ECSA GA4 FINAL REPORT

Data Analysis Report

Vonkeman, D, Miss [26951142]

24 October 2025

Table of Contents

Introduction

This report is based on the ECSA GA4 assessment for Industrial Engineering and aims to investigate various aspects of data analysis, process improvement, and decision-making, using multiple datasets and real-world business and engineering examples. Statistical and R programming tools are applied to analyse and solve the given problems.

To begin with, descriptive statistics are used to analyse the dataset, after which Statistical Process Control (SPC) techniques are applied to assess performance levels and identify variation patterns, specific to the business context. Capability indices such as Cp, Cpk and Cpu are then used to evaluate the process performance and determine whether customer requirements are being met.

The following tasks focus on error analysis – specifically Type I and Type II errors – along with data correction, data comparison of two datasets, and profit optimisation for service-based businesses.

ANOVA tests are then completed, to address a research question derived from the dataset. In the case of this report, it is evaluated whether delivery times differ significantly across months for a given product type from the dataset provided.

Finally, the reliability of service for a car rental agency is investigated, followed by the development of an optimised binomial model to maximise the company's profit through optimal personnel allocation.

The integration of the various components of this report provides a comprehensive analysis of statistical reasoning, R programming, and system performance evaluation, aimed at enhancing operational efficiency, process optimisation, and overall reliability.

Part 1.2: Descriptive Statistics

Introduction

The company provided four datasets to complete a full analysis of its current position by examining Sales, Customers, Products, and Products_Headoffice. The provided data was checked for missing values or duplicates, used to create summary statistics, integrated where needed, and finally used to generate visualizations. This process allowed for thorough insight, enabling a detailed data analysis. Outputs of various kinds are used, analyzed, and discussed. This process provides a better understanding of the company's standing regarding the products, cities, and categories that add the most value to performance, whilst also identifying any underlying strengths, weaknesses, and areas where change is required.

Outputs

Data Quality Report

Data Quality Table

Data Quality Summary			
Dataset	**Rows**	**Columns**	**Total_NAs**
Sales	100000	9	0
Customers	5000	5	0
Products	60	5	0
Products_Headoffice	360	5	0

Table 1.2.1: Data Quality Table

```
$Sales
   CustomerID     ProductID      Quantity     orderTime      orderDay
            0             0             0             0             0
   orderMonth     orderYear  pickingHours deliveryHours
            0             0             0             0

$Customers
CustomerID     Gender        Age     Income       City
         0          0          0          0          0

$Products_Headoffice
   ProductID     Category  Description SellingPrice       Markup
           0            0            0            0            0

$Products
   ProductID     Category  Description SellingPrice       Markup
           0            0            0            0            0
```

Table 1.2.2: Number of missing values per dataset

These outputs demonstrate that the company maintains a clean and complete dataset, facilitating thorough data analysis and subsequent decision-making. Further, no rows need to be removed due to missing data, so any results represent the company's position accurately.

However, it must be noted that even though there are no missing values, i.e., the dataset appears to be complete, it does reveal errors in the identification of the product data. This may, in itself, cause issues for accurate analysis, as the misconception that no missing values indicates a correct dataset persists and could result in making decisions based on inaccurate data. Caution should be taken.

Data Quality based on Integration of Data Sets

CustomerID	ProductID	Category.x	SellingPrice.x	Category.y	SellingPrice.y
CUST1791	CLO011	Keyboard	1070.54	NA	NA
CUST3172	LAP026	Cloud Subscription	18711.72	NA	NA
CUST1022	KEY046	Monitor	708.18	NA	NA
CUST3721	LAP024	Mouse	18366.92	NA	NA
CUST4605	CLO012	Mouse	963.14	NA	NA
CUST2766	MON035	Keyboard	6396.18	NA	NA
CUST4454	MOU052	Monitor	425.14	NA	NA
CUST582	MON032	Cloud Subscription	6634.13	NA	NA
CUST3343	MON040	Monitor	5346.14	NA	NA
CUST4331	KEY049	Software	752.75	NA	NA

Table 1.2.3: Data Integration

The table above is formatted so that the left side (Category.x and SellingPrice.x) shows information regarding the *Products* dataset. In contrast, the right side (Category.y, SellingPrice.y) shows information regarding the *Products_Headoffice* dataset. From the data integration, each row reveals actual product information from the *Products* dataset; however, the integration with the Products_Headoffice dataset shows only NA values. This means that the specific ProductIDs were not found in that CSV file, indicating missing data. As a result, a data integration discrepancy exists.

Any further analysis using Products_Headoffice as a comparison to Products must consider this exclusion, as these items are not valid across both datasets. This could undermine the importance of specific categories where this discrepancy is evident. This table reveals an essential factor: clearly, not all products sold are also sold at the head office. *Sales* and *Customers,* though, still integrate smoothly with no discrepancies, meaning that customer analysis can still be completed unaffected. However, this mishap does indicate a possible inconsistency in the record-keeping or coding systems, which may not align between the various departments.

CustomerID <chr>	ProductID <chr>	Age <dbl>	Income <dbl>	City <chr>	Quantity <dbl>	orderYear <dbl>	orderMonth <dbl>
CUST1791	CLO011	39	100000	Los Angeles	16	2022	11
CUST3172	LAP026	58	90000	Chicago	17	2023	7
CUST1022	KEY046	20	95000	Seattle	11	2022	5
CUST3721	LAP024	66	60000	Miami	31	2023	7
CUST4605	CLO012	70	25000	Chicago	20	2022	2

Table 1.2.4: Customer and ProductID Integration

The table above serves as a visual representation of a clean integration with Customers, demonstrating correct record-keeping with no discrepancies.

Summary Statistics

Sales: Distribution of Quantity

```
    Quantity
Min.    : 1.0
1st Qu.: 3.0
Median : 6.0
Mean   :13.5
3rd Qu.:23.0
Max.    :50.0
```

This distribution shows most orders are small in size, with the mean being much higher than the median, suggesting right-skewed data. A small number of large orders is increasing the average. The maximum of 50 reveals that bulk orders are possible, though they are rare compared to the usual order of 3-6 units. The IQR suggests wide variability.

Customers: Count of unique CustomerID

```
 CustomerID          Gender              Age
Length:5000        Length:5000        Min.    : 16.00
Class :character   Class :character   1st Qu.: 33.00
Mode  :character   Mode  :character   Median : 51.00
                                      Mean   : 51.55
                                      3rd Qu.: 68.00
                                      Max.    :105.00
```

There is a wide range of customers, ranging from 16 to 105 years old. The mean and median are very similar, suggesting a balanced age distribution.

However, the IQR indicates that the majority of the customers are middle-aged and older. This means most of the opportunities should come from middle-aged people; however, the company should still consider younger people in their marketing strategies.

Products & Products_Headoffice: Range and Average of Selling Price and Markup

Products:

```
 SellingPrice        Markup
Min.    :  350.4   Min.    :10.13
1st Qu.:  512.2   1st Qu.:16.14
Median :  794.2   Median :20.34
Mean   : 4493.6   Mean   :20.46
3rd Qu.: 6416.7   3rd Qu.:25.71
Max.    :19725.2   Max.    :29.84
```

Products_Headoffice:

```
 SellingPrice        Markup
Min.    :  290.5   Min.    :10.06
1st Qu.:  495.9   1st Qu.:15.84
Median :  797.2   Median :20.58
Mean   : 4411.0   Mean   :20.39
3rd Qu.: 5843.3   3rd Qu.:24.84
Max.    :22420.1   Max.    :30.00
```

Both datasets reveal very similar figures, indicating a consistent markup policy between the internal departments and head offices.

For the Selling Price, in both datasets, the mean is much higher than the median, indicating a highly right-skewed distribution. Most products are valued at prices between R500 and R800, with perhaps premium products increasing the average price. Due to the Head Office products having lower quartile values, but a higher maximum price, it could indicate that the Head Office also offers extra high-value products.

Histogram: Sales Quantity Distribution

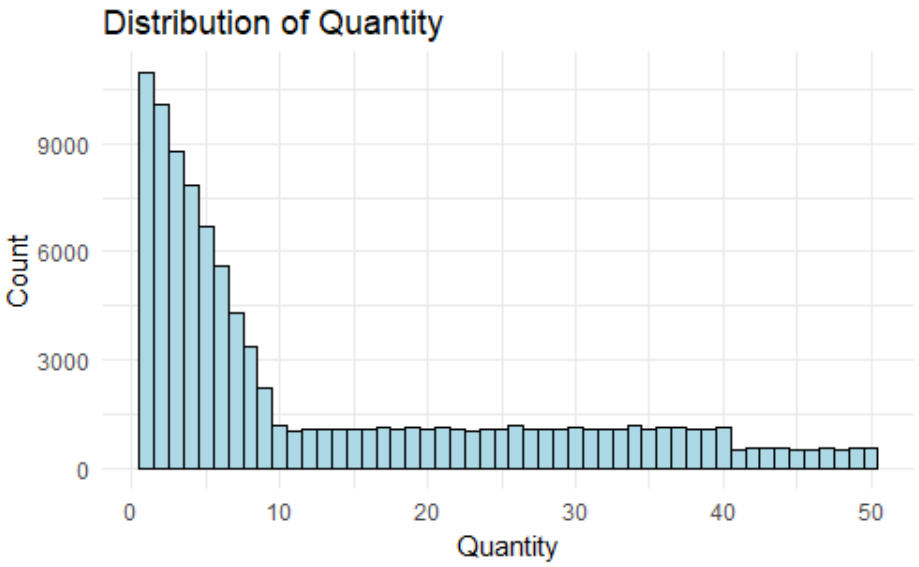

Figure 1.2.1: Histogram to show the Distribution of Quantity

As mentioned earlier, this histogram reveals that most sales are small, ranging from about 1 to 5 units, with fewer sales occurring as the order quantity increases. However, there is still a constant number of medium-to-large-sized orders, ranging from about 15 to 40 units. At the same time, the majority of the demand comes from purchases involving small quantities, likely from individuals or small businesses.

Boxplot: Product Pricing and Markup Spread

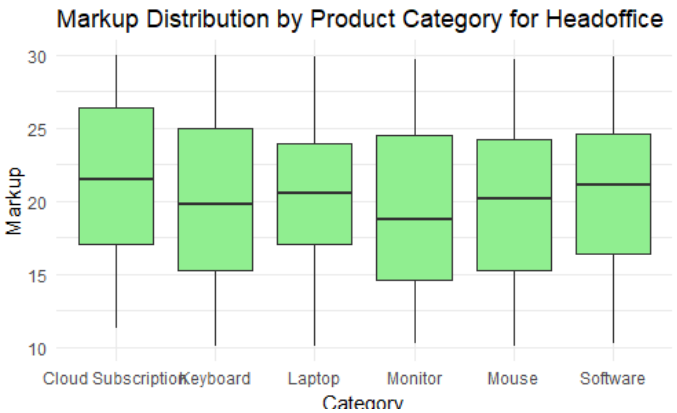

Figure 1.2.2: Boxplot to show Markup Distribution by Product Category: Head office

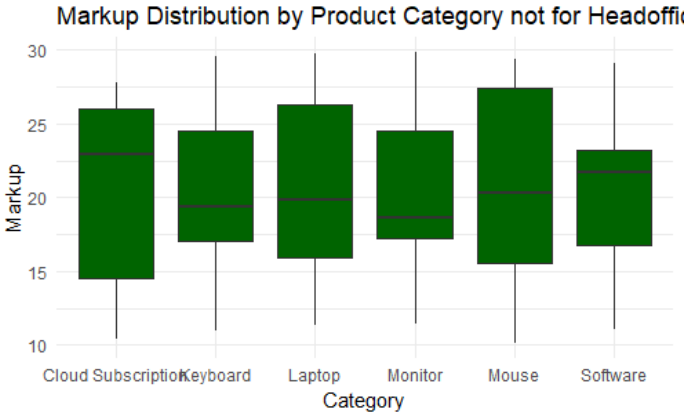

Figure 1.2.3: Boxplot to show Markup Distribution by Product Category: Products

For Products (not Head Office), Cloud Subscriptions, and Mice, the broadest spread in markup values indicates that these products are sold with much higher or lower margins. The median markups for most categories sit around 20% with Cloud Subscriptions being slightly higher, suggesting possible higher demand.

For Products_Headoffice, the spread is more consistent and less wide, indicating better standardized pricing. The median levels are generally higher than those for non-Head office products (suggesting different strategies between the departments), but align more consistently within the various categories, showing more balance, with no one category standing out much more

than the rest. However, Cloud Subscription still shows a higher markup, exhibiting patterns similar to those if it were not in the head office.

Scatterplot: Quantity vs Selling Price

Figure 1.2.4: Scatterplot to show Quantity vs Selling Price

This scatter plot shows that the sales quantity is stacked from 1 to 50 at set price ranges, specifically those seen in the graph. This could mean that the selling price of the company's products has only specific, distinct values, rather than being continuous. The graph shows a concentrated set of data at the lower end of the price range, which suggests that the bulk of the company's sales is from cheaper, more affordable products. There is still a strong, but slightly less concentrated relationship with higher selling prices and increased quantity. Nonetheless, it indicates that premium products sell in higher amounts, showing a variation in demand across different price ranges. This relational information is essential for understanding the company's most effective pricing strategy and the breadth of its customer base, which appears to be quite extensive.

