ECSA Project

Quality Assurance 344

Olivia Juliet Weaver

25938681

Contents

Cover page 1
Introduction 5
1.2. Descriptive Statistics 5
1.2.1 Analysis of the Customer Dataset 6
1.2.2 Analysis of the Product Dataset 8
1.2.3 Analysis of Sales 2022 and 2023 dataset 9
3: Statistical Process Control 12
3.1 Control Charts: X-charts and S-charts 12
3.2 Ongoing Monitoring: sample 31+ 12
3.2.1 Mouse (MOU) 12
3.2.2 Keyboard (KEY) 13
3.2.3 Software (SOF) 14
3.2.4 Cloud Subscription (CLO) 15
3.2.5 Laptops (LAP) 16
3.2.5 Monitors (MON) 17
3.3 Process Capability indices Cp, Cpu, Cpl and Cpk 18
3.4 Control issues 21
3.4.1 Rule A 21
3.4.2 Rule B 21
3.4.3 Rule C 22
4. Risk, Data correction and Optimising 23
4.1 Type I (Manufacturer's) Error 23
4.2 Type II (Consumer's) Error for Bottling Filling Process 23
4.3 Product and Head office Data Correction 24
5. Optimising the Company's Profit 25
5.1 Shop 1 Analysis 25
5.2 Shop 2 Analysis 26
6. DOE and MANOVA or ANOVA 26
6.1 Approach 26
6.2 Results and Interpretation 26
7. Reliability of Service 27

7.1 Determining the Amount of Reliable Service 27
7.1 Optimising profit for the Company 28
Conclusion 28
References:.. 28
Appendix A: 30

List of figures

Figure 1: Summary of the product dataset 5
Figure 2: Summary of the customer dataset 5
Figure 3: Summary of the Sale2022and2023 dataset 6
Figure 4: Scatter plot of age vs income 7
Figure 5: Corelation matrix of age and income 7
Figure 6: Density plot of income vs gender 7
Figure 7: Box Plot of income vs gender 7
Figure 8: ANOVA Test Summary 7
Figure 9: Bar graph of customer age distribution 8
Figure 10: Age distribution across cities 8
Figure 11: Printed summary of the product dataset 8
Figure 12: Corrected Product dataset 8
Figure 13: Box plot of different selling prices per category 9
Figure 14: Box plot of markup per category 9
Figure 15: Summary of the merged dataset 9
Figure 16: histogram of quantities per sale 10
Figure 17: line graph of monthly sales trends 10
Figure 18: Correlation matrix of merged dataset 10
Figure 19: scatter plot of delivery hours vs picking hours 11
Figure 20: density plots of quantity sold per year, month and day 11
Figure 21: Bar graph of total revenue per product category 11
Figure 22: Total revenue per customer 12
Figure 23: X-Chart for Mouse 12
Figure 24: S-Chart for Mouse 13
Figure 25: X-Chart for keyboard 13
Figure 26: S-Chart for Keyboard 14
Figure 27: X-Chart for Software 14
Figure 28: S-Chart for Software 15
Figure 29: X-Chart for Cloud Subscription 15
Figure 30: S-Chart for Cloud subscription 16
Figure 31: X-Chart for Laptops 16
Figure 32: S-Chart for laptops 17

Figure 33: X-Chart for Monitors....17
Figure 34: S-Chart for monitors....18
Figure 35: Process Capability indices for Monitors....18
Figure 36:Process Capability indices for Keyboards....19
Figure 37: Process Capability indices for Software....19
Figure 38: Process Capability indices for Cloud Subscription....20
Figure 39: Process Capability indices for Laptops....20
Figure 40: Process Capability indices for Mouse....21
Figure 41: Rule A....21
Figure 42: Rule B....22
Figure 43: Rule C....22
Figure 44: Distribution of original and new bottle filling process....24
Figure 45: Summary of Product ID found in Product data....24
Figure 46: Summary of ProductID found in HeadOffice dataset....25
Figure 47: Line Graph optimising profit....25
Figure 48: Line graph of service time vs number of Baristas....25
Figure 49: Shop 2 line graph optimising profit....26
Figure 50: Shop 2 line graph of service time vs number of baristas....26
Figure 51: Box Plot of delivery hours per month....27
Figure 52: Box Plot of delivery hours per year....27
Figure 53: MANOVA results....27
Figure 54: Two-way ANOVA....27
Figure 55: Reliabilty calculations....28
Figure 56: Staffing Optimisation....28

Introduction

To demonstrate the data analysis, quality control, and process optimisation skills needed to meet the Engineering Council of South Africa's (ECSA) Graduate Attribute 4 (GA4), this following report was created. Based on a variety of dataset which include the Customer, Product, Sales2022and2023 and lastly the HeadOffice. In order to analysis the data the following was done. The Data was cleaned, combined, and examined using R programming to find trends, patterns, and possible discrepancies. Control charts were utilised to track process stability and identify out-of-control conditions. Standard summary descriptive statistics were employed to reveal the structure and variability of the data sets. Using Type I and Type II error probabilities, analysis also involves calculating Process Capability Indices (Cp, Cpk) to assess each process's capacity to produce VOC standards and identify any hazards that occur. The study goes into additional detail on Design of Experiments (ANOVA) as a supplement to SPC analysis in order to determine variations between service performance and profitability. The final step reflects the integration of data analytics with actual decision-making, as optimisation models are created with the goal of maximising daily revenues in service contexts and minimising unreliability costs. The findings are presented in an industrial engineering framework to offer practical suggestions for enhancing overall operational performance, process capability, and product quality.

1.2. Descriptive Statistics

The first stage of the descriptive statistical analysis involved examining the available datasets which were *customers, product_data* and *sales2022and2023* files. To initiate the analysis, the first few rows of each dataset was examined, and a summary was the created of each dataset (see figures 1,2 and 3). This provided an overview of the data structure, variable types, and potential inconsistencies.

```
  ProductID           Category           Description          SellingPrice          Markup
 Length:60          Length:60          Length:60          Min.   :  350.4   Min.   :10.13
 Class :character   Class :character   Class :character   1st Qu.:  512.2   1st Qu.:16.14
 Mode  :character   Mode  :character   Mode  :character   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
```

Figure 1: Summary of the product dataset

```
  CustomerID           Gender                Age             Income            City
 Length:5000        Length:5000        Min.   : 16.00   Min.   :  5000   Length:5000
 Class :character   Class :character   1st Qu.: 33.00   1st Qu.: 55000   Class :character
 Mode  :character   Mode  :character   Median : 51.00   Median : 85000   Mode  :character
                                       Mean   : 51.55   Mean   : 80797
                                       3rd Qu.: 68.00   3rd Qu.:105000
                                       Max.   :105.00   Max.   :140000
```

