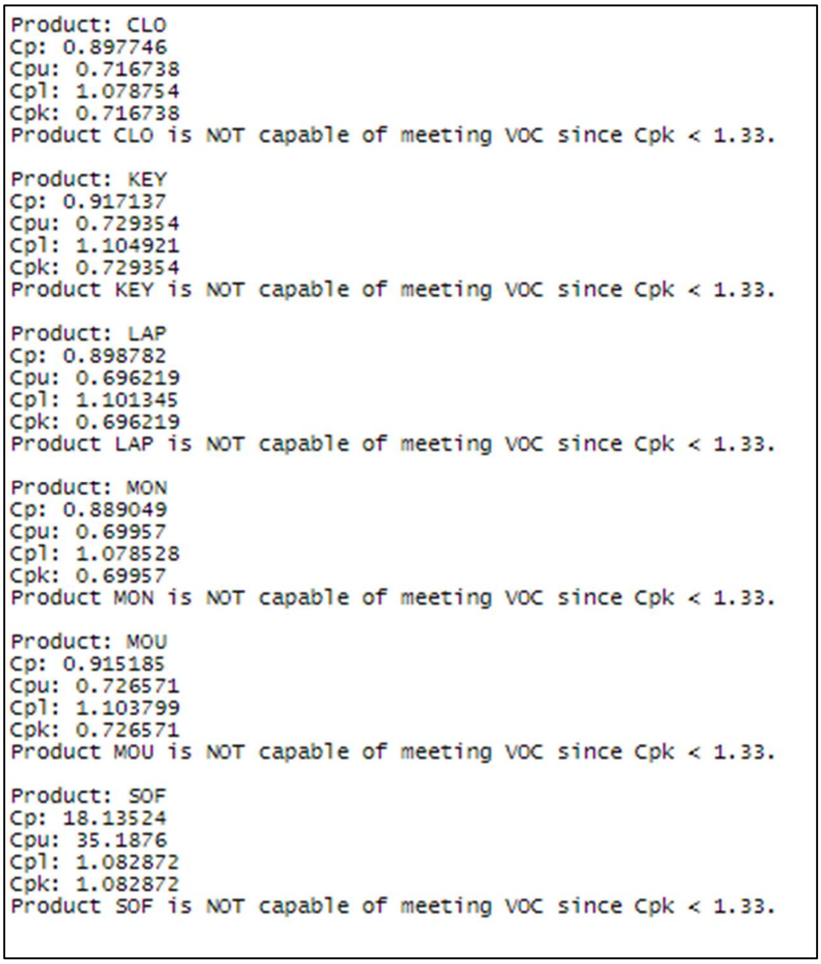


Figure 42: R Chart for SOF

Figures 37 to 42 show the respective R Charts for each product type. After carefully studying these charts, it is clear that they all have similar patterns in their data: all of the points are well within the upper and lower control limits (UCL and LCL), and there are no clear trends, cycles or sudden spikes in the data. The data points all lie scattered along the centreline, which indicates that the variation within each sample remains consistent over time. Even though the X Charts show changes in the mean delivery times, the overall process variability (the spread of individual delivery times within each subgroup) has not changed a lot. This means that the day-to-day or batch-to-batch variation is under control for all product types.

3.4. The Process Capability Indices for Process Delivery Times

The minimum acceptable Cpk value is 1.33 for the Voice of Customer to be met. If the Cpk value is equal to or bigger than 1.33, the process is able to produce outputs that fall inside specified limits, with a very low chance of defects. However, if Cpk is smaller than 1.33, the process cannot consistently produce outputs inside certain limits, which means that the customer requirements (VOC) cannot be met. In this case, the VOC described the customer's expectations for deliveryTime. Looking at figure 43, it is clear that no product is meeting the VOC.



```
Product: CLO
Cp: 0.897746
Cpu: 0.716738
Cpl: 1.078754
Cpk: 0.716738
Product CLO is NOT capable of meeting VOC since Cpk < 1.33.

Product: KEY
Cp: 0.917137
Cpu: 0.729354
Cpl: 1.104921
Cpk: 0.729354
Product KEY is NOT capable of meeting VOC since Cpk < 1.33.

Product: LAP
Cp: 0.898782
Cpu: 0.696219
Cpl: 1.101345
Cpk: 0.696219
Product LAP is NOT capable of meeting VOC since Cpk < 1.33.

Product: MON
Cp: 0.889049
Cpu: 0.69957
Cpl: 1.078528
Cpk: 0.69957
Product MON is NOT capable of meeting VOC since Cpk < 1.33.

Product: MOU
Cp: 0.915185
Cpu: 0.726571
Cpl: 1.103799
Cpk: 0.726571
Product MOU is NOT capable of meeting VOC since Cpk < 1.33.

Product: SOF
Cp: 18.13524
Cpu: 35.1876
Cpl: 1.082872
Cpk: 1.082872
Product SOF is NOT capable of meeting VOC since Cpk < 1.33.
```