Bar Chart: Top Categories by Sales

As seen from this bar chart, there are no drastic differences in the revenue generated by the various categories. However, Laptops and Monitors stand out as the categories providing the most significant income, while Cloud Subscription brings in lower total revenue. This helps management with stocking, marketing, and allocating resources correctly.

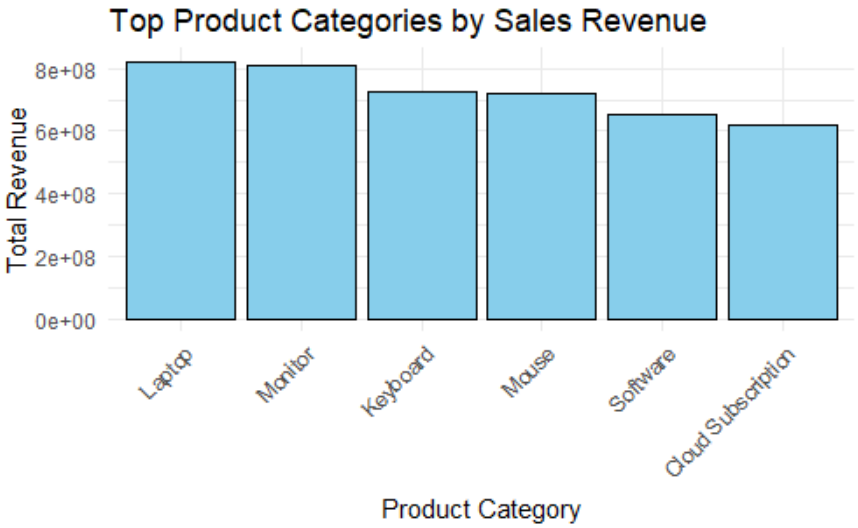

Figure 1.2.5: Bar graph to show the Top Product Categories by Sales Revenue

Bar Chart: Top Cities by Sales

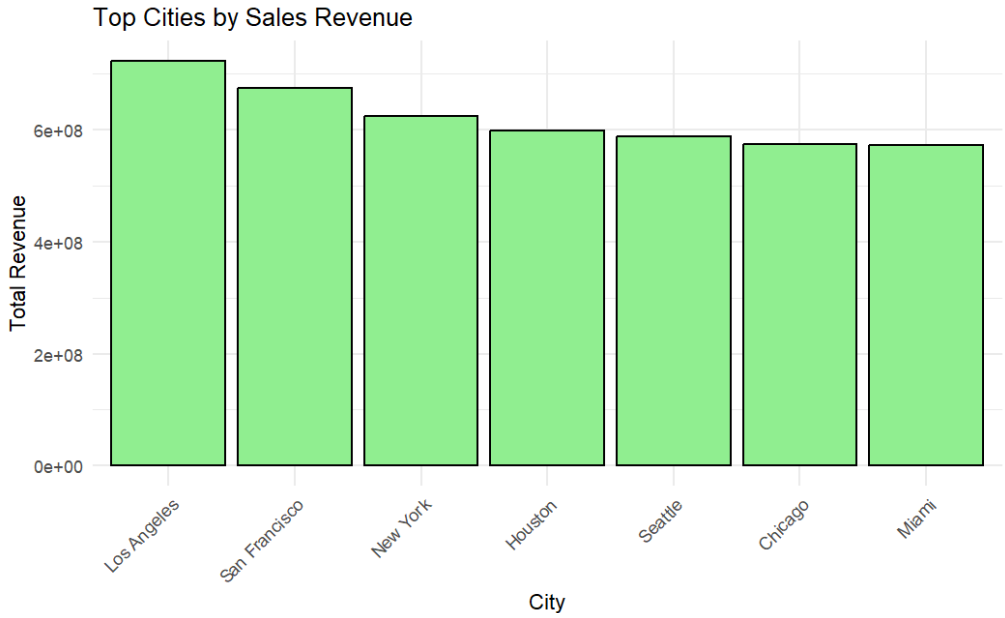

Figure 1.2.6: Bar Graph to show Top Cities by Sales

The bar chart above shows Los Angeles to be the company's most important city in terms of location. This is followed very closely by San Francisco and New York. However, there is no drastic difference between the other cities, showing that all play a relatively equal role in terms of sales. As a result, a wide range of opportunities is created for sales, with a few key areas having a slightly greater impact. This can help marketing and distribution strategies by indicating where emphasis should be placed and how much.

Line Chart: Sales Trends over Time

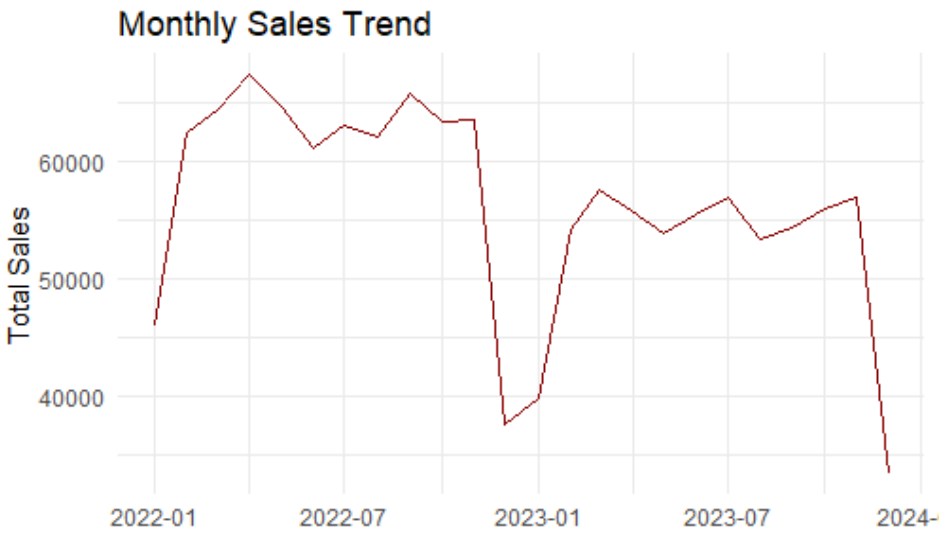

Figure 1.2.7: Line Graph to show the Monthly Sales Trend

This monthly sales trend shows a steep increase in costs during the first half of 2022, followed by steady sales until a deep decline at the beginning of 2023. However, this was quickly recovered with a constant amount of sales, which was generally lower than in 2022. This suggests seasonality or disruptions in operations at the beginning of 2023, indicating that the company should focus more on better demand forecasting and resource planning.

Box Plot to show Selling Price by Customer Income Group

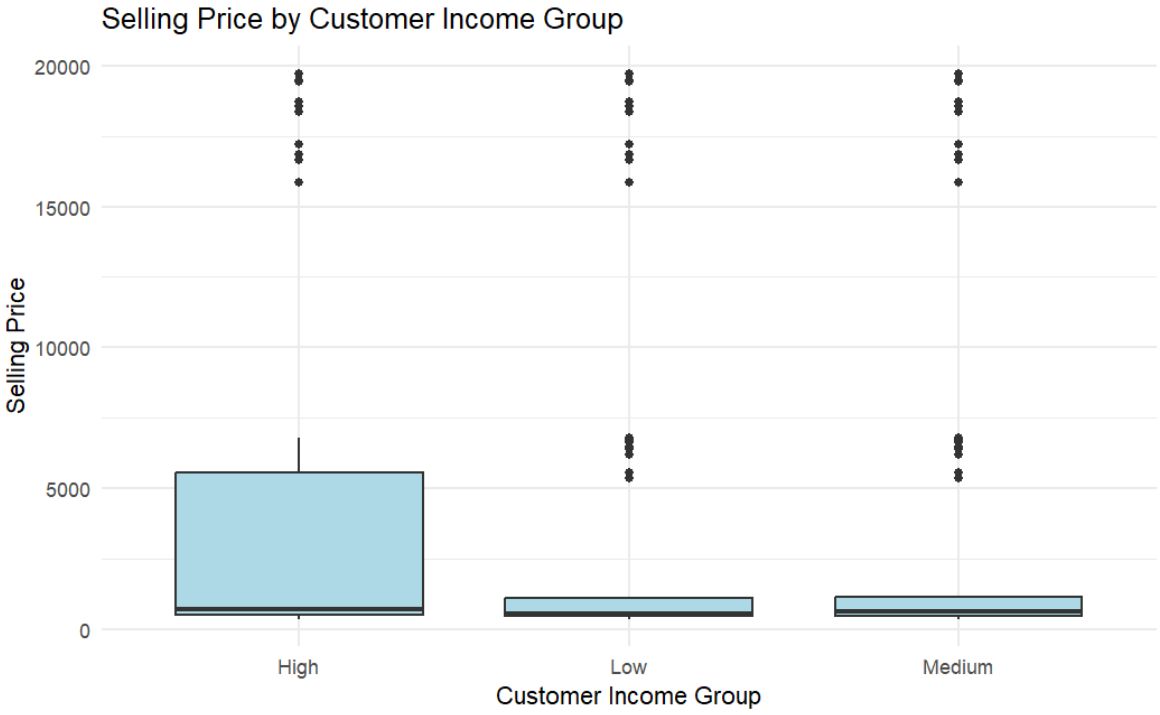

Figure 1.2.8: Box plot to show Selling Price by Customer Income Group

The boxplot above shows that high-income customers make purchases across a wider price range than lower and medium-income customers. The medians across all three income groups are pretty similar, ranging from a lower price of R500 to R600. This indicates that, as can be expected, high-income customers are most likely to buy premium and high-value products.

The plot further shows outliers in all three groups, which indicates that medium and low-income customers also buy premium products on rare occasions. The company can interpret this to mean they should prioritize providing most of the premium products to high-income customers; however, they should still plan for low and medium-income customers to have the opportunity to purchase these higher-priced products.

Part 3: Statistical Process Control

Introduction

This part of the report aims to conduct Statistical Process Control (SPC) on a set of delivery-time data related to the 2026 and 2027 sales of various product types. The data was cleaned and chronologically ordered, after which it was grouped into samples of size 24. This was to represent how real-time monitoring and process control would work. In addition to this, X-charts and S-charts were created based on the samples to analyze process stability by calculating the process capability indices, including Cp, Cpu, Cpl, and Cpk. This would allow for analysis of whether each process can meet customer requirements related to a VOC of 0-32 hours.

Data Preparation

To ensure analysis was done on correct and clean data, missing or unrealistic delivery times were cleared from the dataset and organized chronologically by date. Each product type was revealed to have between 700 and 900 samples, each consisting of 24 instances. The table below from the RMarkdown file summarises the statistics regarding the delivery times of each product type.

Table 3.1: Descriptive Statistics for Delivery Times per Product Prefix

prod_prefix	n_deliveries	mean_hours	sd_hours	min_hours	max_hours
CLO	14763	21.06	5.60	5.54	31.55
KEY	16939	21.07	5.55	5.54	31.55
LAP	9662	21.13	5.54	5.54	31.55
MON	14067	21.08	5.52	5.55	31.55
MOU	19528	21.12	5.62	5.54	31.55
SOF	20749	1.09	0.31	0.28	1.90

Control Charts

3.1 Control Charts for the initial 30 Samples per Product Type

SOF Products:

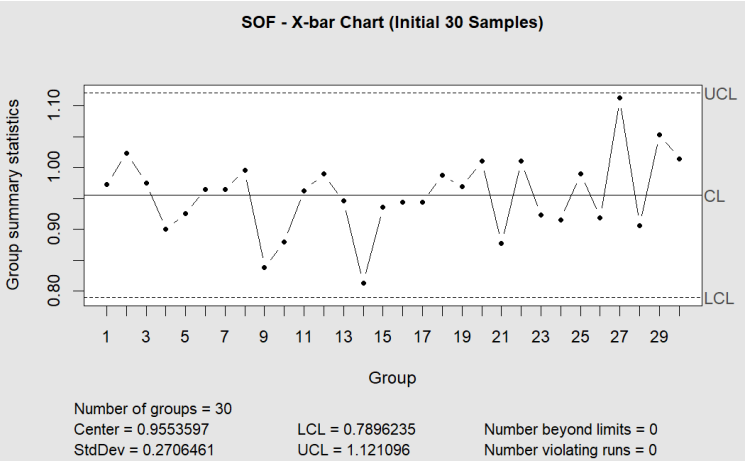

Figure 3.1.1: X-bar Chart for SOF Products (Initial 30 Samples)