Figure 2: Summary of the customer dataset

```
  CustomerID          ProductID           Quantity       orderTime          orderDay   
 Length:100000      Length:100000      Min.   : 1.0   Min.   : 1.00    Min.   : 1.0  
 Class :character   Class :character   1st Qu.: 3.0   1st Qu.: 9.00    1st Qu.: 8.0  
 Mode  :character   Mode  :character   Median : 6.0   Median :13.00    Median :15.0  
                                       Mean   :13.5   Mean   :12.93    Mean   :15.5  
                                       3rd Qu.:23.0   3rd Qu.:17.00    3rd Qu.:23.0  
                                       Max.   :50.0   Max.   :23.00    Max.   :30.0  
   orderMonth       orderYear     pickingHours      deliveryHours    
 Min.   : 1.000   Min.   :2022   Min.   : 0.4259   Min.   : 0.2772  
 1st Qu.: 4.000   1st Qu.:2022   1st Qu.: 9.3908   1st Qu.:11.5460  
 Median : 6.000   Median :2022   Median :14.0550   Median :19.5460  
 Mean   : 6.448   Mean   :2022   Mean   :14.6955   Mean   :17.4765  
 3rd Qu.: 9.000   3rd Qu.:2023   3rd Qu.:18.7217   3rd Qu.:25.0440  
 Max.   :12.000   Max.   :2023   Max.   :45.0575   Max.   :38.0460  
```

Figure 3: Summary of the Sale2022and2023 dataset

1.2.1 Analysis of the Customer Dataset

The first dataset which was analysed was the *Customer* dataset. The figure below illustrates the average income per city. This highlighted the income distribution patterns, which is valuable information in terms of future targeted marketing and product pricing strategies. The top two cities with the highest average incomes were Miami and Chicago, which suggest these locations could have stronger purchasing power over other cities.

Table 1: Average income per city

City <chr>	Average_Income <dbl>
Chicago	82244.48
Houston	80248.62
Los Angeles	80475.21
Miami	83346.21
New York	79752.07
San Francisco	79852.56
Seattle	79947.99

The relationship between the customer age and income was examined next using two methods, correlation matrix and scatter plot. As illustrated by figure 5, there is a weak positive correlation between age and income with a Pearson coefficient of 0.16. This indicates that income tends to increase slightly due to age, however, it is not the primary reason income increase. The relationship is not particularly strong. To further confirm this, a scatter plot was plotted seen in figure 4. This was visually shown by the red regression line that trends upwards with a gentle slope. These results suggest that while age has some influence on income levels, other demographic or socioeconomic factors likely play a more substantial role in determining customer income.

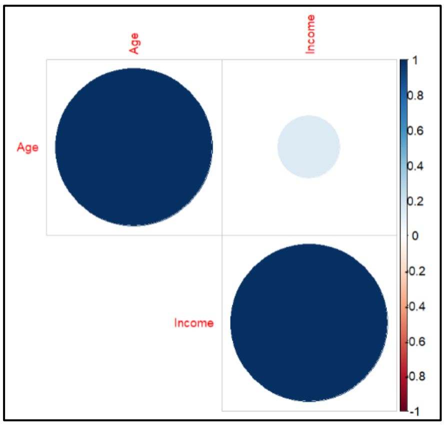


Figure 5: Corelation matrix of age and income

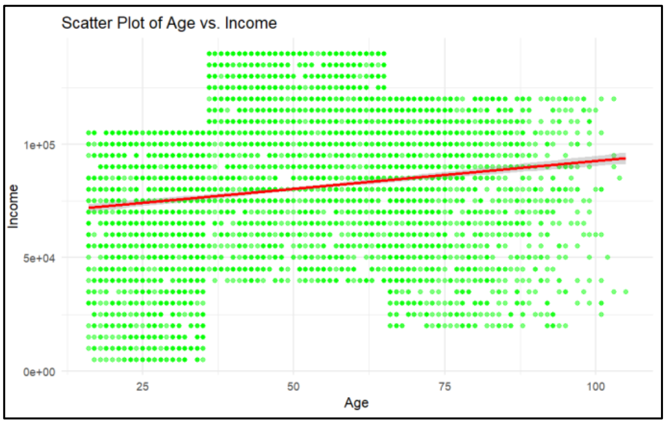


Figure 4: Scatter plot of age vs income

The relationship between income and gender was examined next using density plots, box plots and one-way ANOVA test. The density plot (figure 6) illustrates that the distribution of income across the 3 gender categories, *Female, Male* and *Other.* The shape of the curves is very similar which indicates very little variation in income distribution among genders.

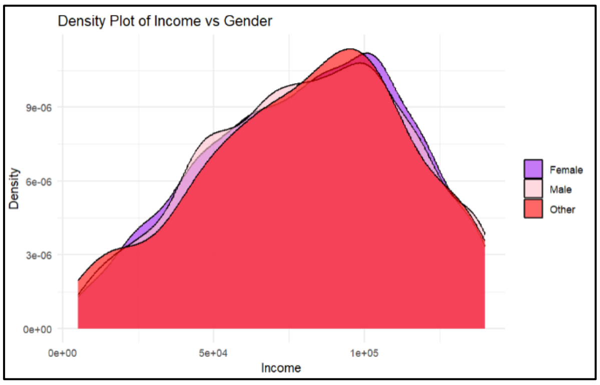


Figure 6: Density plot of income vs gender

The box plot (figure 7) provides a clearer comparison of central tendencies and dispersion across genders. It also reveals that all gender groups have comparable median income levels with slight variation in spread of values. This visual consistency suggests that gender does not play a major role in determining income levels within the dataset.

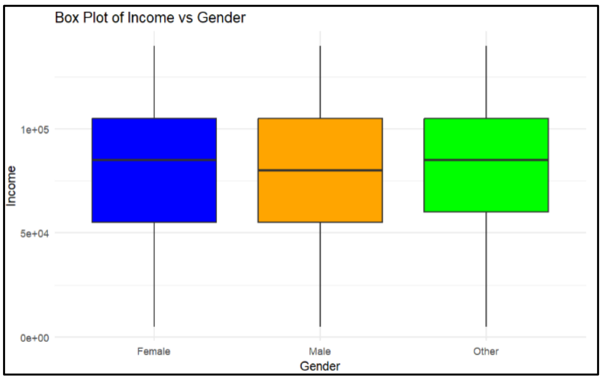


Figure 7: Box Plot of income vs gender

To statistically validate these observations, the ANOVA test was conducted. The results in figure 8, how an F value = 0.002 and p-value =0.98, which further confirms that there is no statistically significant difference between gender and income.

	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		

Figure 8: ANOVA Test Summary

The analysis of the customers age distribution will provide us will provide valuable insight into the demographic composition of the dataset. This will allow us to develop age-specific product development and target marketing strategies.

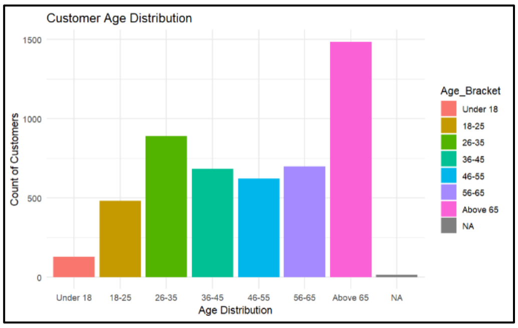


Figure 9: Bar graph of customer age distribution

Based on the bar graph in figure 9, the 65 years and older age bracket contain the highest number of customers (1484 individuals). This means that a significant amount of the customers belongs to an older demographic. This suggests marketing should be target to the preferences and needs of this group.