Figure 43: Process Capability Indices for Process Delivery Times

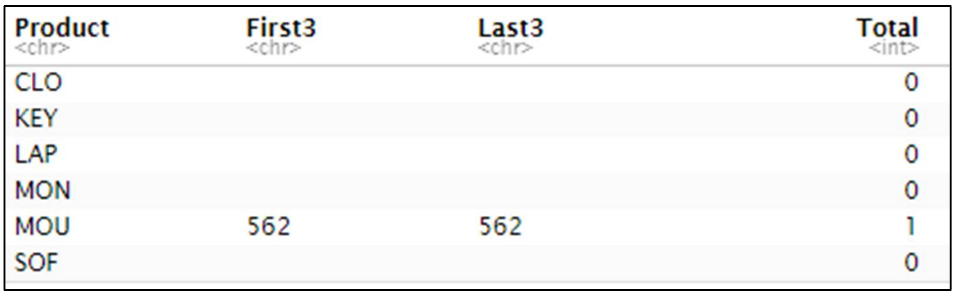
3.5. Samples Showing Processes Out of Control

Three rules were used to identify processes out of control:

Rule A: Flags single S-chart samples above the +3 sigma limit, indicating high variability.

Rule B: Finds the longest run of S-chart samples within ±1 sigma, showing stable process control.

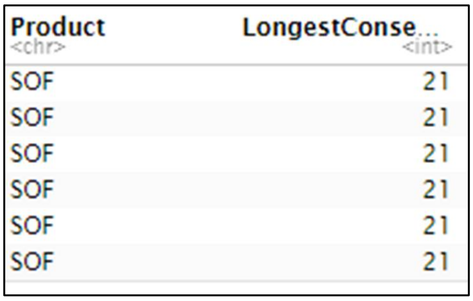
Rule c: Finds four consecutive X-bar samples beyond the upper second control limit, signalling shifts in the process mean.



Product <chr>	First3 <chr>	Last3 <chr>	Total <int>
CLO			0
KEY			0
LAP			0
MON			0
MOU	562	562	1
SOF			0

Table 5: Out-of-Control Samples for S-Chart (Rule A)

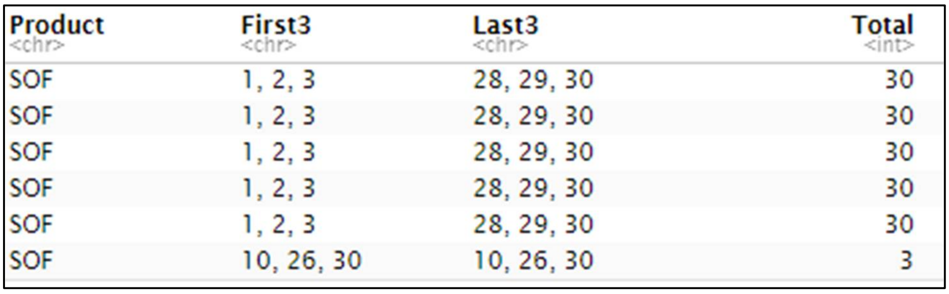
After looking at table 5, it is clear that most of the product types (CLO, KEY, LAP, MON, SOF) have no samples exceeding the +3σ limit on the S Chart, which indicates that the process variability for these products is stable and under control. MOU has a single sample that is out of control (at sample 562), which means that there is a rare spike in variability that may require investigation. Overall, the S-chart shows that process variation is well-managed for most products.



Product	LongestConse...
<chr>	<int>
SOF	21
SOF	21
SOF	21
SOF	21
SOF	21
SOF	21

Table 6: Longest Consecutive Run within ±1 Sigma (Rule B)

Table 6 identifies the longest consecutive sequence of samples where the S values remain within ±1 sigma of the centreline, which represents periods of good process control. CLO has the longest stable period with 14 consecutive samples, followed by MON (12 samples). LAP, KEY, MOU, and SOF have shorter runs, indicating that these processes experience more frequent minor fluctuations in variability. The longer the consecutive run, the more consistent and controlled the process is.