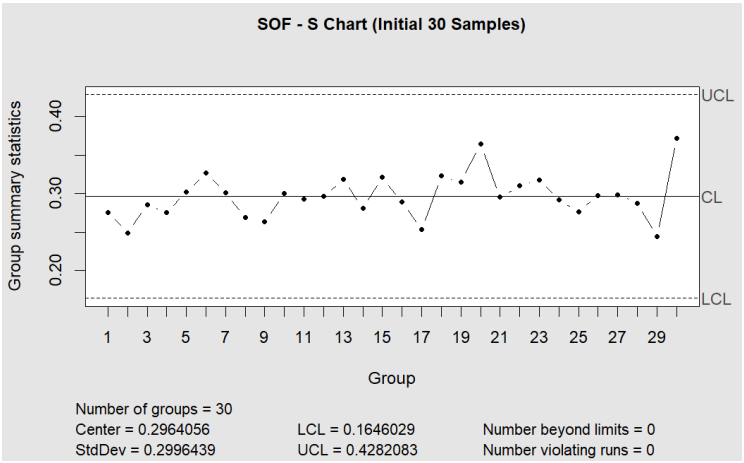

Figure 3.1.2: S Chart for SOF Products (Initial 30 Samples)

CLO Products:

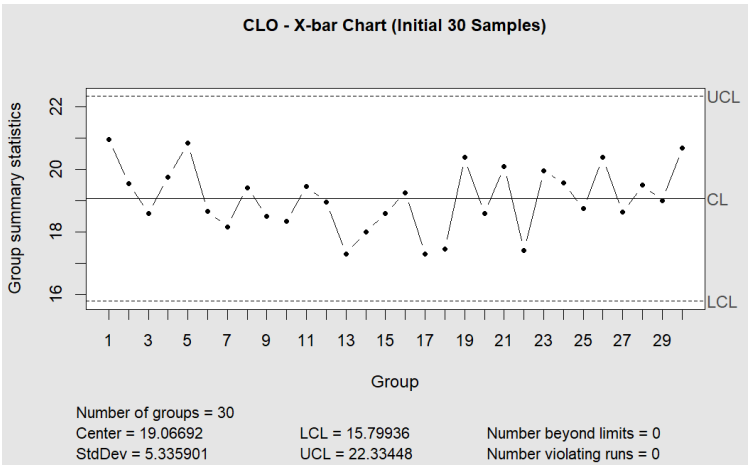

Figure 3.1.3: X-bar Chart for CLO Products (Initial 30 Samples)

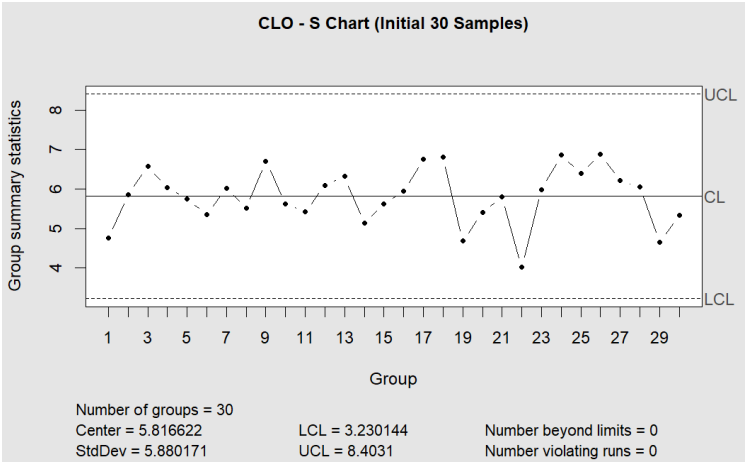

Figure 3.1.4: S Chart for CLO Products (Initial 30 Samples)

KEY Products:

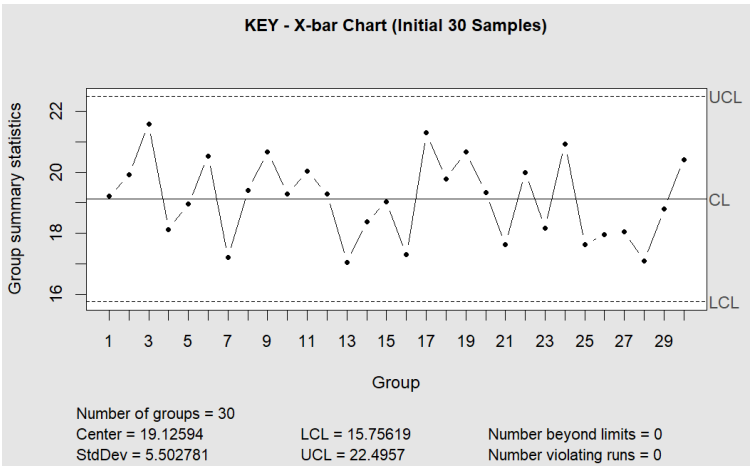

Figure 3.1.5: X-bar Chart for KEY Products (Initial 30 Samples)

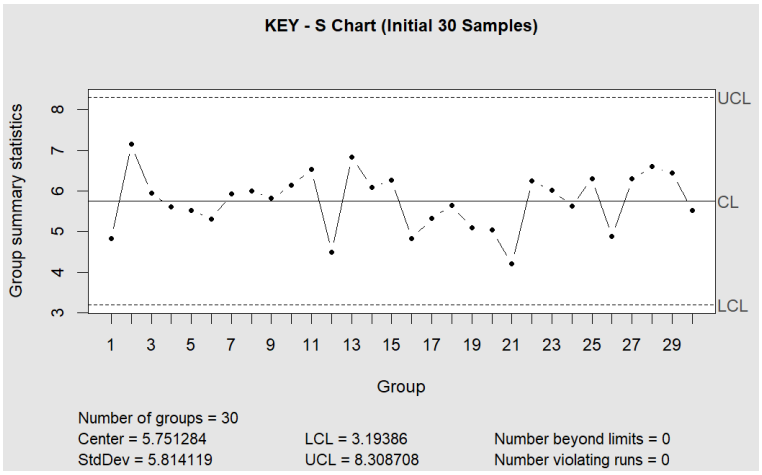

Figure 3.1.6: S Chart for KEY Products (Initial 30 Samples)

MOU Products:

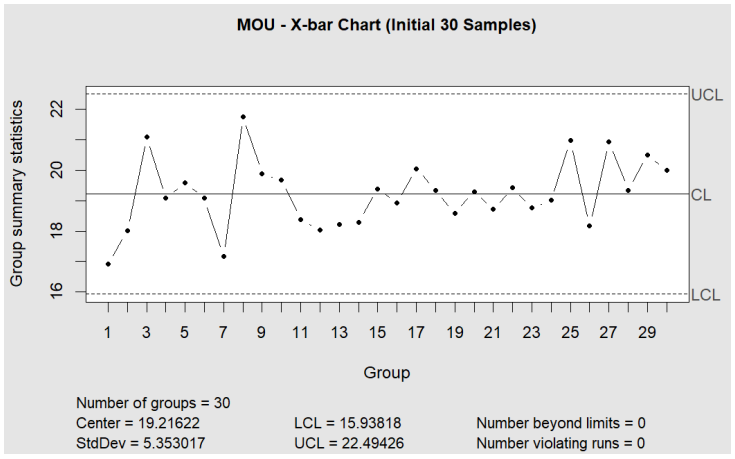

Figure 3.1.7: X-bar Chart for MOU Products (Initial 30 Samples)

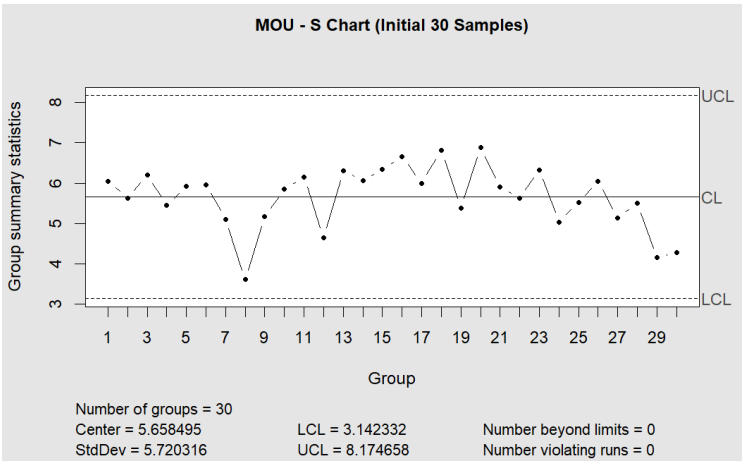

Figure 3.1.8: S Chart for MOU Products (Initial 30 Samples)

MON Products:

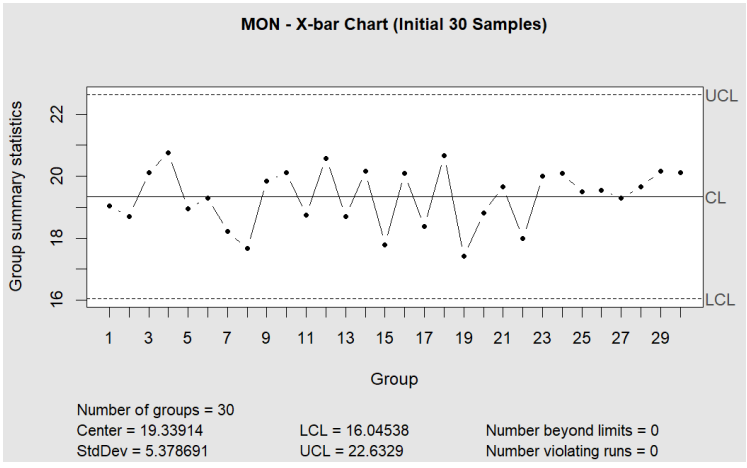

Figure 3.1.9: X-bar Chart for MON Products (Initial 30 Samples)

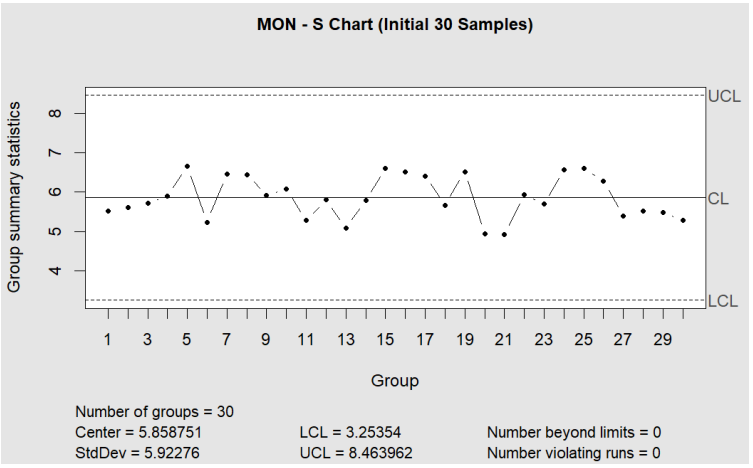

Figure 3.1.10: S Chart for MON Products (Initial 30 Samples)

LAP Products:

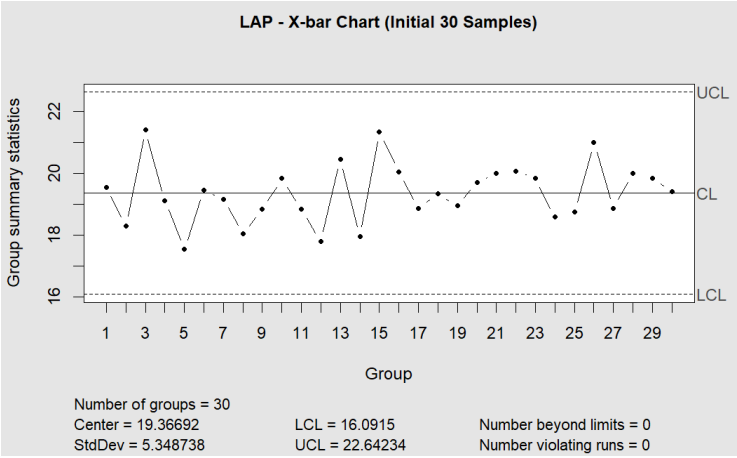

Figure 3.1.11: X-bar Chart for LAP Products (Initial 30 Samples)

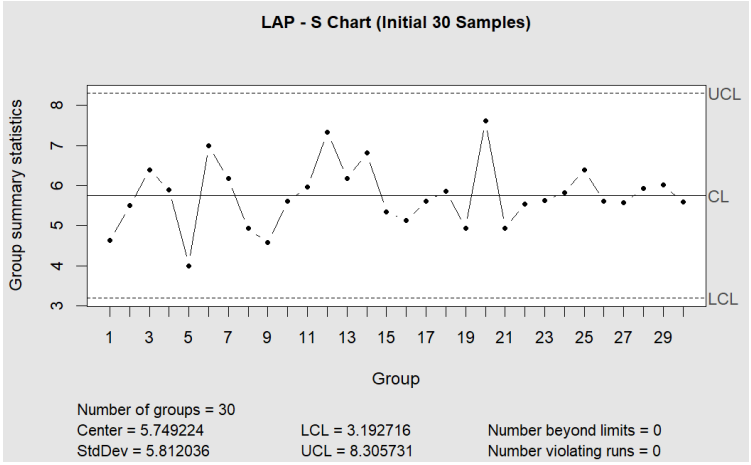

Figure 3.1.12: S Chart for LAP Products (Initial 30 Samples)

The X-bar charts and corresponding S charts for the various product types, using the first 30 samples (720 observations), are displayed above. The control limits at ±1σ, ±2σ, and ±3σ shown in the graph are evaluated using the qcc package in RStudio. All the charts show a fairly similar pattern, with all samples falling neatly within the control limits. This indicates good control.