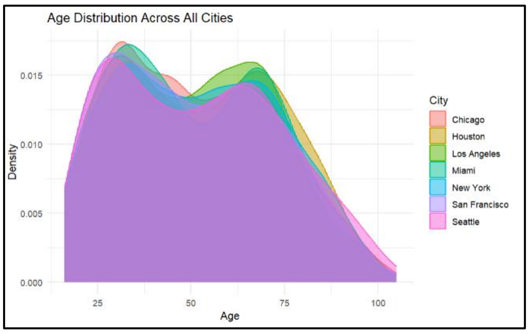


Figure 10: Age distribution across cities

The density plot illustrates the age distribution across the cities. All the density plots have similar shapes which indicates a consistent demographic pattern across cities. This uniformity suggests that the customer age composition does not vary significantly between cities and highlights the dominance of the 65+ age group is a general trend across all cities. Further highlighting the importance of 65+ age group.

1.2.2 Analysis of the Product Dataset

The product dataset was first inspected by printing the first 6 rows. A data-validation check revealed that there was an error in the dataset. It revealed that there was a ProductID-Category misalignment see figure 11.

	ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
1	SOF001	Software	coral matt	511.53	25.05
2	SOF002	Cloud Subscription	cyan silk	505.26	10.43
3	SOF003	Laptop	burlywood marble	493.69	16.18
4	SOF004	Monitor	blue silk	542.56	17.19
5	SOF005	Keyboard	aliceblue wood	516.15	11.01
6	SOF006	Mouse	black silk	478.93	16.99

Figure 11: Printed summary of the product dataset

The inconsistencies were corrected by reconciling each record to link with the ProductID prefix (e.g. SOF – software, LAP – laptop, MON - monitor, KEY – keyboard, MOU – mouse, CLO – cloud subscription). This allowed a reliable basis for analyses (figure 12).

ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
CLO011	Cloud Subscription	burlywood silk	1070.54	16.41
CLO012	Cloud Subscription	azure silk	963.14	10.13
CLO013	Cloud Subscription	chartreuse silk	1067.54	16.80
CLO014	Cloud Subscription	burlywood silk	1083.11	21.25
CLO015	Cloud Subscription	azure silk	728.26	27.70
CLO016	Cloud Subscription	coral bright	959.51	19.55

Figure 12: Corrected Product dataset

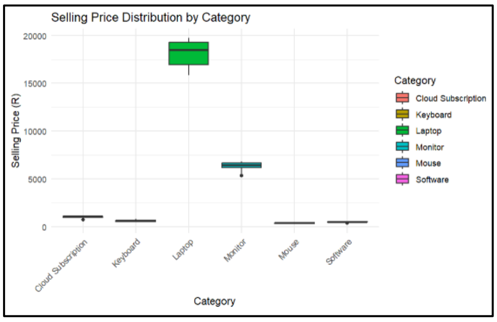


Figure 13: Box plot of different selling prices per category

The boxplot shows (figure 13) that prices are strongly stratified by category. Laptops have the highest median prices with a wide range of prices; monitors follow. Cloud subscription, keyboards, mice and software cluster far lower with tight dispersion. This indicates a more standardised pricing in those products.

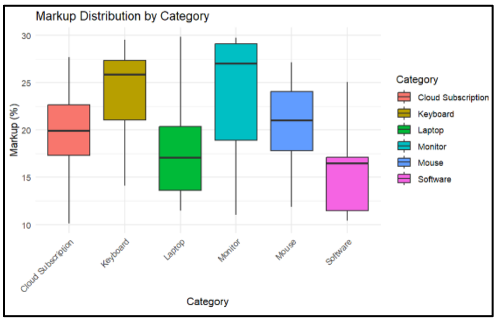


Figure 14: Box plot of markup per category

As seen in figure 14, the Markup % does not mirror price. Monitors and keyboards show the highest median markup while software is the lowest. Markup spread is widest for monitor suggesting mixed positioning between standard and premium products

1.2.3 Analysis of Sales 2022 and 2023 dataset

To ensure a consistent analysis across the 2022-2023 sales dataset a unified and analysis-ready dataset was created by integrating the sales, customer and product tables. The sales data with customer demographics and product data ensure that a full analysis can be done. The sales were then ordered in ascending order to establish a reliable temporal baseline.

```
  ProductID           CustomerID           Quantity       orderTime         orderDay    
 Length:100000      Length:100000      Min.   : 1.0   Min.   : 1.00   Min.   : 1.0  
 Class :character   Class :character   1st Qu.: 3.0   1st Qu.: 9.00   1st Qu.: 8.0  
 Mode  :character   Mode  :character   Median : 6.0   Median :13.00   Median :15.0  
                                       Mean   :13.5   Mean   :12.93   Mean   :15.5  
                                       3rd Qu.:23.0   3rd Qu.:17.00   3rd Qu.:23.0  
                                       Max.   :50.0   Max.   :23.00   Max.   :30.0  
   orderMonth       orderYear     pickingHours      deliveryHours       Gender         
 Min.   : 1.000   Min.   :2022   Min.   : 0.4259   Min.   : 0.2772   Length:100000     
 1st Qu.: 4.000   1st Qu.:2022   1st Qu.: 9.3908   1st Qu.:11.5460   Class :character  
 Median : 6.000   Median :2022   Median :14.0550   Median :19.5460   Mode  :character  
 Mean   : 6.448   Mean   :2022   Mean   :14.6955   Mean   :17.4765                     
 3rd Qu.: 9.000   3rd Qu.:2023   3rd Qu.:18.7217   3rd Qu.:25.0440                     
 Max.   :12.000   Max.   :2023   Max.   :45.0575   Max.   :38.0460                     
      Age             Income           City             Category          Description       
 Min.   : 16.00   Min.   :  5000   Length:100000      Length:100000      Length:100000     
 1st Qu.: 33.00   1st Qu.: 55000   Class :character   Class :character   Class :character  
 Median : 51.00   Median : 85000   Mode  :character   Mode  :character   Mode  :character  
 Mean   : 51.57   Mean   : 80699                                                           
 3rd Qu.: 69.00   3rd Qu.:105000                                                           
 Max.   :105.00   Max.   :140000                                                           
  SellingPrice        Markup     
 Min.   :  350.4   Min.   :10.13  
 1st Qu.:  493.7   1st Qu.:16.18  
 Median :  627.9   Median :20.44  
 Mean   : 3243.8   Mean   :20.42  
 3rd Qu.: 5346.1   3rd Qu.:25.56  
 Max.   :19725.2   Max.   :29.84  
```

Figure 15: Summary of the merged dataset

By looking at all the means and median, one can identify patterns in the data in terms of skewing. The median is significantly lower than the mean of 13.5 when looking at order quality. This suggests a right-skewed sales distribution. It also indicates that there are a few very high-value transactions mixed in with many low-value transactions. When compared to the mean, the standard deviation is similarly large, indicating a significant amount of variation in the sales. As a result, there are a lot of low sales, and some much bigger sales distort the average.

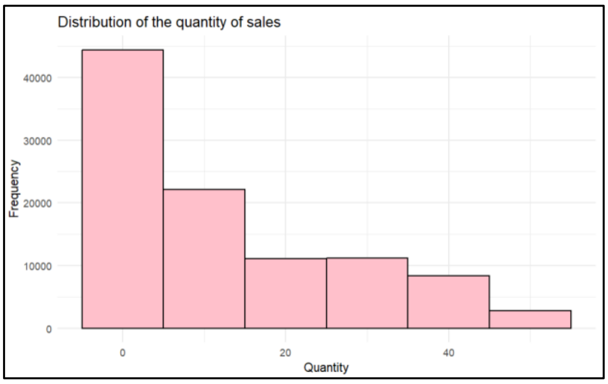


Figure 16: histogram of quantities per sale

The histogram of the quantity confirms further that the merged data is right-skewed. The distribution of majority of sales fall within lower range while higher sales are less common.