Product	First3	Last3	Total
<chr>	<chr>	<chr>	<int>
SOF	1, 2, 3	28, 29, 30	30
SOF	1, 2, 3	28, 29, 30	30
SOF	1, 2, 3	28, 29, 30	30
SOF	1, 2, 3	28, 29, 30	30
SOF	1, 2, 3	28, 29, 30	30
SOF	10, 26, 30	10, 26, 30	3

Table 7: Out-of-Control Samples for X-bar Chart (Rule C)

Table 7 shows the samples that are outside the upper or lower ±2σ control limits on the X-bar chart. CLO has no out-of-control points, reflecting a stable mean delivery time. KEY has 2 samples exceeding the control limits, suggesting moderate deviations in average delivery times. LAP and MON each have a single out-of-control sample, while MOU and SOF show three points outside the limits, indicating occasional process instability. The presence of out-of-control points should prompt further analysis to identify causes of variability and maintain delivery performance.

Part 4: Risk, Data Correction and Optimising for Maximum Profit

4.1. Probability of Making a Type I Error

A Type I error means rejecting the null hypothesis when it's actually true. (Scribbr, 2021) It can negatively affect profit when products are unnecessarily rejected.

For Rule A (when only one sample falls outside the ±3σ limits on the s-chart), the probability of a false alarm is given by $P(Z > 3) = 1 - \Phi(3) = 0.00135$, or approximately 0.27 %.

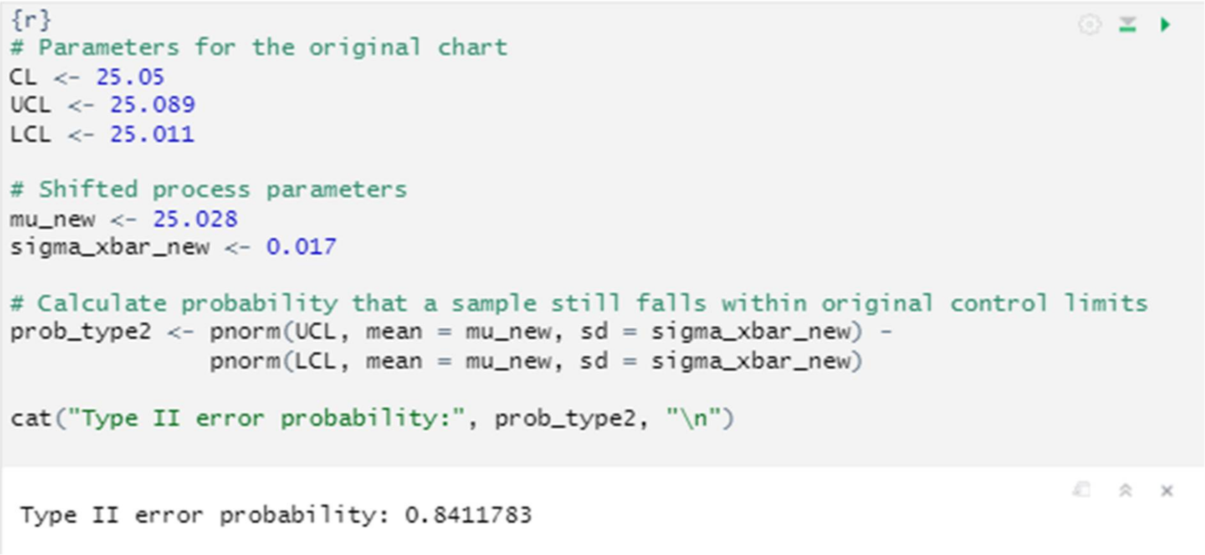
For Rule B, the probability that seven consecutive points fall above the centreline is $0.5^7 = 0.0078$, roughly 0.78 %.

For Rule C (when four consecutive X-bar points are outside of the +2σ warning limit), the probability of any one subgroup goes beyond +2σ is $P(Z > 2) = 0.0228$. The pobability that this happened 4 times in a row is $0.0228^4 = 2.7 \times 10^{-7}$.

These calculations indicate that Rule A would produce a false alarm about once every 370 samples, Rule B about 8 times in 1,000 weeks of sampling, and Rule C almost never. In summary, Type I errors are rare but still possible, especially when a large number of subgroups are being monitored over long periods of time.