3.2 Control Charts for all Samples of each Product Type

SOF Products:

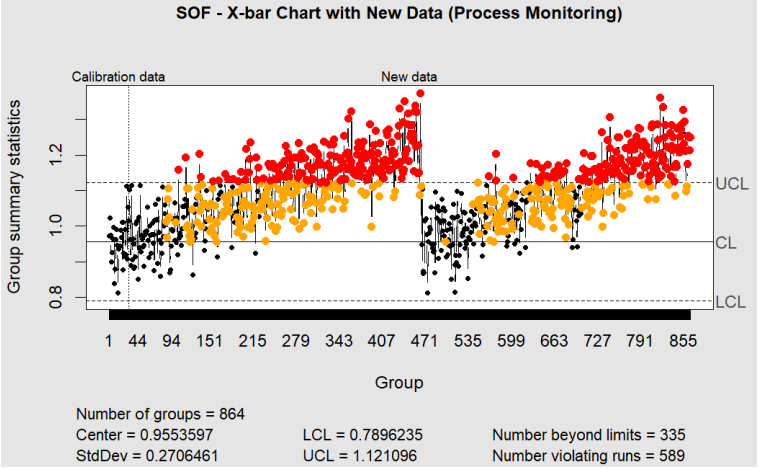

Figure 3.2.1: X-bar Chart for SOF Products (All Samples)

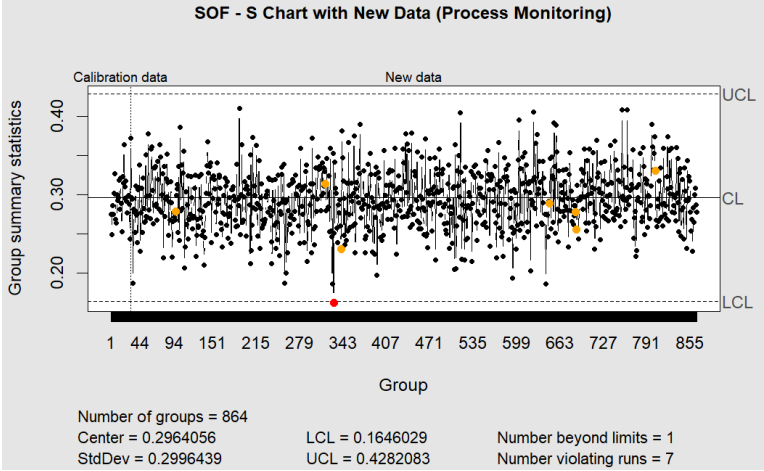

Figure 3.2.2: S Chart for SOF Products (All Samples)

CLO Products:

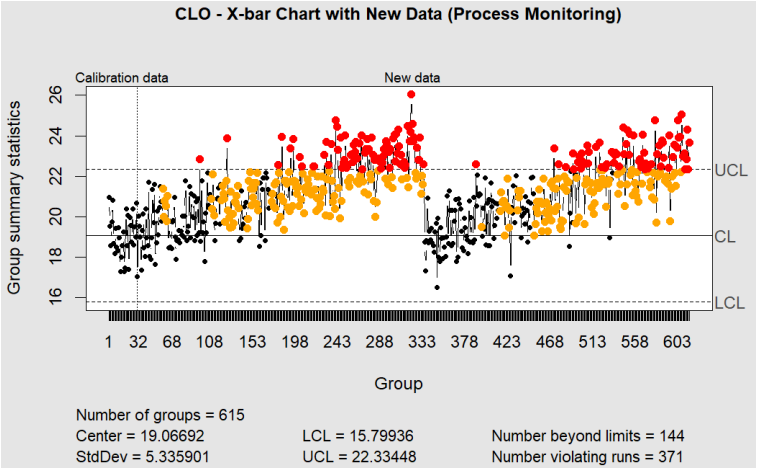

Figure 3.2.3: X-bar Chart for CLO Products (All Samples)

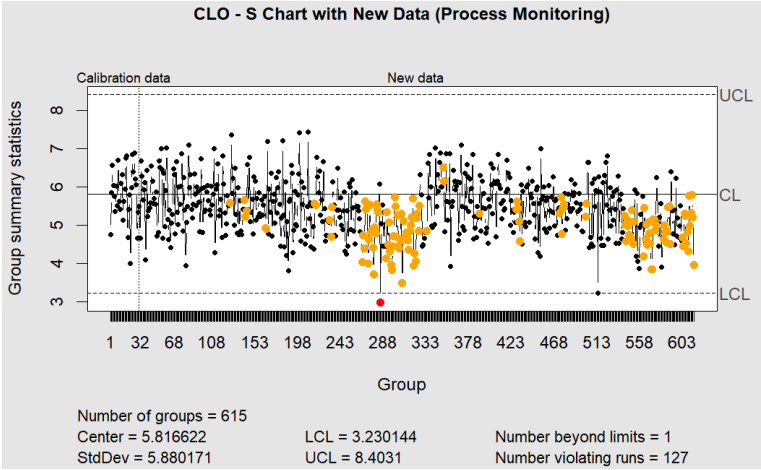

Figure 3.2.4: S Chart for CLO Products (All Samples)

KEY Products:

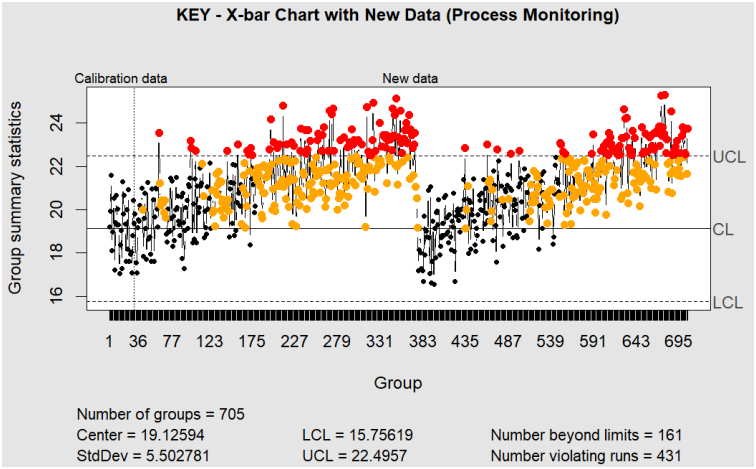

Figure 3.2.5: X-bar Chart for KEY Products (All Samples)

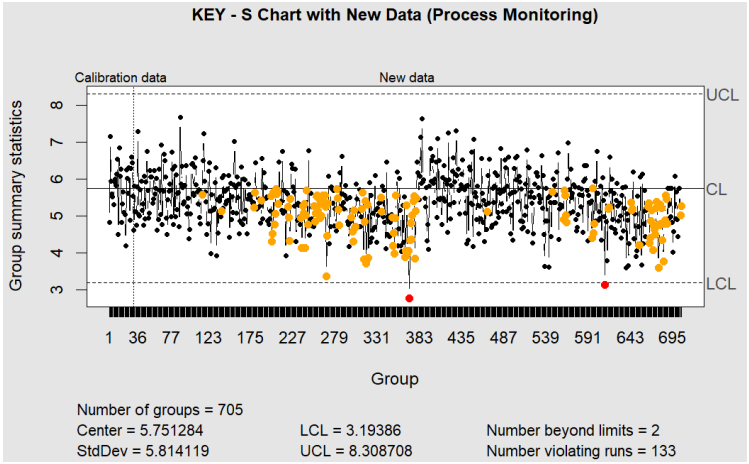

Figure 3.2.6: S Chart for KEY Products (All Samples)

MOU Products:

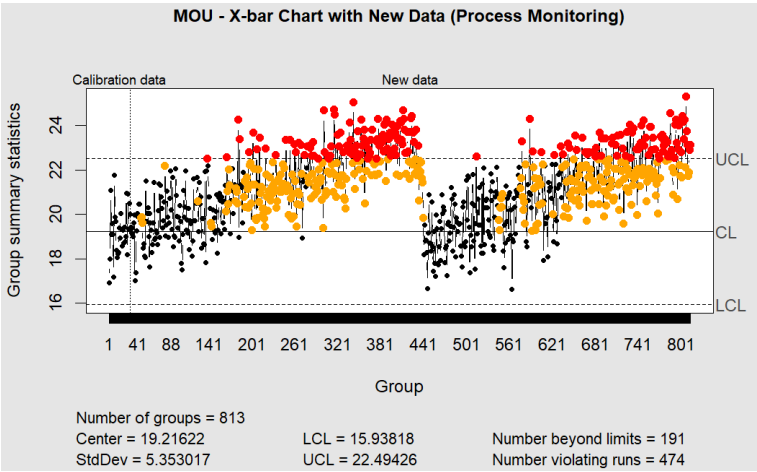

Figure 3.2.7: X-bar Chart for MOU Products (All Samples)

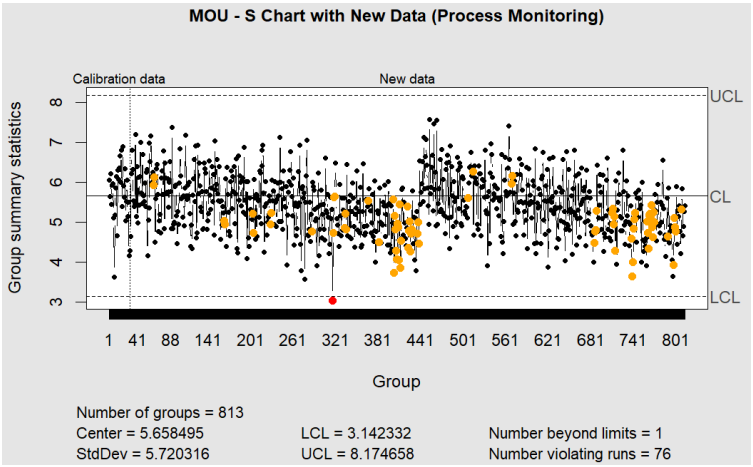

Figure 3.2.8: S Chart for MOU Products (All Samples)

MON Products:

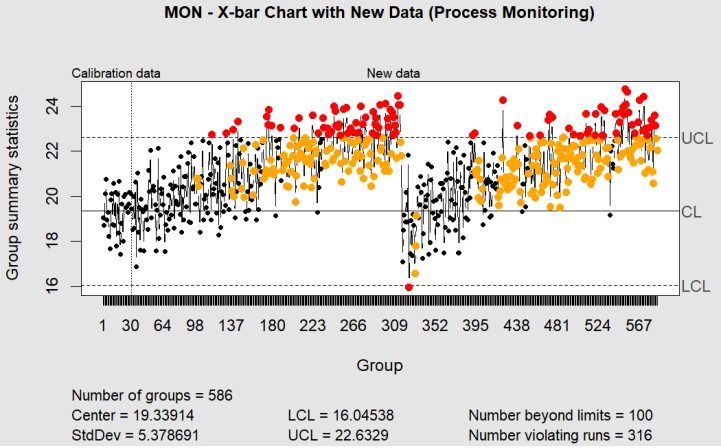

Figure 3.2.9: X-bar Chart for MON Products (All Samples)

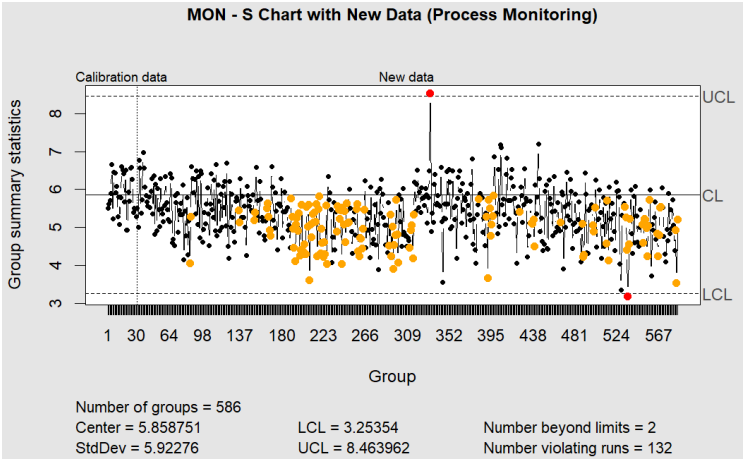

Figure 3.2.10: S Chart for MON Products (All Samples)

LAP Products:

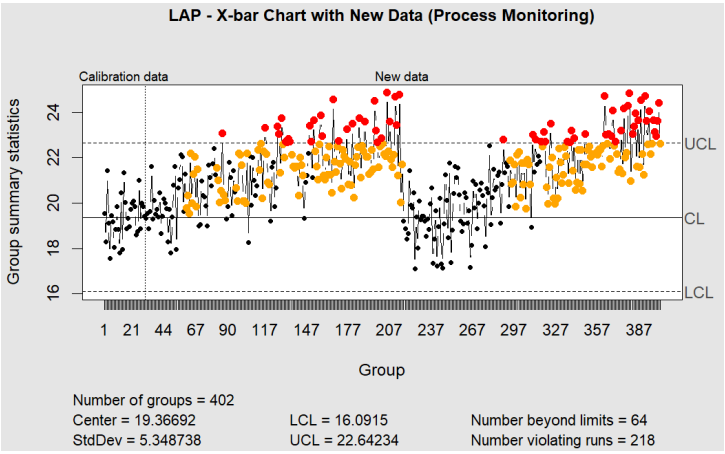

Figure 3.2.11: X-bar Chart for LAP Products (All Samples)

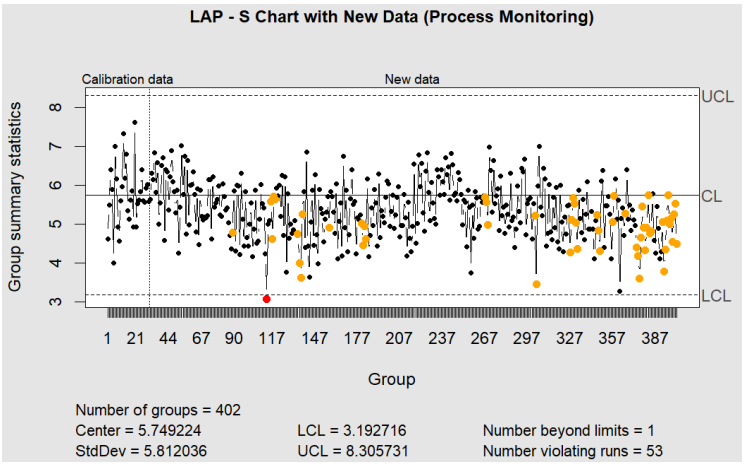

Figure 3.2.12: S Chart for LAP Products (All Samples)

For all the product types above, the X-bar charts generally show a very similar pattern, with an upward shift in the process mean, as well as many samples lying above the upper control limit. This means the overall performance of the company has changed, with the processes no longer being stable.

Regarding the S charts, however, these show better control, with all samples lying within the control limits. Therefore, while variation on a day-to-day basis has not necessarily become worse, the company's overall process has slowly started drifting away from the intended target hours. Therefore, even though the system functions smoothly, albeit at a slower pace, it is clear that it is due to systematic problems in the operations, rather than random variation.

3.3 Process Capability Results

The process capability indices were calculated for the first 1000 deliveries per product type. The table below shows the results of these calculations, stating whether the processes are capable, marginally capable, or not capable at all.

Table 3.3.1: Process Capability Indices per Product Prefix

Product_Prefix	Cp	Cpk	Cpu	Status
SOF	18.144	1.085	35.202	Marginal
CLO	0.922	0.747	0.747	Not Capable
KEY	0.931	0.746	0.746	Not Capable
MOU	0.928	0.741	0.741	Not Capable
MON	0.906	0.721	0.721	Not Capable
LAP	0.917	0.718	0.718	Not Capable

SOF is the only marginally capable process, with the other product types having a Cpk less than 1. This means the processes cannot meet customer requirements, or the mean is possibly not centred.

The significant Cp level for SOF can be interpreted as indicating that the process spread is very small relative to the tolerance range. This is probably because SOF has shorter delivery hours of about 1 hour, compared to other products, which have around 20 hours. The significant differences and process capability results are primarily due to the varying time ranges of the products.

Lastly, even though the set specification range is between 0 and 32 hours, the lower limit of 0 hours does not add any value as delivery hours are usually at least an hour due to the physical operations included in deliveries. Therefore, it is not considered a real operating limit, making the use of the Cpk – which compares how centred the process is between both limits – although still relevant, less useful.

On the other hand, the Cpu, which measures how close the process mean is to the upper specification limit of 32 hours will be more useful for the purpose of this analysis. A higher Cpu value indicates that the process mean is well below 32 hours, meaning that there is a low risk of late deliveries.

It is evident then, that that the SOF product type, with a high Cpu of 35.202 ensures that deliveries consistently occur below the 32-hour limit. In contrast, however, the other product types have

extremely low Cpu values of about 0.7, meaning they are cutting it very fine regarding making the deliveries within the 32-hour limit. The risk of exceeding the delivery time window is very high and can impact customer satisfaction and service reliability quite negatively.

3.4 Control Rule Violations

According to the SPC rules, the control charts were analysed along with additional calculations to establish whether there are any violations:

Rule A identified samples outside of the upper +3 sigma-control limits for all product types

Rule B identified the most consecutive samples of s within the -1 and +1 sigma-control limits for all product types, indicating reasonable control.

Rule C detected four consecutive X-bar samples outside of the upper control limits, the second for all product types.

	Product_Prefix	RuleA_Total	RuleA_First3	RuleA_Last3	RuleB_LongestRun	RuleC_Total	RuleC_First3	RuleC_Last3
1	SOF	0			17	0		
2	CLO	0			7	0		
3	KEY	0			9	0		
4	MOU	0			8	0		
5	MON	0			19	0		
6	LAP	0			10	0		

Table 3.4.1: SPC Rule Violations and Control Performance per Product Prefix

Rule A shows that all processes are stable with no spikes in the S Charts.

Rule B shows strong control, specifically regarding MON and SOF of 19 and 17 longest runs, respectively. This means the processes are consistent over more extended periods.

Rule C triggers no X-bar chart samples above the second control limits, again indicating stable processes.

In summary, although the processes are stable, they cannot meet customer requirements.

Concluding Remarks

In conclusion, the SPC analysis reveals that the delivery processes are stable and under statistical control, with no products signalling specific concern. Nonetheless, the capability calculations show that only SOF is capable of meeting customer specifications, and even so, only marginally. The other product categories are incapable of meeting the VOC requirements of 0-32 hours.

A distinction must be made between the results concerning stability and those concerning capability. All processes are stable and consistent; however, they need further improvement to enhance their abilities by reducing variability to ensure meeting the target.

In the future, the company should focus on altering certain aspects of various processes, ensuring operator consistency, and optimizing delivery schedules, all of which will aid in improving process capability without increasing variability.

Part 4: Risk and Data Correction

4.1 Likelihood of making a Type I Error for A, B, and C

Probability of Type I Error (α): 0.0027 ≈ 0.27 % (See code for specifications of calculations)

4.2 Likelihood of making a Type II Error for a bottle filling process

Type II Error (β): 0.8412 → Probability of missing a real mean shift

Power (1 − β): 0.1588 → Probability of detecting the shift

There is about 84.12 % chance of failing to detect this small mean shift, and about 15.88 % chance that the control chart will correctly detect it.

4.3 Re-analysis after Data Corrections

After correcting the designated errors, the integration between the two new datasets — products_data2025.csv and products_Headoffice2025.csv — was redone. This allowed for a re-analysis of the company's position, using more accurate data.

Data Quality Summary

Data Quality Summary			
Dataset	**Rows**	**Columns**	**Total_NAs**
Sales	100000	9	0
Customers	5000	5	0
Products	60	6	0
Products_Headoffice	1860	6	0

Table 4.3.1: Data Quality Table

As shown in Table 4.3.1 above, the updated datasets show that for all four files – Sales, Customers, Products, and Products_Headoffice – there are no missing values.

Furthermore, the output below, using setdiff(), shows that the incorrect identifications of the various products and their corresponding pricing errors, as mentioned in Section 3.1, have been corrected. This means that the ProductIDs that initially had NA values in Category.y and Selling Price.y (related to the Products_Headoffice file) were corrected.

Output:

[1] "NA011" "NA012" "NA013" "NA014" "NA015" "NA016" "NA017" "NA018" "NA019" "NA020"

[11] "NA021" "NA022" "NA023" "NA024" "NA025" "NA026" "NA027" "NA028" "NA029" "NA030"

[21] "NA031" "NA032" "NA033" "NA034" "NA035" "NA036" "NA037" "NA038" "NA039" "NA040"

[31] "NA041" "NA042" "NA043" "NA044" "NA045" "NA046" "NA047" "NA048" "NA049" "NA050"

[41] "NA051" "NA052" "NA053" "NA054" "NA055" "NA056" "NA057" "NA058" "NA059" "NA060"

Table 4.3.2: Verification Output Showing Incorrect Product IDs Removed after Correction

Data Quality based on Integration of corrected Datasets

Table 4.3.3: Sample of Successfully Joined Product and Head Office Data

CustomerID	ProductID	Category.x	SellingPrice.x	Category.y	SellingPrice.y
CUST1791	CLO011	Clothing	1070.54	Clothing	1070.54
CUST3172	LAP026	Laptop	18711.72	Laptop	18711.72
CUST1022	KEY046	Keyboard	708.18	Keyboard	708.18
CUST3721	LAP024	Laptop	18366.92	Laptop	18366.92
CUST4605	CLO012	Clothing	963.14	Clothing	963.14
CUST2766	MON035	Monitor	6396.18	Monitor	6396.18
CUST4454	MOU052	Mouse	425.14	Mouse	425.14
CUST582	MON032	Monitor	6634.13	Monitor	6634.13
CUST3343	MON040	Monitor	5346.14	Monitor	5346.14
CUST4331	KEY049	Keyboard	752.75	Keyboard	752.75

Table 4.3.3: Data Integration of Corrected Datasets

After correcting the *products_Headoffice* file, the integration of the datasets was redone, with the new table above showing the Category.x and Category.y, as well as the SellingPrice.x and SellingPrice.y (representing the different datasets) to have equal values. There are no missing values, as observed in the initial integrated dataset. This further confirms that the errors of

mismatched product types and selling prices have been corrected, making any further analysis more reliable and valuable to the company.

Summary Statistics using the corrected Product Data

Table 4.3.4: Summary Statistics for Corrected Products_Headoffice2025 Data

Category	Avg_SellingPrice	Avg_Markup	n
Clothing	1019.06	19.96	10
Cloud Subscription	4514.40	20.44	300
Keyboard	4253.03	20.68	320
Laptop	5343.15	20.33	320
Monitor	4607.15	20.67	320
Mouse	4237.41	20.46	320
Software	4046.42	20.14	270

The summary statistics above relate to the corrected information for the updated products_Headoffice2025.csv file. This dataset includes all product lines for each type, providing an accurate product catalog with corresponding selling prices and markups, reflecting the company's offerings.

Products2025 & Products_Headoffice2025: Range and Average of Selling Price and Markup

Products2025

```
  SellingPrice         Markup
Min.   :  350.4   Min.   :10.13
1st Qu.:  512.2   1st Qu.:16.14
Median :  794.2   Median :20.34
Mean   : 4493.6   Mean   :20.46
3rd Qu.: 6416.7   3rd Qu.:25.71
Max.   :19725.2   Max.   :29.84
```

Products_Headoffice2025

```
  SellingPrice         Markup
Min.   :  350.4   Min.   :10.13
1st Qu.:  512.2   1st Qu.:16.14
Median :  794.2   Median :20.34
Mean   : 4493.6   Mean   :20.46
3rd Qu.: 6416.7   3rd Qu.:25.71
Max.   :19725.2   Max.   :29.84
```

The above summary statistics reveal that the two datasets, *Products2025* and *Products_Headoffice2025,* now precisely align compared to the original summary statistics, where there were slight but distinct differences between the datasets. This confirms that the errors regarding data recording have been rectified.

Boxplot: Product Pricing and Markup Spread using updated Datasets

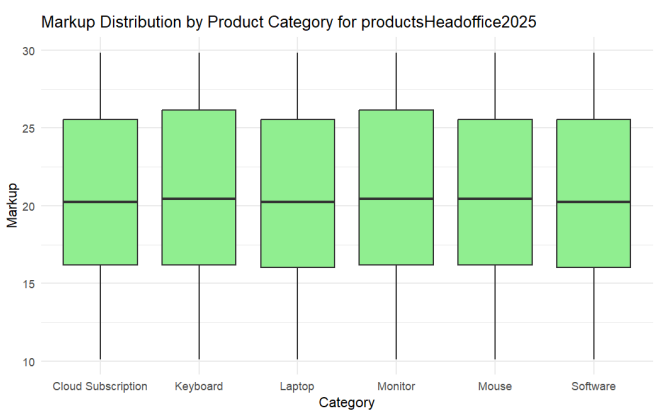

Figure 4.3.1: Boxplot to show Markup Distribution by Product Category: Updated

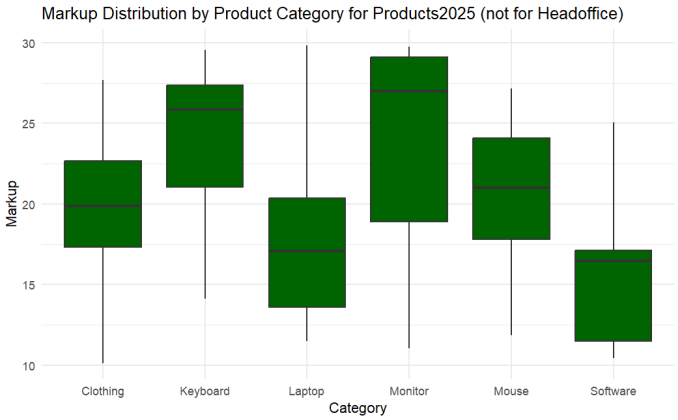

Figure 4.3.2: Boxplot to show Markup Distribution by Product Category: Updated

The box plots above are based on the updated, corrected datasets, which, when compared to the initial plots, are evidently quite different.