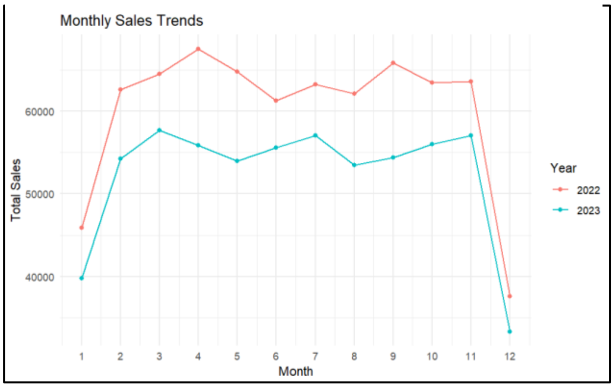


Figure 17: line graph of monthly sales trends

The line graph shows the monthly sales trends for 2022 and 2023. The sales decrease from 2022 to 2023. Sales in both years increase in February and remain relatively stable until November and then decrease in December.

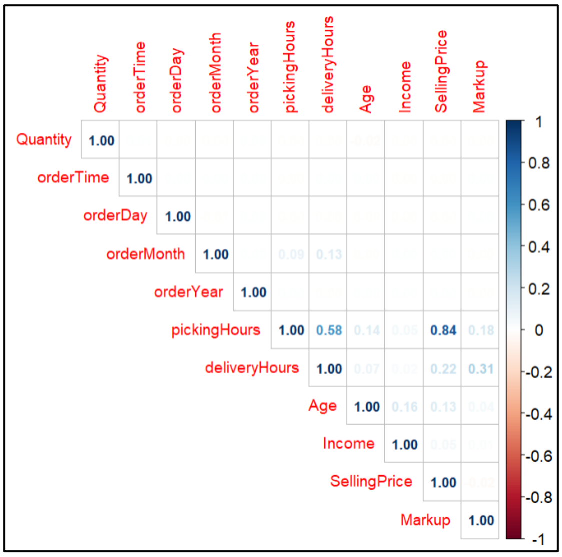


Figure 18: Correlation matrix of merged dataset

The correlation matrix shows that most variables are weakly related, this indicates that there is independence among factors. However, there is a strong association between picking hours and selling price (r=0.84). This suggests that higher priced items take longer to process. Picking and delivery hours (r=0.58) indicating that longer picking times lead to longer deliveries.

Figure 19: scatter plot of delivery hours vs picking hours

The scatter plot compares picking hours and delivery hours by product type for 2022 and 2023. The Scatter plot shows that's all deliveries fall between 10-30 hours, while picking times cluster in distinct bands by category. The weak visual association between picking and delivery hours suggests that warehouse processing and delivery are largely independent. This suggests that improvements in one is unlikely to improve the other. Operationally, this suggest that one should focus on prioritising high effort products such as monitors and laptops.

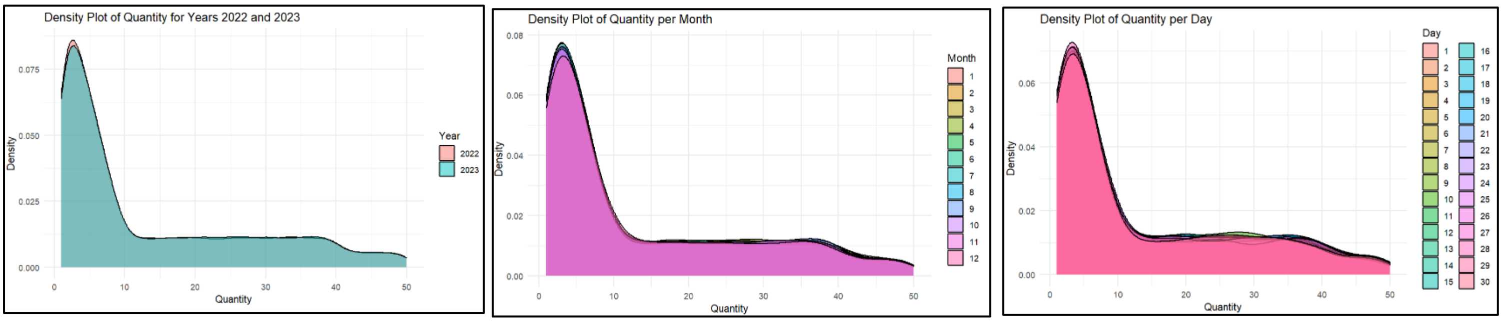


Figure 20: density plots of quantity sold per year, month and day

It is clear from the density plots of quantity by year, month, and day that the ordered quantities don't change across various time periods. This is crucial because, when predicting future demand, steady demand is preferred. Making as many sales as possible, preventing stock outs, and reducing inventory and handling expenses all depend on accurately predicting future demand.

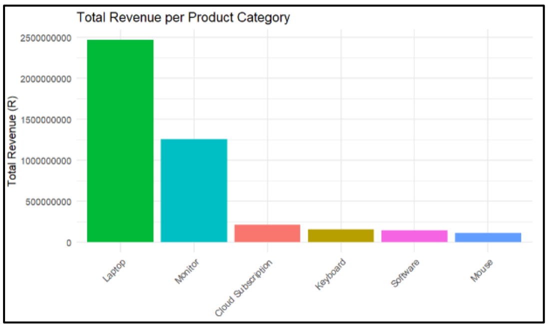
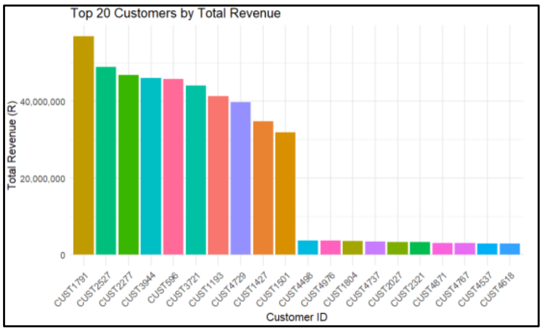


Figure 21: Bar graph of total revenue per product category

The bar chart of total revenue by product category which is strongly right skewed. The revenue is concentrated in laptops and monitors. This concentration is consistent with earlier selling price, categories with much higher average selling price dominant the revenue.



The bar chart of the top 20 customers exhibits a Pareto pattern. A few accounts contribute a large proportion of revenue, with a sharp drop after the tenth customer. The concentration also implies that there is marketing opportunity as well as risk in terms of dependency.

Figure 22: Total revenue per customer

3: Statistical Process Control

3.1 Control Charts: X-charts and S-charts

The X-charts and S-charts for each product type are included in Appendix A. No points depart from the control limits, and the S charts in each graph have a standard deviation of about 5.4. The average delivery time, as seen in the X-bar chart, is about 19 hours. No point exceeds the predetermined control boundaries. For the charts in 3.2, the centre lines, outer control limits, 2-sigma control limits, and 1-sigma-control limits are determined using these first 30 samples.