4.2. Probability of Making a Type || Error

A Type II error means failing to reject the null hypothesis when it's actually false. (Scribbr, 2021) This can be very problematic, since the process might be seen as if it is in control, but it is not. This can lead to unsatisfied customers.



```
{r}
# Parameters for the original chart
CL <- 25.05
UCL <- 25.089
LCL <- 25.011

# Shifted process parameters
mu_new <- 25.028
sigma_xbar_new <- 0.017

# Calculate probability that a sample still falls within original control limits
prob_type2 <- pnorm(UCL, mean = mu_new, sd = sigma_xbar_new) -
              pnorm(LCL, mean = mu_new, sd = sigma_xbar_new)

cat("Type II error probability:", prob_type2, "\n")
```

```
Type II error probability: 0.8411783
```

Figure 44: Probability of Type || Error

As can be seen in figure 44, the probability of making a Type || Error is 0.8411783. This means that there is an 84.12% likelihood of making a Type || Error. In other words, even though the process mean has shifted to 25.028 litres, there is an 84.12% probability that the sample mean will still lie within the control limits (LCL = 25.011 *l*, UCL = 25.089 *l*), leading to the mistaken conclusion that the process is under control. This high probability indicates that the control limits are not sensitive enough to detect the change in the process means from 25.05 *l* to 25.028 *l*, using the new standard deviation of 0.017 *l*.

4.3. Correcting Head-Office Data

The head-office data was corrected by updating product codes, repeating the correct pricing and markup for rows 11–60 per product type, and updating the category column to match product IDs. The updated files were saved as products_Headoffice2025.csv and products_data2025.csv.

Part 5: Optimising Profit

Shop 1

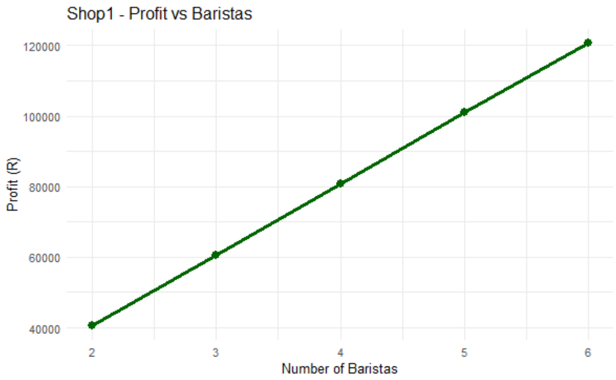


Figure 45: Shop 1 - Profit vs Baristas

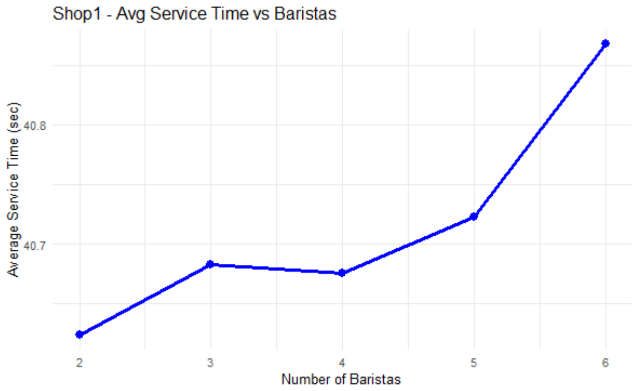


Figure 46: Shop 1 - Avg Service Time vs Baristas

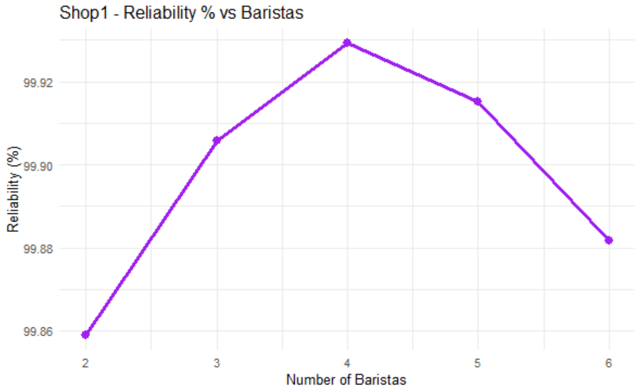


Figure 47: Shop 1 - Reliability % vs Baristas

Ideally, for any shop, you want the profit to be the highest, the reliability % to be the highest, and the average service time (how quickly customers are being served on average at a given number of baristas) to be the lowest. Looking at figure 45, it is clear that profit increases almost linearly with the number of baristas, so the maximum of 6 baristas would give the most profit. However, figure 46 shows that the average service time drastically increases after 5 baristas, meaning it would no longer be optimal to use 6 baristas because we want customers to be served quickly.

Finally, figure 47 shows that reliability % drops after 4 baristas. Considering all three factors, the optimal number of baristas is therefore 4.