Firstly, for the Head Office Markup Distribution, it is clear that the new plot shows much more consistency across the various product types, with each medians markup sitting comfortable around 20-21%, compared to the initial plot, where there was a slightly wider range for the median of about 22-23% for Cloud Subscription and Keyboard, and 19-21% for the other categories. Furthermore, the spread is much more similar across the categories, with no extreme outliers or variance present and a consistent IQR. This contrasts with the initial plot, which showed a more uneven distribution and IQR, indicating inconsistencies in the data entries. Due to this change, it is evident that the new dataset, with its cleaned and corrected values, has aligned markups. This indicates improved control within the department and suggests a more reliable and stable pricing structure, which would help operational efficiency in the future.

For the Products2025 updated boxplot, a great deal of markup variability can be seen between the categories. The range is also much larger, which indicates slightly less control over the pricing structure. This boxplot, unlike that of the Head office products, shows decentralized datapoints, meaning this department is likely more susceptible to environmental changes that could alter the pricing of the products.

The initial boxplot for the Products dataset showed slightly more moderate variations across the categories, compared to the more extreme differences in the markups seen in the new boxplot. This indicates the success of cleaning the data, whereby any inconsistencies, as well as incorrect sales prices and markups, were removed and corrected, allowing for a better understanding of the

company's actual pricing structure. It is also more reflective of how local pricing control works. This updated boxplot reveals a more accurate reflection of the selling price and markup, giving the company the necessary information to make future decisions based on accurate data.

Part 5: Optimising Profit for Coffee Shop Operations

5.1 Outputs for Shop 1

5.1.1 Summary Statistics of TimetoServe

```
   Baristas       ServeTime     
Min.   :1.00   Min.   : 13.00  
1st Qu.:5.00   1st Qu.: 33.00  
Median :5.00   Median : 38.00  
Mean   :5.16   Mean   : 41.22  
3rd Qu.:6.00   3rd Qu.: 45.00  
Max.   :6.00   Max.   :227.00  
```

5.1.2 head() output table for TimetoServe

Baristas <dbl>	ServeTime <dbl>
5	47
6	32
6	32
5	32
4	48
5	43

Table 5.1.1: head() Output

This table shows an example of an output that compares the number of baristas at any given time with the time it took to serve customers, in seconds, with that many baristas on hand. Each row represents one service event, acting as a performance record for different staffing levels.

5.1.3 Profit and Reliability per Number of Baristas

Table 5.1.2: Profit and Reliability per Number of Baristas

Baristas	Mean Service Time	Reliability (% of customers served within 45 seconds)	N	Profit
1	200.16	0.00	417	NA
2	100.17	0.00	3556	15250.51
3	66.61	0.03	12126	35912.06
4	49.98	20.72	29305	65147.14
5	39.96	86.57	56701	103103.14
6	33.36	99.62	97895	149415.97

The table above shows how the mean service time, reliability, and profitability of the coffee shop change as the number of baristas on hand increases.

As can be expected, the mean service time decreases with an increase in the number of baristas available. While only one barista on the staffing team takes an average of about 200 seconds, using six baristas reduces the average to only 33 seconds per customer. This is further confirmed by the boxplot on the right, which visually plots the values shown in the table. It indicates that having six baristas on staff is the optimal number for achieving the lowest average service time, while one barista results in a significantly longer service time compared to other staffing levels.

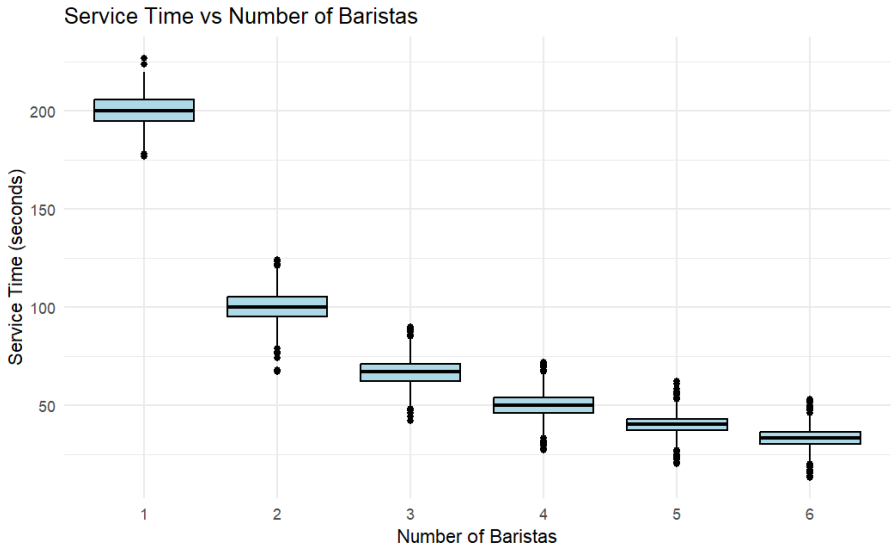

Figure 5.1.1: Boxplot to show Service time vs Number of Baristas for Coffee Shop 1

To calculate reliability, a reliability threshold of 45 seconds was chosen as an average benchmark, based on typical service standards for busy cafes. The reliability of serving customers in under 45 seconds is directly proportional to the number of baristas on staff at the time. With only one to three

baristas available, the reliability is about as low as 0%, i.e., no customers are served within the reliability threshold of 45 seconds. However, the reliability increases to 20.72% as soon as a fourth barista is added to the staff and shows a drastic increase to 86.57% reliability with five baristas. At a staffing level of 6 baristas, a nearly perfect reliability of 99.62% is reached.

The profit is calculated based on a revenue per customer of R30 and a daily cost per barista of R1000. The fewer baristas on staff, the lower the overall profit, with one barista making no profit for the coffee shop, two baristas making only R15 250.51, and six baristas, significantly increasing the profit to R149 415.97, as can be further viewed in the line graph below. An increased service speed allows for more customers, leading to higher total profit. This means the increased costs of adding more baristas are offset by the benefits of nearly 100% reliability and increased throughput.

These results reveal that the optimal staffing level is six baristas, achieving a mean service time of 33 seconds and a high reliability of 99.6% with a maximum profit of R149 416. It is clear that the coffee shop has not yet reached the risk of overstaffing. However, adding more baristas could potentially yield a smaller profit, although this is unlikely to occur in the near future, when considering the clear upward trend in the graph below.

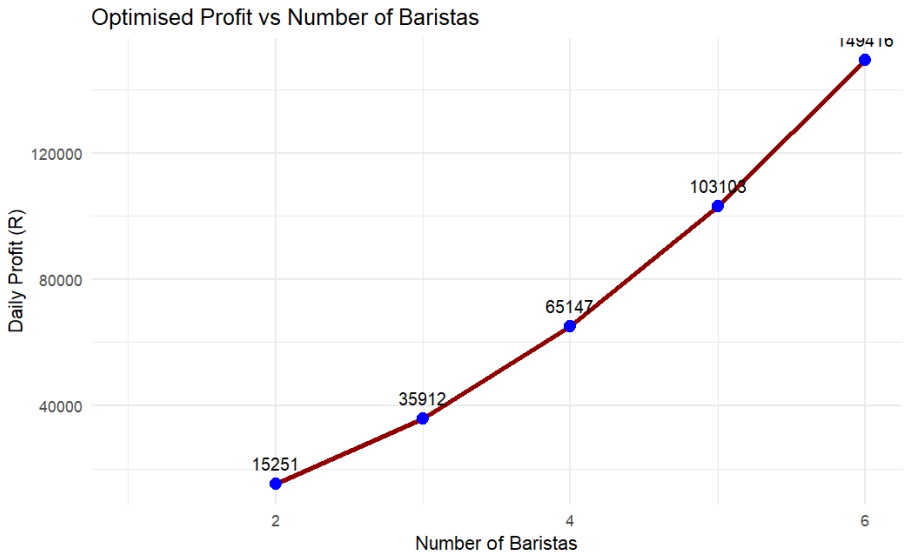

Figure 5.1.2: Line Graph to show Optimised Profit vs Number of Baristas for Coffee Shop 1

5.2 Outputs for Shop 2

Table 5.2.1: Shop 2 – Profit and Reliability per Number of Baristas

Baristas	Mean Service	Reliability (% of customers served within 45 seconds)	N	Profit
1	200.17	0	2196	NA
2	141.51	0	8859	10210.75
3	115.44	0	19768	19453.04
4	100.02	0	35289	30554.72
5	89.44	0	54958	43302.71
6	81.64	0	78930	57496.17

5.2.1 Summary Statistics of TimetoServe2

```
    Baristas        ServeTime
Min.   :1.000   Min.   : 62.00
1st Qu.:4.000   1st Qu.: 83.00
Median :5.000   Median : 89.00
Mean   :4.844   Mean   : 94.32
3rd Qu.:6.000   3rd Qu.:100.00
Max.   :6.000   Max.   :235.00
```

5.2.2 head() output table for TimetoServe2

	Baristas <int>	ServeTime <int>
1	4	97
2	5	88
3	6	82
4	6	83
5	6	87
6	6	84

Table 5.2.2: head() Output for TimeServe2

5.2.3 Profit and Reliability per Number of Baristas

Table 5.2.2 shows that service speed increases with the number of baristas, with a decrease in the mean service speed from about 200 seconds to 82 seconds, as the staffing levels increase from one to six baristas. This can be expected as each additional worker reduces the total workload per person.

Reliability remains at 0%, indicating that even with six baristas, this coffee shop is unable to meet the reliability threshold of serving customers within 45 seconds. Even with the maximum number of baristas, the average service time does not drop below 82 seconds. Shop 2 is not necessarily considered to be failing in its operations, but rather simply requires improvement in its operational efficiency, due to it being generally slower. This could be due to several reasons, whether it be more sophisticated drinks, a higher demand, or poor layout within the coffee shop. Further investigation would benefit Shop 2 to determine the primary reasons for its slower service.

Regarding profit, this increases as do the number of baristas on staff. This once again shows that the increased costs of higher staff levels are offset by the benefits of being able to serve more customers.

The output of the dataset found that 6 baristas proved to be the optimal staffing level to produce maximum profit for Coffee Shop 2. This supports the concept that the optimal choice gives the largest profit, with 6 baristas being the limit.

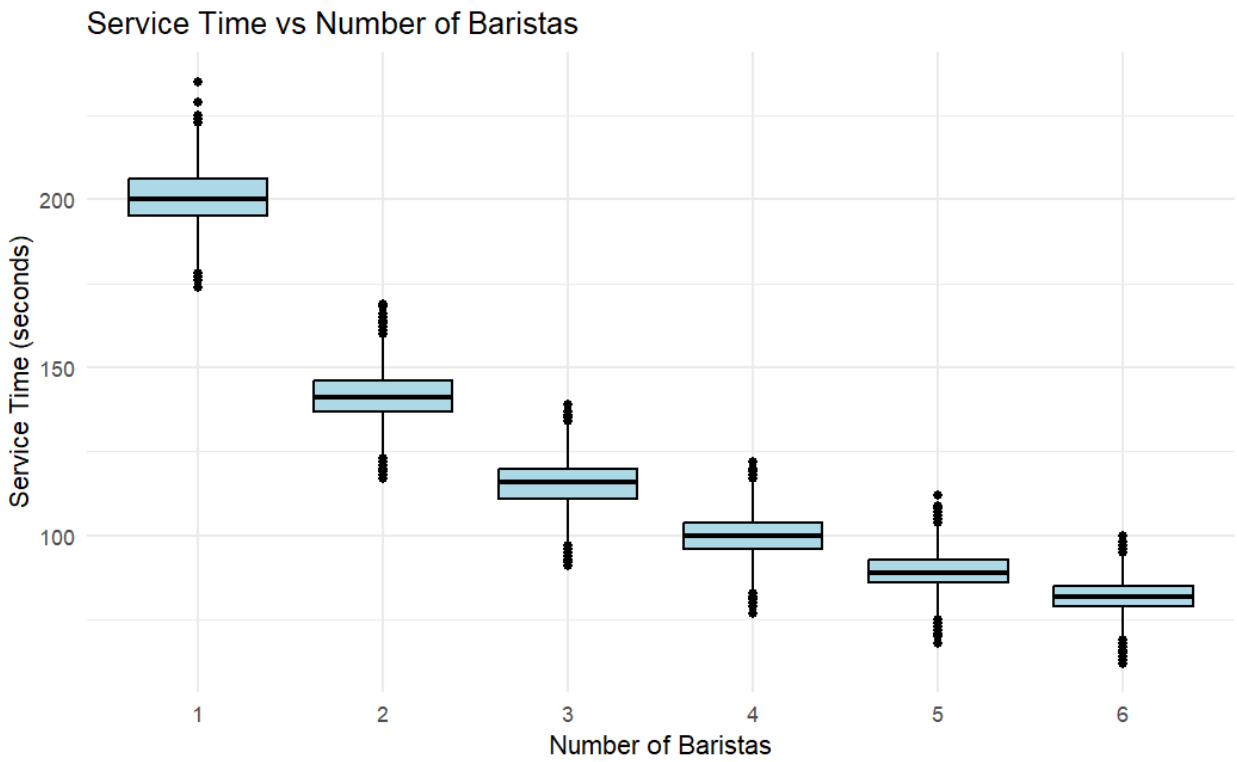

Figure 5.2.1: Boxplot to show Service time vs Number of Baristas for Coffee Shop 2

The boxplot above shows a less drastic decrease in service time between a staffing level of one barista and a staffing level of two baristas, compared to coffee shop 1. The degree of the decline in service time gradually decreases as the number of baristas on staff increase.