3.2 Ongoing Monitoring: sample 31+

3.2.1 Mouse (MOU)

There is a distinct drift pattern visible in the $\bar{X}$-chart. The subgroup means values from samples roughly 1–450 show an upward trend from around 19–22 hours to roughly 24–26 hours. There are occasional peaks near +3σ and frequent excursions above the +2σ range. The series ends with a step down (reset to an earlier start position) at about 450, followed by another slow increase that once more approaches the high zones. A non-stationary process mean and likely special-cause inputs over time are indicated by this behaviour.

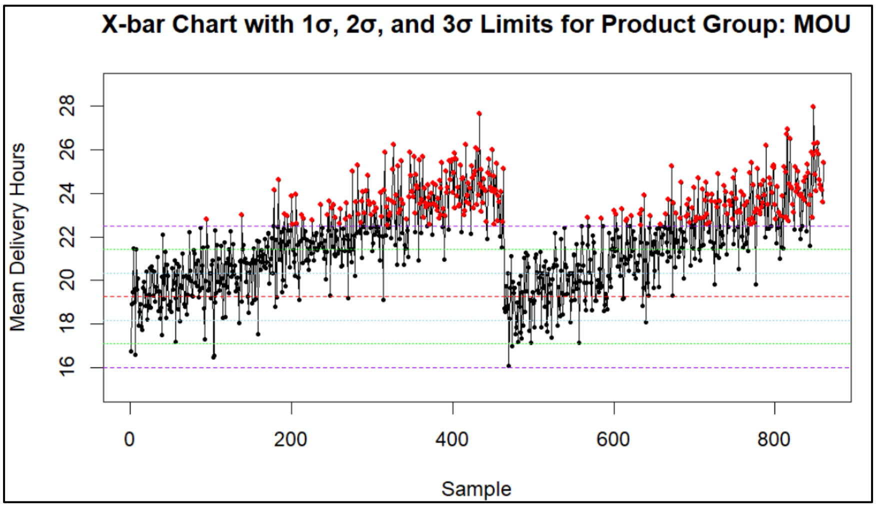


Figure 23: X-Chart for Mouse

Aside from a single significant outlier that exceeds the upper control limit close to sample roughly 590, the S-chart is generally consistent around s = 5.5–6.5h. Otherwise, dispersion stays within limits. In summary, while spread is mostly under control, location is not. Operational activity should concentrate on determining the causes of the solitary high-variability incident and the upward drift in mean delivery time.

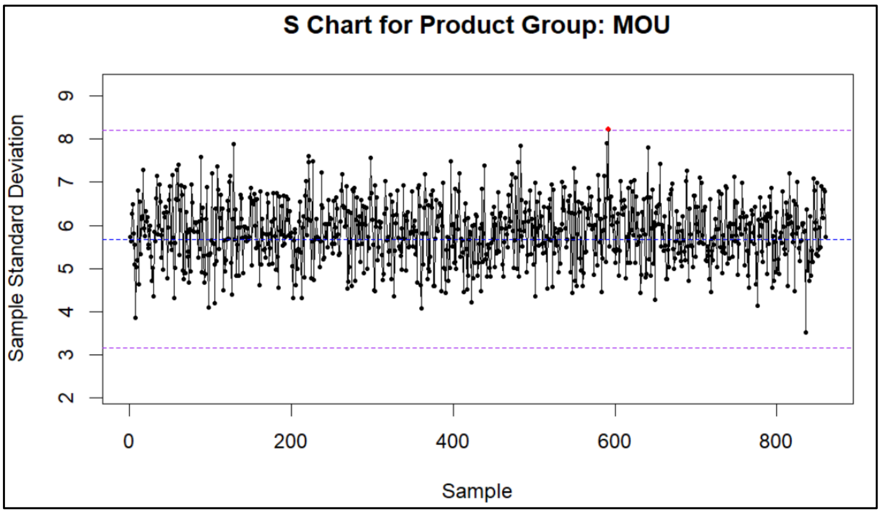


Figure 24: S-Chart for Mouse

3.2.2 Keyboard (KEY)

Like MOU, the $\bar{X}$-chart shows a sustained upward drift into the +2σ area from approximately 18 to 21 hours, followed by a mid-sequence downward step (around sample roughly 400) and a renewed ascent back into the high zones. Many subgroups exceeding +3σ are indicated by red dots. Thus, indicating persistent mean shifts rather than noise.

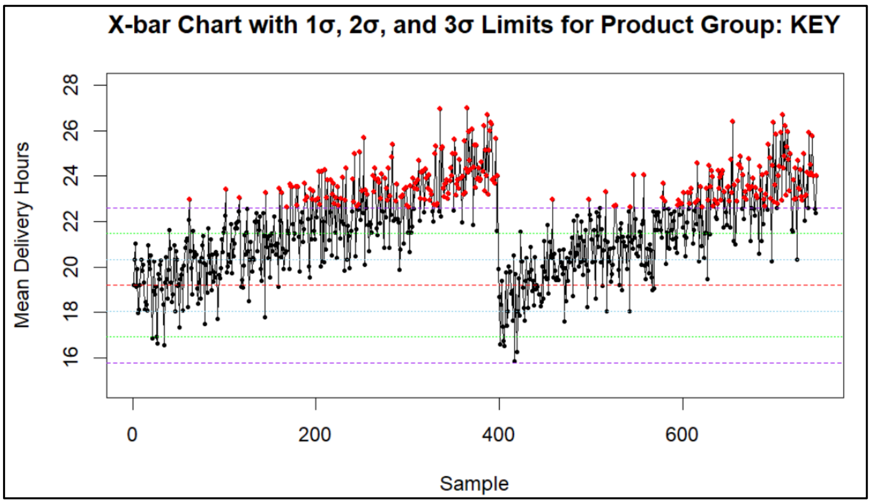


Figure 25: X-Chart for keyboard

There are no clear runs or cycles in dispersion, and the S-chart stays centred at s = 5.8–6.0 h with no violations of the ±3σ bounds. The process mean is unstable, which is consistent with the capacity result (Cp ≈ 1.33, Cpk < 1 due to poor centring), however, variability is within control.

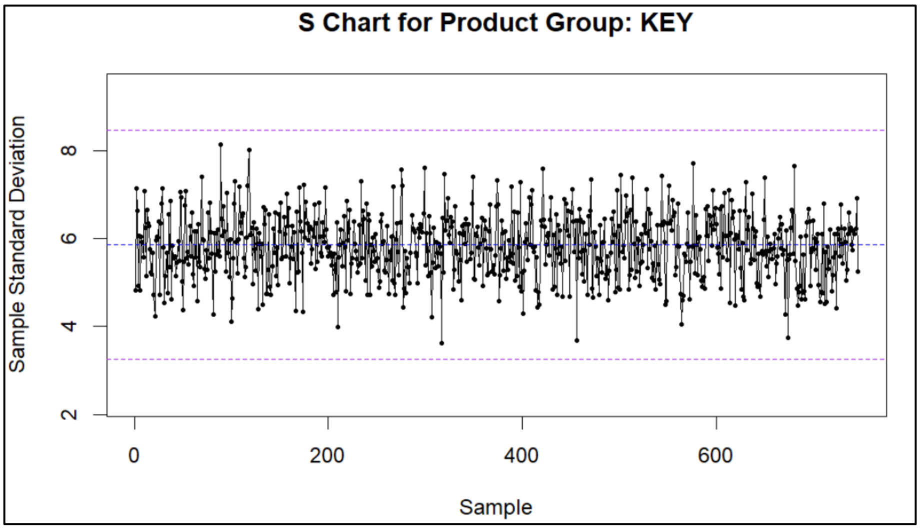


Figure 26: S-Chart for Keyboard

3.2.3 Software (SOF)

The X̄-chart shows a noticeable step drop in the middle of the period, followed by a slow rebound to the initial level, and is closely grouped around 1.0 to 1.2 hours. Most samples fall well outside of the ±3σ area and within the ±2σ band. For the specified LSL/USL, this is a very competent, well-centred procedure (0–32 h).