Shop 2

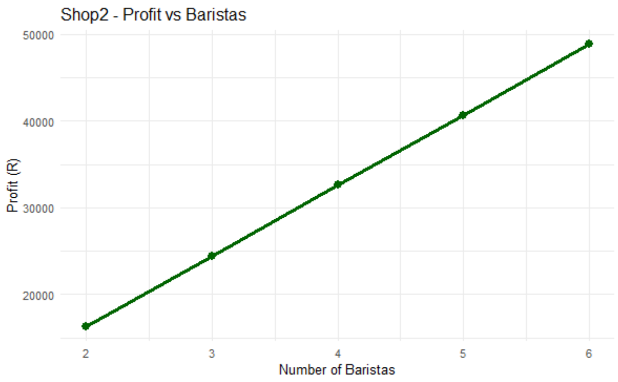


Figure 48: Shop 2 - Profit vs Baristas

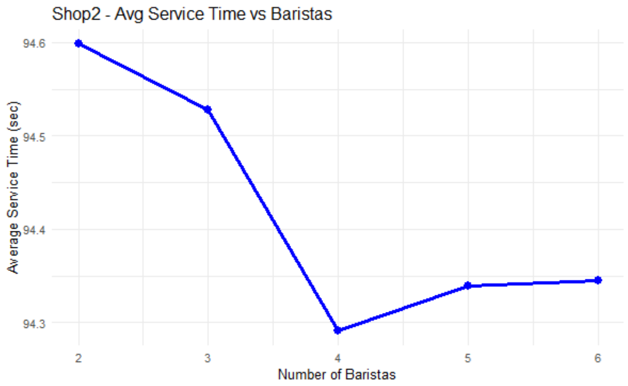


Figure 49: Shop 2 - Avg Service Time vs Baristas

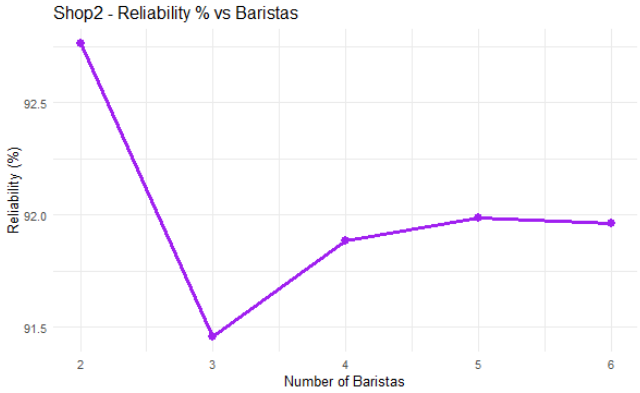


Figure 50: Shop 2 - Reliability % vs Baristas

After looking at figure 48, it is clear that for shop 2, profit also increases almost linearly with the number of baristas, so the maximum of 6 baristas would be optimal from a profit perspective. However, figure 49 shows that the average service time is lowest at 4 and 5 baristas, which is more desirable since we want customers to be served quickly. Figure 50 shows that the highest reliability % occurs with 2 baristas, but that is not optimal for profit or service time. Considering all three factors, 5 baristas provide a good balance, with the second highest reliability % and the second lowest average service time, making it the optimal number of baristas for shop 2.

Part 6: DOE, MANOVA/ANOVA

6.1. Introduction

In this part of the project, the way that the delivery performance differs between 2026 and 2027 is analysed, and whether there are noticeable differences across product types. A Two-Way ANOVA was used to explore this, with orderYear and ProductType as factors and deliveryHours as the response variable. This analysis helps the company to determine if delivery times are consistent across years and product types.

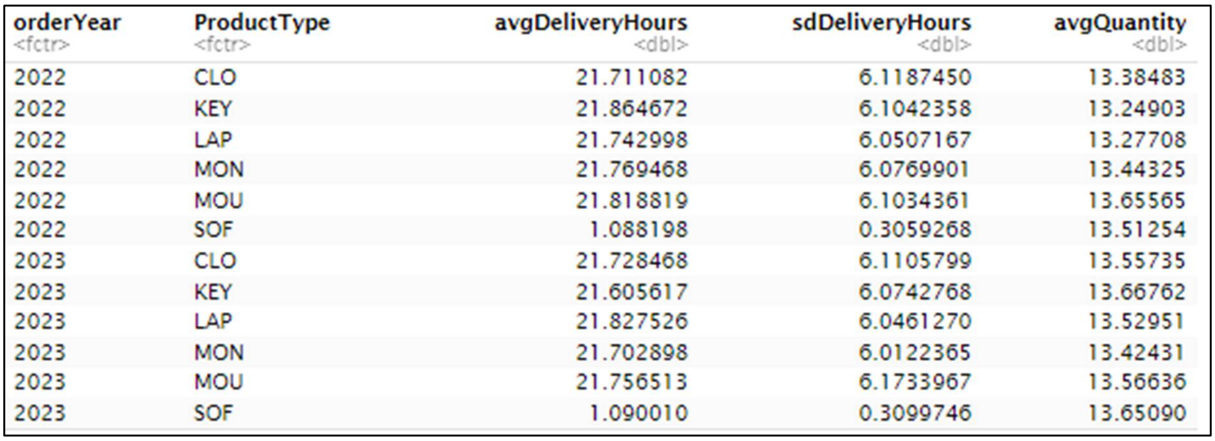
Hypotheses:

1. **Year effect**
   - **Null (H0):** Mean delivery hours are the same for 2026 and 2027.
   - **Alternative (H1):** Mean delivery hours differ between 2026 and 2027.
2. **Product Type effect**
   - **Null (H0):** Mean delivery hours are the same for all product types.
   - **Alternative (H1):** Mean delivery hours differ between product types.
3. **Interaction effect (Year × Product Type)**
   - **Null (H0):** The effect of year on delivery hours is the same across product types.
   - **Alternative (H1):** The effect of year on delivery hours differs by product type.

6.2. Summary Statistics

There is a mistake in the data – 2022 should be 2026 and 2023 should be 2027.

The following table shows the average delivery hours, standard deviation, and quantity for 2026 and 2027 across all product types:



orderYear	ProductType	avgDeliveryHours	sdDeliveryHours	avgQuantity
<fctr>	<fctr>	<dbl>	<dbl>	<dbl>
2022	CLO	21.711082	6.1187450	13.38483
2022	KEY	21.864672	6.1042358	13.24903
2022	LAP	21.742998	6.0507167	13.27708
2022	MON	21.769468	6.0769901	13.44325
2022	MOU	21.818819	6.1034361	13.65565
2022	SOF	1.088198	0.3059268	13.51254
2023	CLO	21.728468	6.1105799	13.55735
2023	KEY	21.605617	6.0742768	13.66762
2023	LAP	21.827526	6.0461270	13.52951
2023	MON	21.702898	6.0122365	13.42431
2023	MOU	21.756513	6.1733967	13.56636
2023	SOF	1.090010	0.3099746	13.65090

Table 8: Summary Statistics

The mean delivery hours between 2026 and 2027 are very similar, indicating a consistent overall delivery performance.

6.3. Two-Way ANOVA Results

The Two-Way ANOVA results for deliveryHours ~ orderYear * ProductType are as follows:

Term	Df	Sum Sq	Mean Sq	F value	Pr(>F)
orderYear	1	139	139	4.707	0.0300 *
ProductType	5	7,022,849	1,404,570	47,696.755	< 2e-16 ***
orderYear:ProductType	5	273	55	1.852	0.0991 .
Residuals	99,988	2,944,438	29		

Table 9: Two-Way ANOVA Results

A Two-Way ANOVA was used to test how year (2026 vs. 2027) and product type affect delivery hours. The results showed a small but statistically significant difference between the two years (F(1, 99 988) = 4.707, p = 0.0300), with average delivery times of 17.50 hours in 2026 and 17.43 hours in 2027.

The product type had a much stronger effect (F(5, 99 988) = 47 696.76, p < 0.001), meaning that delivery times vary noticeably between different types of products.

The interaction between year and product type was not significant (F(5, 99 988) = 1.852, p = 0.0991), indicating that the difference in delivery hours between 2026 and 2027 was similar across all product types.

Checks of the model assumptions showed slight deviations from normality (Shapiro-Wilk p < 0.001) and unequal variances between groups (Bartlett's test p < 0.001). However, this is common with very large datasets and is unlikely to affect the overall conclusions.

Product type is thus the main factor influencing delivery time, while year has only a minor effect, and there is no evidence that the year affects specific product types differently.

6.4. Graphical Analysis

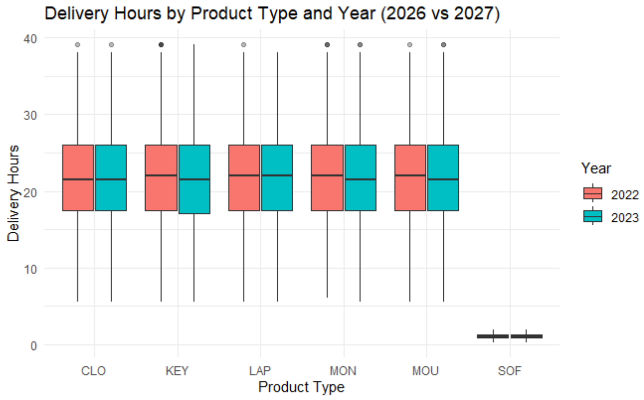


Figure 51: Delivery Hours by Product Type and Year

Figure 51 shows the distribution of delivery hours for each product type in 2026 and 2027. Some outliers are visible, but the median delivery hours remain almost consistent. The product types KEY, MOU and MON had a slight decline in the median of delivery hours from 2026 to 2027, which means that the delivery performance for these products slightly increased from 2026 to 2027. For the rest of the product types the delivery performance between years was consistent.