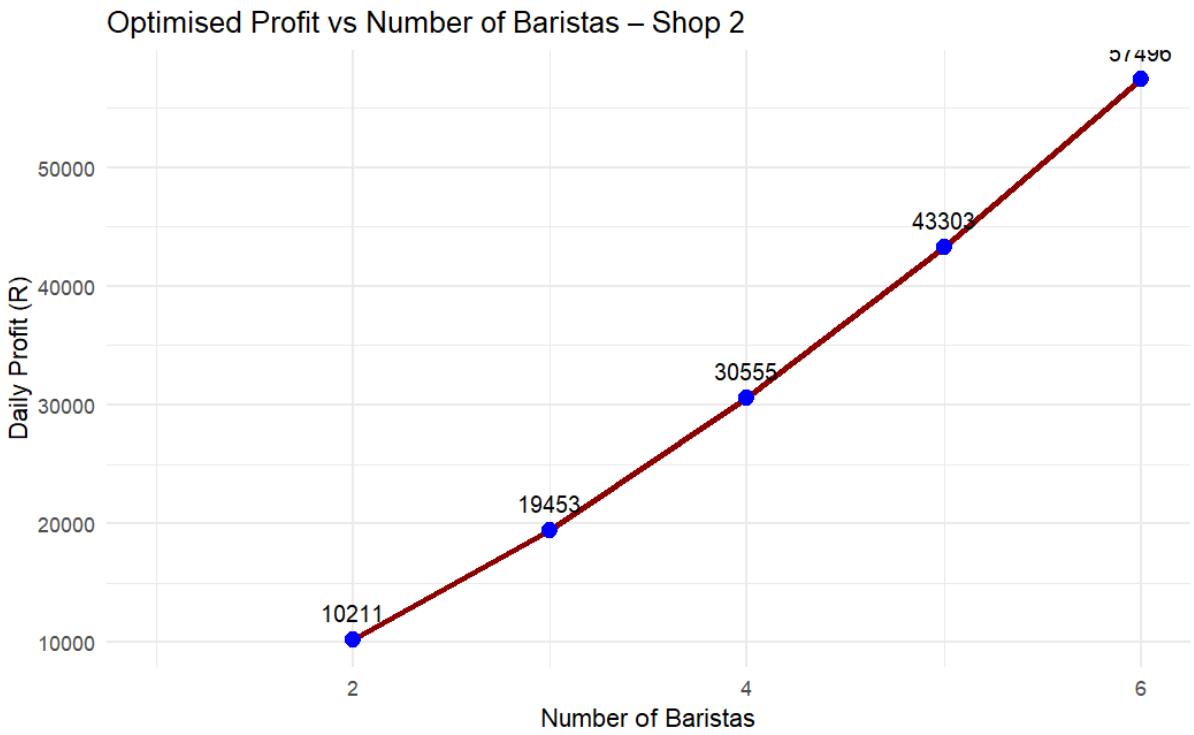

Figure 5.2.2: Line Graph to show Optimised Profit vs Number of Baristas for Coffee Shop 2

Comparison of Performance Levels between Coffee Shop 1 and Coffee Shop 2

The outputs shown above reveal distinct differences in performance levels between the two coffee shops in consideration. Although both coffee shops show similar patterns of a directly proportional relationship between profit and the addition of baristas, Coffee Shop 1 clearly reveals a stronger relationship between not only profit and the number of baristas available, but also the reliability levels.

Compared to Coffee Shop 2, Coffee Shop 1 achieves a faster service time and higher profit at each varying staffing level. The average service time for Coffee Shop 1 drops from about 200 seconds to 33 seconds from a staffing level of one barista to a staffing level of six baristas, whereas Coffee Shop 2, although still showing a decrease in mean service time, reveals a generally lower level of performance, with its lowest average service time dropping to about only 80 seconds with a maximum of six baristas on staff. This indicates Coffee Shop 1 to have a much better level of operational efficiency, compared to Shop 2.

The reliability levels further support Coffee Shop 1 as the better performer of the two shops. Shop 1 reaches nearly a 100% reliability level at six baristas, with shop 2 achieving 0%, regardless of the number of baristas on staff. This clearly indicates Coffee Shop 2 to have certain operational issues,

such as bottlenecks or process inefficiencies, such as mentioned above, that simply increasing the number of baristas does not fix.

Both coffee shops, however, show an increase in profit with the number of baristas available, achieving maximum profit at an optimal staffing level of six baristas – this also being the maximum number of baristas possible. Coffee Shop 2 still achieved an overall lower profit and poor reliability level, indicating inefficiencies that should be investigated.

Therefore, the results conclude that that even though increasing staff theoretically improves service levels, regarding speed and profit achieved, the overall performance depends primarily on how efficient the processes within the coffee shops are. As recommended by the analyst, Coffee Shop 2 should focus on redesigning their process, either through some form of pre-preparation, a change in the design of the workstation layout, or improved staff coordination. This will have a bigger impact on their performance levels than simply changing the staffing levels. Maximum optimisation will require a combination of speed, quality and cost considerations, rather than just the addition of extra baristas.

Part 6: DOE and ANOVA

6.1 ANOVA – SOF Monthly Delivery Times

Research Question

Is there a statistically significant difference in the *average delivery times* for the SOF product type across the 12 months of the year?

Hypothesis

H_0 (null): There is no significant difference in mean delivery times between months.

H_1 (alternative): At least one month's mean delivery time differs significantly.

6.2 Results

ANOVA Results

Table 6.2.1: ANOVA Summary for SOF Monthly Delivery Times

term	df	sumsq	meansq	statistic	p.value
factor(order_month)	11	138.191	12.5628	142.5489	0
Residuals	20737	1827.549	0.0881	NA	NA

The ANOVA results table above shows a F-statistic value of 142.5, with p<0.001.

This means that the differences in mean delivery times between the months are, to a large degree, statistically significant.

```
        Shapiro-Wilk normality test

data:  sample(res, min(5000, length(res)))
W = 0.99175, p-value < 2.2e-16

Levene's Test for Homogeneity of Variance (center = median)
         Df F value Pr(>F)
group    11  0.8002 0.6402
      20737
  Tukey multiple comparisons of means
    95% family-wise confidence level
```

In order to ensure the assumptions made to retrieve valid results from the ANOVA table, were true, assumption checks were completed, using the Shapiro-Wilk normality test and Levene's Test for Homogeneity of Variance.

Both tests confirmed that ANOVA remains valid for this dataset:

For the Shapiro-Wilk normality test, $p < 0.05$ and $W = 0.9918$, meaning there are slight deviations from normality, however due to the large sample size of the dataset, these become negligible.

For Levene's test, $p > 0.05$ and $F = 0.80$, therefore equal variances can be assumed, and the homogeneity assumption holds true.

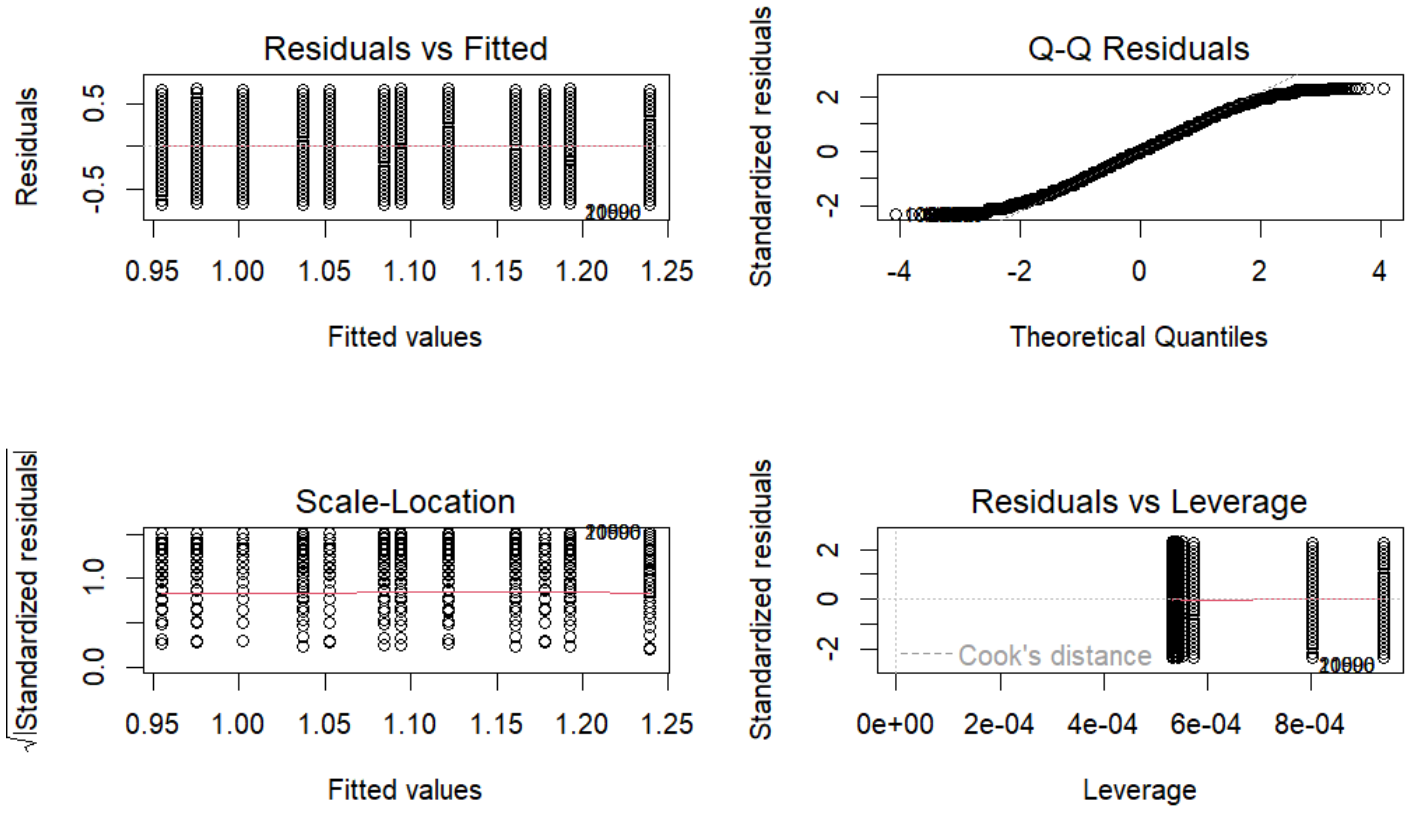

Figure 6.2.1: Diagnostic Plots to Confirm Reliability of ANOVA results

The diagnostic plots were used as a confirmation regarding the reliability of the ANOVA results. The findings were as follows:

Residuals vs Fitted: These data points are spread evenly around zero, indicating no clear pattern and thus confirming the variation is consistent across all months.

Normal Q-Q Plot: This plot shows the data to be quite normal due to most of the points falling close to the line. The few that deviate on either tail are not considered significant enough to be an issue.

Scale-Location Plot: This graph further supports equal variance due to the points once again being spread evenly across the fitted values.

Residuals vs Leverage: There are no outliers or data points that are considered to be influencing the results in any significant manner.

These graphs therefore support the statement that the assumptions for ANOVA are met, thus making the results trustworthy.