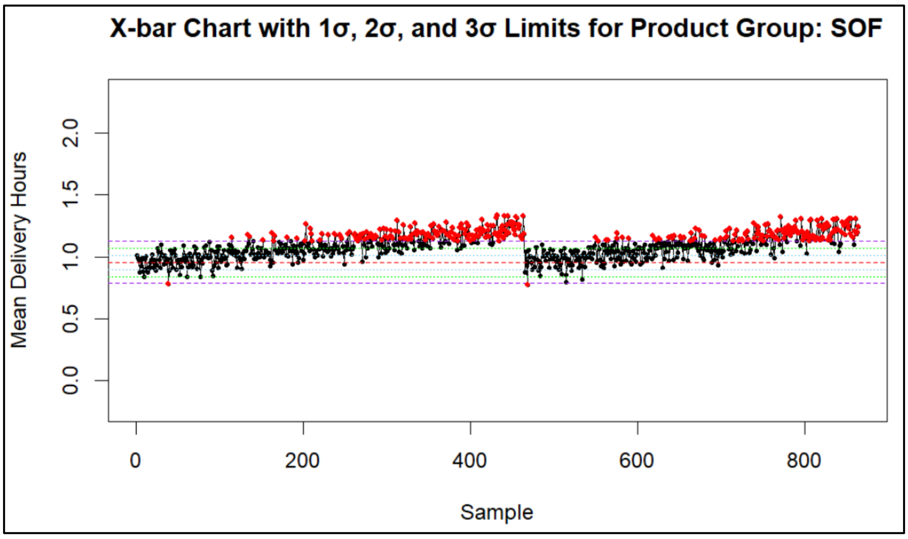


Figure 27: X-Chart for Software

Throughout, the S-chart stays remarkably low and constant (s = 0.25–0.4 h), with every point easily falling inside the control ranges. The charts support your very strong capability indices (very large Cp and Cpk), which show that SOF can meet the VOC with a wide margin.

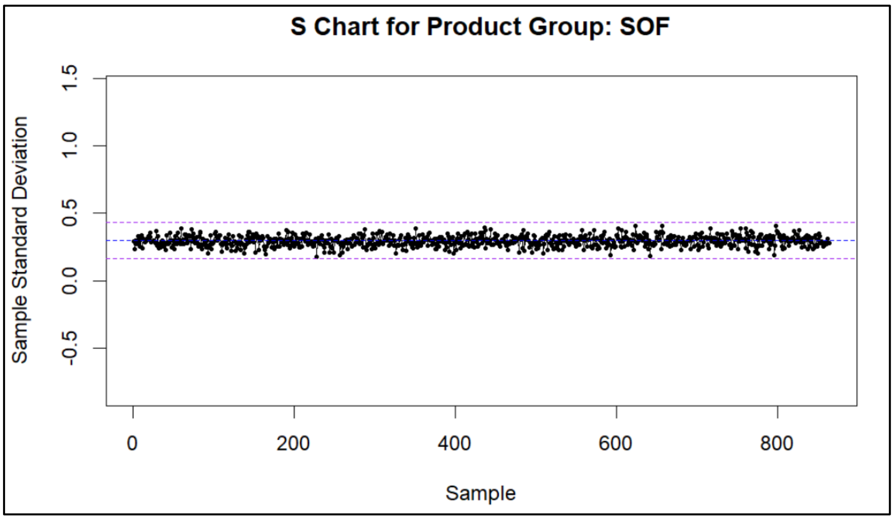


Figure 28: S-Chart for Software

3.2.4 Cloud Subscription (CLO)

With numerous points over the +3σ band, the X̄-chart displays a progressive increase in subgroup means from the upper teens into the low to mid-20s. Midway through the series, there is a drop down, and then there is a gradual increase again. This pattern suggests ongoing pressure on delivery times.

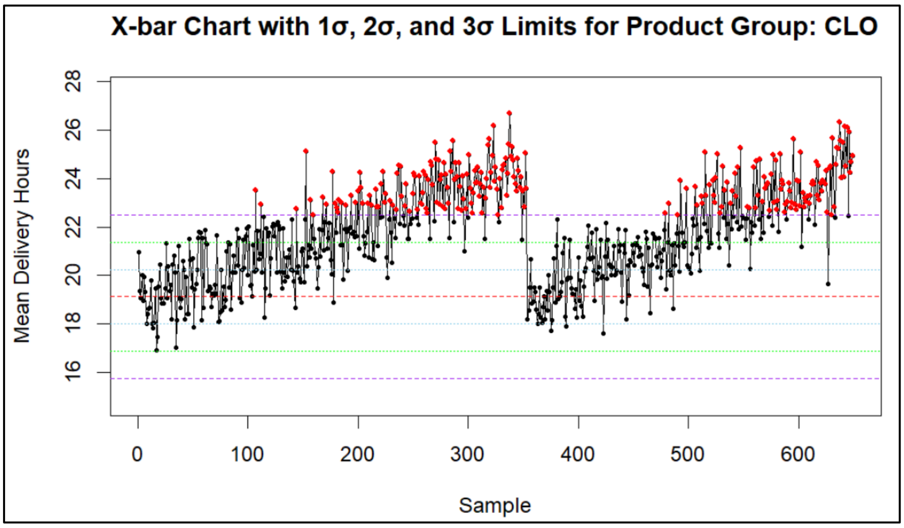


Figure 29: X-Chart for Cloud Subscription

With healthy containment within the control limits, the S-chart remains centred at s ≈ 5.8–6.0 h. The combination of shifting means and stable dispersion fits the capability picture which is that CLO cannot combat the VOC without re-centring the process because the spread is sufficient, but the centring is inadequate (Cpk < 1).