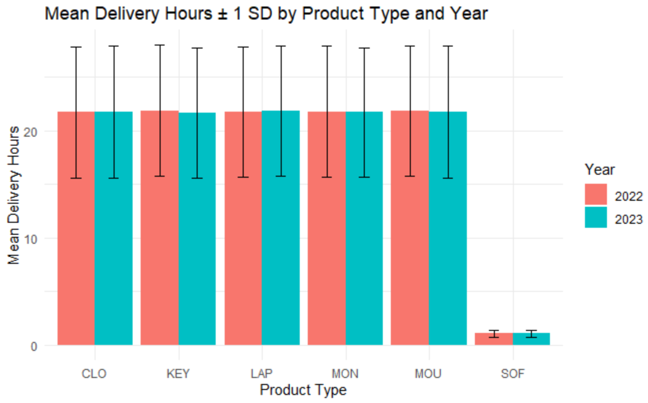


Figure 52: Mean Delivery Hours +- SD by Product Type and Year

Figure 52 shows the average delivery hours for each product type along with their corresponding standard deviation. This is useful for management to monitor delivery performance across different product categories. The product type LAP has slightly higher delivery hours in 2027, suggesting possible delays in handling or shipping for this product type. On the other hand, KEY and MOU's mean delivery hours improved, indicating improvements in efficiency or faster processing times. The rest of the product types displayed almost no change between the two years, reflecting consistent delivery performance over time.

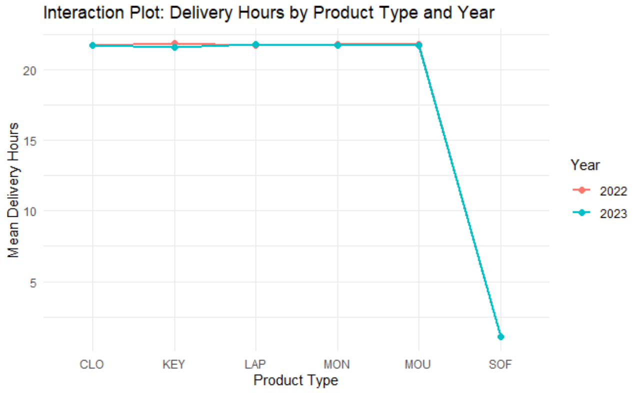


Figure 53: Interaction Plot

Figure 53 shows the relationship between year and product type on delivery hours. Nearly parallel lines indicate that delivery trends for most product types are consistent across 2026 and 2027.

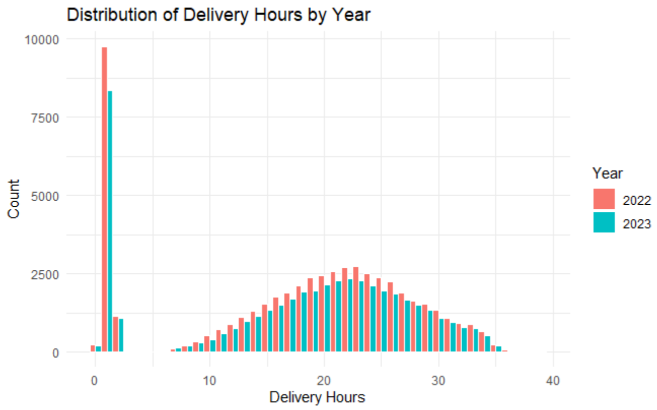


Figure 54: Distribution of Delivery Hours by Year

Figure 54 confirms that the overall delivery time distribution is very similar between 2026 and 2027. There is only an overall decline in the total delivery hours from 2026 to 2027, which is a very good thing, because it means that the company's overall delivery performance increased.