Tukey Post-hoc Analysis

	diff	lwr	upr	p adj
2-1	0.01980224	-0.0156140629	0.05521854	0.8030888
3-1	0.04700985	0.0115254888	0.08249421	0.0009118
4-1	0.08153275	0.0461127038	0.11695280	0.0000000
5-1	0.09757769	0.0620318392	0.13312355	0.0000000
6-1	0.12884594	0.0928800714	0.16481181	0.0000000
7-1	0.13917517	0.1036716975	0.17467864	0.0000000
8-1	0.16673657	0.1312100364	0.20226311	0.0000000
9-1	0.20583666	0.1702713955	0.24140193	0.0000000
10-1	0.22278643	0.1872055617	0.25836730	0.0000000
11-1	0.23751733	0.2018335536	0.27320110	0.0000000
12-1	0.28399440	0.2435079347	0.32448088	0.0000000
3-2	0.02720761	-0.0044902044	0.05890542	0.1775429
4-2	0.06173052	0.0301047092	0.09335632	0.0000000
5-2	0.07777546	0.0460088148	0.10954210	0.0000000
6-2	0.10904371	0.0768077682	0.14127964	0.0000000
7-2	0.11937293	0.0876537222	0.15109215	0.0000000
8-2	0.14693433	0.1151893120	0.17867936	0.0000000
9-2	0.18603443	0.1542460629	0.21782279	0.0000000
10-2	0.20298419	0.1711783766	0.23479001	0.0000000
11-2	0.21771509	0.1857941964	0.24963598	0.0000000
12-2	0.26419217	0.2269797233	0.30140461	0.0000000
4-3	0.03452291	0.0028209023	0.06622491	0.0192513
5-3	0.05056785	0.0187253449	0.08241035	0.0000137
6-3	0.08183610	0.0495254002	0.11414679	0.0000000

7-3	0.09216532	0.0603701392	0.12396051	0.0000000
8-3	0.11972672	0.0879057906	0.15154766	0.0000000
9-3	0.15882682	0.1269626448	0.19069099	0.0000000
10-3	0.17577658	0.1438950000	0.20765817	0.0000000
11-3	0.19050748	0.1585110923	0.22250387	0.0000000
12-3	0.23698456	0.1997073341	0.27426178	0.0000000
5-4	0.01604494	-0.0157258833	0.04781576	0.8900792
6-4	0.04731319	0.0150731311	0.07955325	0.0001033
7-4	0.05764242	0.0259190179	0.08936582	0.0000001
8-4	0.08520382	0.0534546111	0.11695302	0.0000000
9-4	0.12430391	0.0925113677	0.15609645	0.0000000
10-4	0.14125368	0.1094436837	0.17306367	0.0000000
11-4	0.15598457	0.1240595186	0.18790963	0.0000000
12-4	0.20246165	0.1652456371	0.23967766	0.0000000
6-5	0.03126825	-0.0011099712	0.06364647	0.0699774
7-5	0.04159748	0.0097336753	0.07346128	0.0012022
8-5	0.06915888	0.0372693821	0.10104837	0.0000000
9-5	0.10825897	0.0763263291	0.14019161	0.0000000
10-5	0.12520874	0.0932587217	0.15715875	0.0000000
11-5	0.13993963	0.1078750590	0.17200421	0.0000000
12-5	0.18641671	0.1490809440	0.22375248	0.0000000
7-6	0.01032923	-0.0220024598	0.04266092	0.9966649
8-6	0.03789063	0.0055336185	0.07024764	0.0072336
9-6	0.07699072	0.0445911882	0.10939025	0.0000000
10-6	0.09394049	0.0615238309	0.12635714	0.0000000
11-6	0.10867138	0.0761418115	0.14120095	0.0000000
12-6	0.15514846	0.1174125947	0.19288433	0.0000000
8-7	0.02756140	-0.0042808481	0.05940365	0.1677817
9-7	0.06666149	0.0347760350	0.09854695	0.0000000
10-7	0.08361126	0.0517084018	0.11551412	0.0000000
11-7	0.09834216	0.0663245704	0.13035974	0.0000000
12-7	0.14481923	0.1075238136	0.18211465	0.0000000
9-8	0.03910009	0.0071889597	0.07101123	0.0036041
10-8	0.05604986	0.0241213405	0.08797838	0.0000006
11-8	0.07078076	0.0387376010	0.10282391	0.0000000
12-8	0.11725783	0.0799404604	0.15457521	0.0000000
10-9	0.01694977	-0.0150218470	0.04892138	0.8532661
11-9	0.03168066	-0.0004054325	0.06376676	0.0567145
12-9	0.07815774	0.0408034892	0.11551199	0.0000000
11-10	0.01473090	-0.0173724889	0.04683428	0.9410846
12-10	0.06120797	0.0238388703	0.09857708	0.0000056
12-11	0.04647708	0.0090099815	0.08394418	0.0029453

Table 6.2.2: TukeyHSD

The first column of the TukeyHSD table above displays the various comparisons of the two months being compared for that specific row. Column 2 shows the difference in the means between the two months specified, with each difference always being calculated by subtracting the second month from the first month shown in Column 1. The third and fourth column displays the lower limit and upper limit, respectively, of the 95% confidence interval. If 0 is not within the range between the lower and upper bound, it is an indication of the difference between the two months being statistically significant. Lastly, the final column shows the adjusted p- values calculated as a means to control Type 1 errors. In other words, it is the probability that the difference between the months occurred purely by chance and is therefore adjusted for multiple comparisons. An adjusted p value greater than 0.05 indicates a significant difference, whereas an adjusted p value less than 0.05 can be considered to reveal no significant difference.

According to the TukeyHSD table based off this dataset, there were numerous pairwise combinations to show significant differences, especially in the months near the beginning and end of the year. This means that delivery times increase from January to December, with the largest difference being found between Month 12 and Month 1, showing a difference in mean value of 0.28 hours.

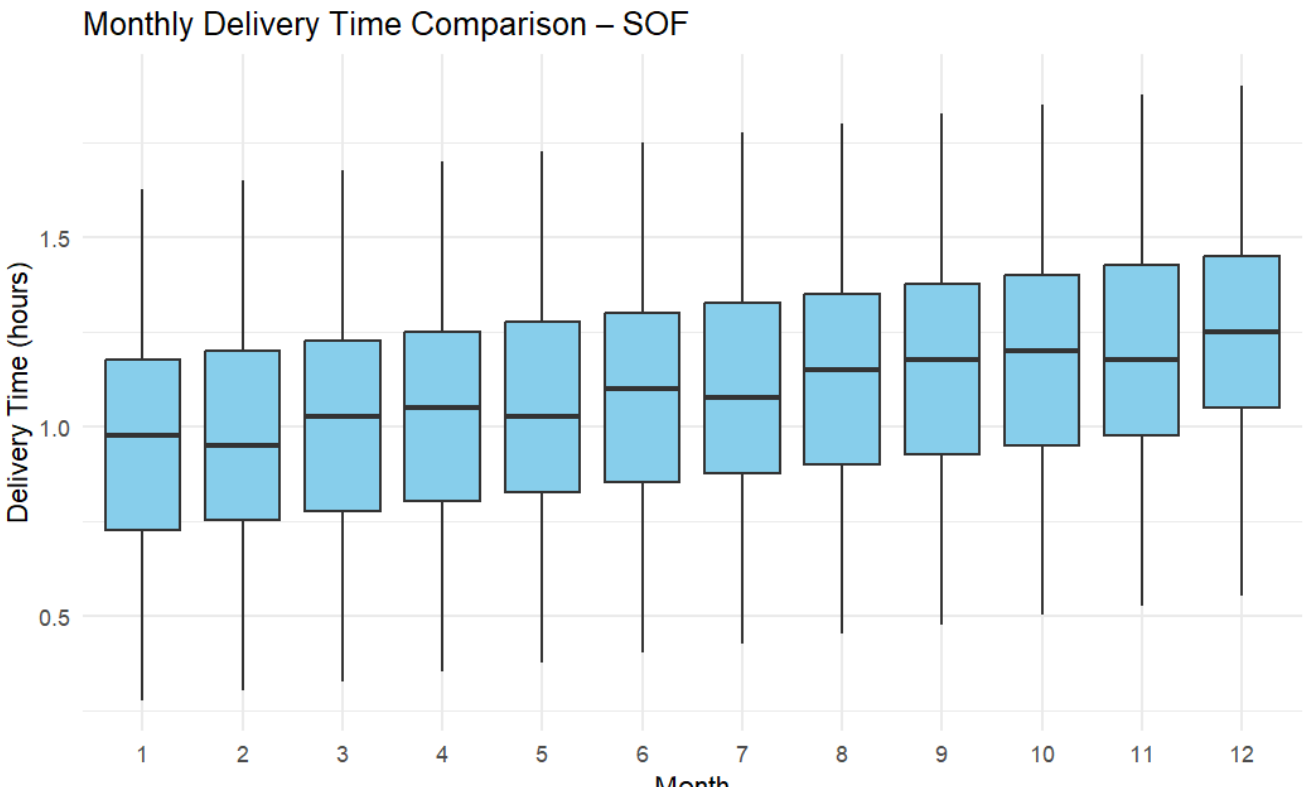

Figure 6.2.2 Boxplot of Monthly Delivery Time Comparison - SOF

The boxplot above supports the findings discussed above, revealing an increasing trend in the delivery times over the year, with an increase in the mean and median delivery times from month 1

to month 12. At the same time, the variance continues to remain relatively consistent, further confirming Levene's test results.

In conclusion, it is clear that delivery times for SOF products differ significantly across the various months. Performance levels are slightly reduced in the later months of the year. It is recommended that whether it be an increase in workload, a result of an increase in demand, or struggling with seasonal fluctuations, these inefficiencies should be focused on improving the overall process of this specific product type.

Part 7: Reliability of a Service

7.1 Estimation of Reliable Service Days

Based on the information provided, reliable service is assumed to occur when 15 or more workers are on duty. This would result in 366 of the 397 days being considered to offer reliable service, thus producing a reliability rate of 92.2% of days per year, when current staffing patterns are considered. Unreliable service occurs approximately 8% of the time, when fewer than 15 staff members are on duty.

7.2 Optimising Profit for the Company

Using the information from the graph and assuming a binomial model (n=16, p), as specified, the maximum likelihood estimation (MLE) for the *weighted* attendance probability is 0.974024. This estimate considers the fact that higher-frequency outcomes, such as 16 workers on 270 days, have a larger impact on the attendance probability.

	scheduled	prob_reliable	reliable_days_per_year	problem_days_per_year	lost_sales_per_year	added_persons	annual_staff_cost	net_change
1	16	0.9363690	341.8	23.2	464506	0	0	0
2	17	0.9909288	361.7	3.3	66220	1	300000	98286
3	18	0.9989599	364.6	0.4	7593	2	600000	-143087
4	19	0.9998986	365.0	0.0	740	3	900000	-436234
5	20	0.9999913	365.0	0.0	63	4	1200000	-735557

Table 7.2.1: Table to show Optimisation Results

The table above shows how the annual profit changes depending on the number of personnel hired, taking into consideration that it cannot be less than 15 people. The net_change column displays the expected net benefit that would result from hiring additional people. A positive net change indicates that hiring that many people would improve the annual profit, whereas a negative net change indicates a reduction in profit. Therefore, based on this table of results, hiring an additional one person to have a total of 17 workers would result in an increase in expected annual profit, as the

benefit of the reduction in lost sales due to the extra personnel would outweigh the costs of extra staff.

Therefore, it is concluded that the optimal number of personnel assigned is 17 workers. Hiring beyond this amount would result in a reduction in profit.

The table below summarizes the findings of this reliability and optimization problem.

Table 7.2.2: Reliability and Profit Optimisation Summary

Metric	Value
Total days observed	397
Reliable days (≥15 workers)	366
Reliability rate	92.2%
Estimated attendance probability ($\hat{p}$)	0.974
Annual expected loss (current)	R 570,025
Annual loss if reliability ≥95%	R 365,000
Estimated annual savings	R 205,025
Recommended action	Hire 1 additional worker to improve reliability

Conclusion

This project highlights the crucial role of statistical tools and data-driven methods in performing effective analysis and implanting process improvements within industrial systems. Complex datasets were created, analysed and interpreted using R programming, as well as through the application of control charts, capability indices and ANOVA tables.

The findings discussed in the report demonstrate the importance of understanding the data, to make informed decisions that support the overall goals of the company by improving processes and enhancing service reliability. The various real-world business examples – such as the Coffee Shop and Car Rental agency optimization models – illustrate how the correct use of data analytics can directly lead to increased profitability and customer satisfaction.

In conclusion, through the use of practical business examples, the report effectively demonstrates the need for analytical thinking, in improving operational efficiency. It further emphasizes the value of combining quantitative, contextual and qualitative analyses to optimize profits, process efficiency and service reliability in any business setting.

References

Hammond, T. (2025). *24 types of charts and graphs for data visualization*. [online] ThoughtSpot. Available at: https://www.thoughtspot.com/data-trends/data-visualization/types-of-charts-graphs [Accessed 25 Sep. 2025].

QA 344 Coding Samples – Provided by lecturer

QA 344 Statistics Notes – Provided by lecturer (no details regarding specifics of the notes)

*NOTE: Use of AI was made for grammar corrections and improved writing