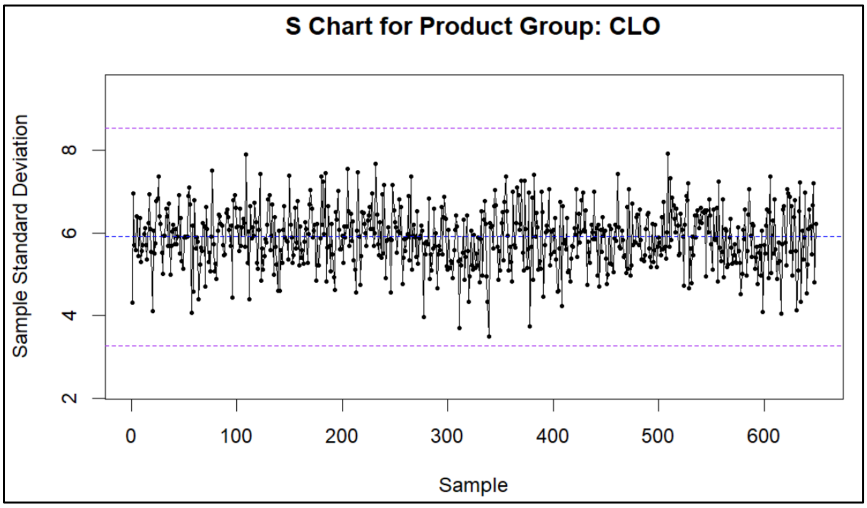


Figure 30: S-Chart for Cloud subscription

3.2.5 Laptops (LAP)

With frequent red spots above +3σ, the X̄-chart once more shows an upward mean drift, a mid-series downward step, and a subsequent climb back into the top control zones. This demonstrates clearly how special-cause behaviour influences the process mean over time.

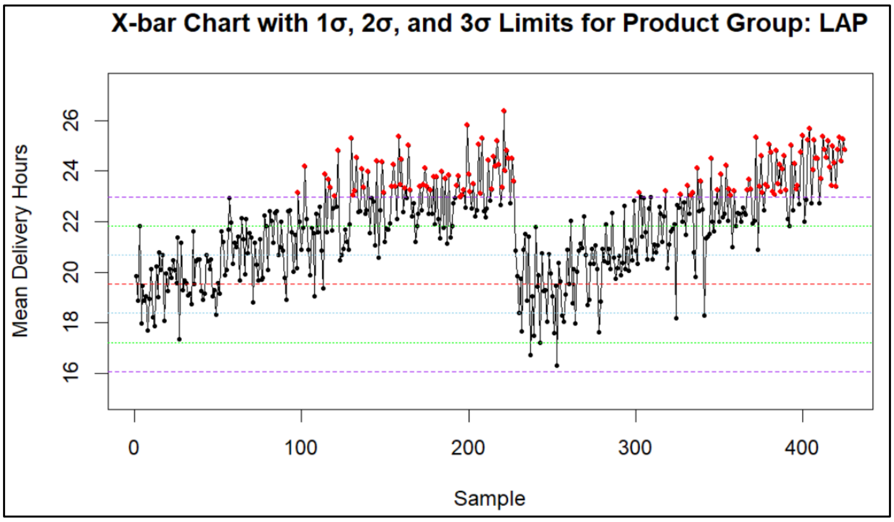


Figure 31: X-Chart for Laptops

There is a single low outlier (one subgroup at the –3σ limit), but no consistent pattern of instability in distribution. The S-chart is primarily contained near s ≈ 5.7–6.2 h. The main problem is mean instability which is seen throughout the other categories.

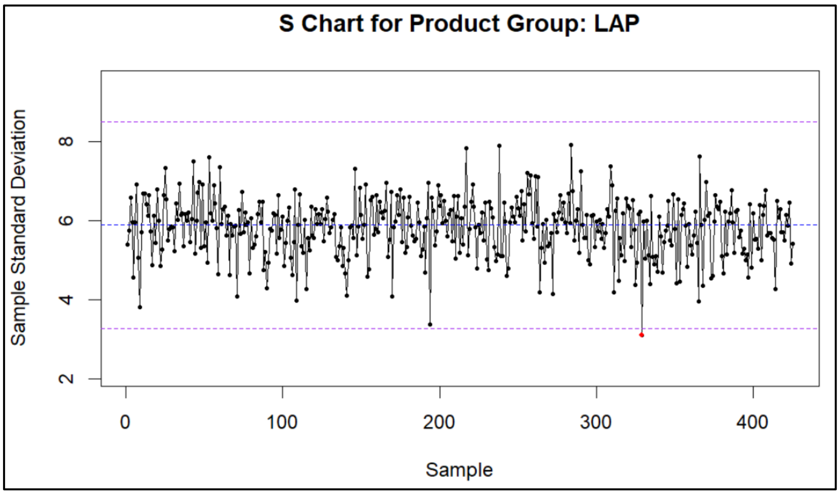


Figure 32: S-Chart for laptops

3.2.5 Monitors (MON)

Like other product categories, the Phase II X̄-chart shows a consistent increase in subgroup means from around 18 to 20 hours into the upper zones which is followed by a noticeable decline around sample ~340 and a new ascent that once more reaches the +3σ region.

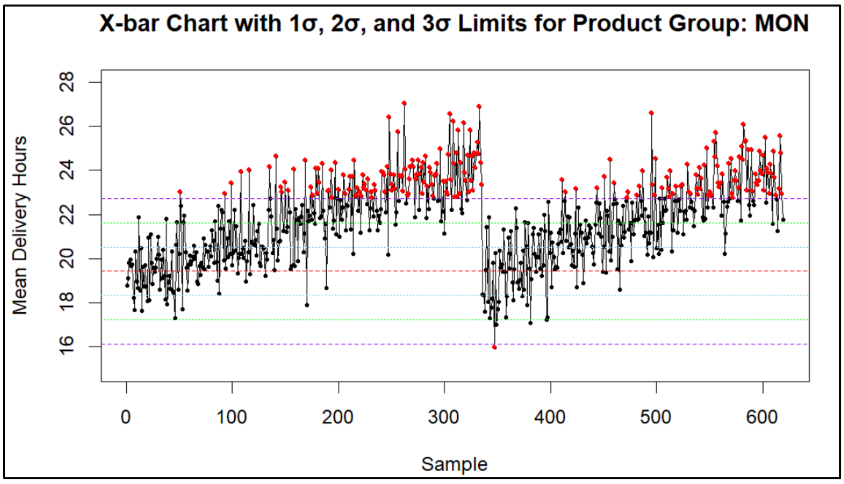


Figure 33: X-Chart for Monitors

On the other hand, the S-chart behaves well. The dispersion is still centred at s = 5.8–6.1 h, with no continuous runs or cycles that fall above/below the ±3σ bounds. The spread is under control, but the process mean is unstable.

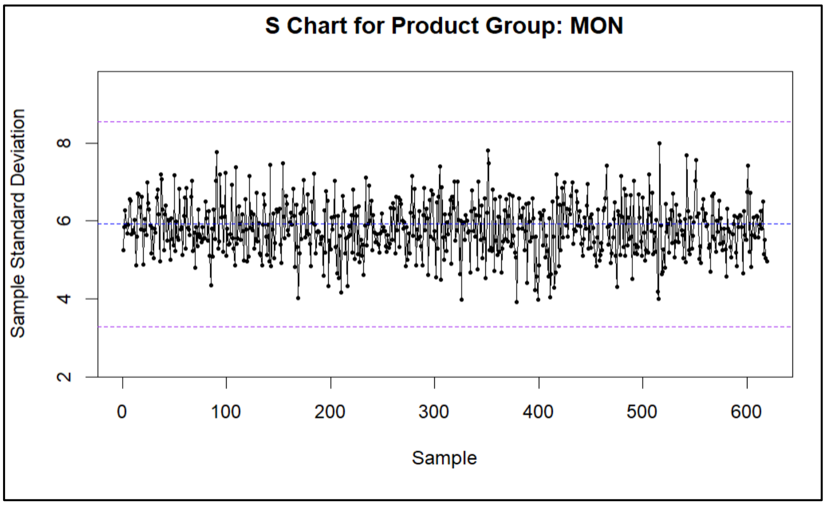


Figure 34: S-Chart for monitors

3.3 Process Capability indices Cp, Cpu, Cpl and Cpk

For Monitors (MON), the Cp value is 0.89 this indicates that the process spread is below the acceptable benchmark. Additionally, the Cpk value is 0.70 demonstrates that the process mean is not well-centred. The process stays stable below the lower limit, as indicated by the Cpl value of 1.08, while the CPU value of 0.70 indicates frequent risks of breaching the upper limit. The VOC cannot be met by MON (Monitors) since stable and predictable supply times require improvements in both process centring and spread.

```
Product Category: MON
Mean Delivery Hours: 19.41 hours
Standard Deviation: 6 hours
Cp: 0.89
Cpu: 0.7
Cpl: 1.08
Cpk: 0.7

Product Category MON is NOT capable of satisfying the VOC.
```