6.5. Extra Month Analysis

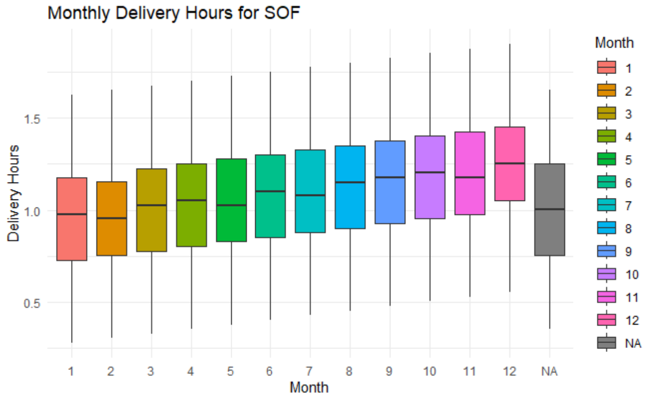


Figure 55: Monthly Delivery Hours for SOF

For the selected product type SOF, a Month ANOVA was conducted to determine if delivery hours differ across months. The results show that average delivery hours generally increase over time, although the monthly medians fluctuate rather than showing a consistent upward trend. Thus, the boxplots confirm seasonal variations in delivery performance for SOF products.

6.6. Discussion

Overall, the delivery performance between 2026 and 2027 stayed quite consistent. There were only small differences between product types, but the company can confidently plan delivery schedules using these averages. The bar and interaction plots provide valuable insights for management to track product-specific delivery performance and identify any deviations over time. While delivery hours were generally consistent across years, some product types required slightly longer delivery times. The parallel trends across product types suggest stable and predictable performance.

Part 7: Reliability of Service

7.1. Expected Days of Reliable Service

The following graph reflects the number of people on duty over 397 days:

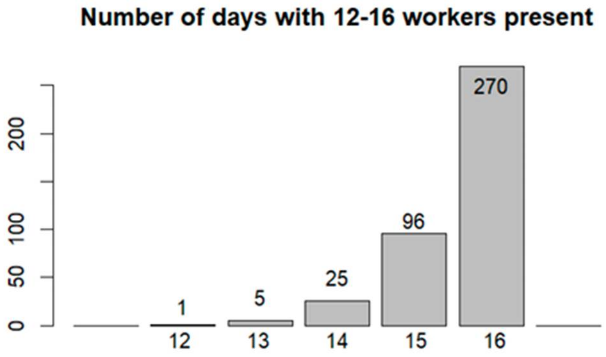


Figure 56: Number of Days with 12-16 Workers Present

Service problems occur when fewer than 15 staff members are on duty. Using the distribution in Figure 56, the expected number of reliable service days per year were calculated as follows:

$$\text{Reliable days} = 96 + 270 = 366 \text{ days over 397 recorded days}$$

Now, scaling to a year:

$$\text{Expected reliable days} \approx \frac{366}{397} \times 365 \approx 336.2 \text{ days/year}$$

Thus, with the current staffing, the agency can expect reliable service for roughly 336 days per year.

7.2 Optimising Profit

To optimise profit, daily losses from insufficient staffing were compared with monthly costs of additional personnel:

Daily loss if less than 15 workers were used: R20,000

Cost per additional worker: R25,000 per month

Currently 14 staff are present on 25 days, causing a potential revenue loss of:

$$25 \text{ days} \times 20{,}000 \text{ R/day} = 500{,}000 \text{ R/year}$$

Hiring 2 additional workers increases the minimum staffing to 16, eliminating almost all problem days (only 1 day with 12 workers would remain in the model). The total annual cost for these additional workers is:

$$2 \text{ workers} \times 25{,}000 \text{ R/month} \times 12 \text{ months} = 600{,}000 \text{ R/year}$$

Thus, 16 workers will give the most optimal combination of reliable service and personnel costs.

Conclusion

This project illustrated how data analysis can inform and enhance business processes. The SPC and capability studies showed how variability and process performance can be monitored and controlled, while the exploration of Type I and Type II errors highlighted the importance of interpreting results carefully. The optimisation analyses provided insights into balancing staff levels, service reliability, and costs to improve profitability. Finally, the ANOVA and monthly analyses demonstrated how performance can differ across products and time periods. Overall, the project highlighted the value of using data-driven methods to understand, manage, and improve both operational processes and business outcomes.

References

Engineering Council of South Africa (ECSA), 2025. *Project ECSA 2025: Stellenbosch University Quality Assurance Module* [pdf]. Stellenbosch University. Available at: ProjectECSA2025Final.pdf [Accessed 24 October 2025].

Scribbr, 2021. *Type I and Type II errors* [online]. Available at: Type I & Type II Errors | Differences, Examples, Visualizations [Accessed 24 October 2025].

STHDA, 2022. *MANOVA Test in R: Multivariate Analysis of Variance – Easy Guides* [online]. Available at: MANOVA Test in R: Multivariate Analysis of Variance - Easy Guides - Wiki - STHDA [Accessed 24 October 2025].