Figure 35: Process Capability indices for Monitors

A Cp value of 0.92 for keyboards (KEY) indicates that the process has a medium chance of meeting the specification constraints. However, a Cpk score of 0.73 suggests that the mean delivery time is not centred, creating irregular performance. A Cpl score of 1.10 suggests steady performance below the lower limit, whilst a Cpu value of 0.73 indicates that deliveries occasionally above the upper limit. The process is nearly capable, but in order to continuously satisfy consumer expectations, average delivery times need to be properly managed.

```
Product Category: KEY
Mean Delivery Hours: 19.28 hours
Standard Deviation: 5.82 hours
Cp: 0.92
Cpu: 0.73
Cpl: 1.1
Cpk: 0.73

Product Category KEY is NOT capable of satisfying the VOC.
```

Figure 36:Process Capability indices for Keyboards

The Cp score of 18.14 for software (SOF) indicates a very tight and efficient process because the process variation is incredibly small when compared to the specification limitations. With a Cpk value of 1.08, the process appears to be nearly within the bounds, albeit just below the usual capability threshold of 1.30. The Cpl value of 1.08 verifies consistent control above the lower limit, while the CPU value of 35.19 indicates that the process rarely surpasses the higher limit. Despite barely missing the capability level, SOF (software) is still a very reliable and effective process with good overall performance.

```
Product Category: SOF
Mean Delivery Hours: 0.96 hours
Standard Deviation: 0.29 hours
Cp: 18.14
Cpu: 35.19
Cpl: 1.08
Cpk: 1.08

Product Category SOF is capable of satisfying the VOC.
```

Figure 37: Process Capability indices for Software

The Cp value of 0.90 for cloud services (CLO) indicates that the process spread is more than the limitations of the standard, indicating uneven delivery performance. The process mean is not well-centred, as shown by the Cpk value of 0.72, which could lead to deliveries that are over the maximum limit. This is supported by the CPU value of 0.72, which indicates a greater likelihood of delays that go above the allowable range, and the Cpl value of 1.08, which indicates comparatively steady management below the lower limit. The VOC cannot be satisfied by CLO (Cloud Services), and process modifications are needed to lower variance and enhance centring.

```
Product Category: CLO
Mean Delivery Hours: 19.23 hours
Standard Deviation: 5.94 hours
Cp: 0.9
Cpu: 0.72
Cpl: 1.08
Cpk: 0.72

Product Category CLO is NOT capable of satisfying the VOC.
```

Figure 38: Process Capability indices for Cloud Subscription

Deliveries for laptops (LAP) vary more than intended, as indicated by the Cp value of 0.90, which indicates that the process has a greater spread than the predetermined boundaries. Delivery delays are a result of the process mean being off-centre, as shown by the Cpk value of 0.70. While the Cpl value of 1.10 indicates good control below the lower limit, the CPU value of 0.70 indicates a recurrent tendency to exceed the upper limit. As a result, LAP (laptops) cannot meet the VOC and would benefit from process improvement to produce more reliable outcomes.

```
Product Category: LAP
Mean Delivery Hours: 19.61 hours
Standard Deviation: 5.93 hours
Cp: 0.9
Cpu: 0.7
Cpl: 1.1
Cpk: 0.7

Product Category LAP is NOT capable of satisfying the VOC.
```

Figure 39: Process Capability indices for Laptops

Mouse (MOU) has a considerable spread in relation to the limits, as indicated by its Cp value of 0.92, but it is not effectively centred, as indicated by its Cpk value of 0.73. While the Cpl value of 1.10 indicates little variation below the lower limit, the CPU value of 0.73 indicates that the process occasionally surpasses the higher specification limit. Since mouse (MOU) cannot fulfil the VOC, minor centring adjustments and variance reduction could assist the process operate within the intended range.

```
Product Category: MOU
Mean Delivery Hours: 19.3 hours
Standard Deviation: 5.83 hours
Cp: 0.92
Cpu: 0.73
Cpl: 1.1
Cpk: 0.73

Product Category MOU is NOT capable of satisfying the VOC.
```

Figure 40: Process Capability indices for Mouse

When evaluating whether a delivery process can satisfy the demands, expectations, and performance standards of a customer for a service or product, process capability indices are crucial. Assuming that a process is exactly centred between the Lower and Upper Specification Limits, its potential capability is represented by the Cp value. It shows how close or how far the process spread is from these boundaries. However, by taking into account how well the mean is centred inside the specification boundaries, the Cpk value assesses the process's true capabilities. The procedure may be able to satisfy customer needs if the Cpk value is 1.30 or greater. Performance in relation to the upper and lower specification bounds is shown by Cpu and Cpl, which are essential components of Cpk. poor performance. Crucial elements of Cpk, Cpu and Cpl, demonstrate performance with respect to the upper and lower specification bounds. subpar work.

3.4 Control issues

3.4.1 Rule A

MOU had a single s-chart point beyond the +3σ UCL at sample 592 (Total = 1; first/last = 592), which was the only breach found across all product categories. No additional products displayed exceedances of +3σ. This points to an uncommon but significant dispersion inflation in MOU during that period, which justifies a brief special-cause investigation surrounding sample 592.

Product <chr>	**Total** <int>	**First3** <chr>	**Last3** <chr>
MOU	1	592	592

Figure 41: Rule A

3.4.2 Rule B

Consecutive samples of s between the -1σ and +1 σ sigma-control limits for all product types (ordered from highest to lowest). This signifies good control.

- CLO: 35 consecutive samples in-control (samples 474–508).
- MON: 34 samples (238–271).

- SOF: 21 samples (659–679).
- LAP: 19 samples (116–134).
- MOU: 16 samples (672–687).
- KEY: 15 samples (730–744).

These runs indicate the CLO and MON processes maintained the most consistent short-term variation, while KEY and MOU had the shortest stable stretches and may benefit from variance-reduction efforts.

Product <chr>	LongestRun <int>	Start <dbl>	End <int>
CLO	35	474	508
MON	34	238	271
SOF	21	659	679
LAP	19	116	134
MOU	16	672	687
KEY	15	730	744

Figure 42: Rule B

3.4.3 Rule C

4 consecutive X-bar samples outside of the upper, second control limits for all product types.

- SOF: 27 windows (first three: [133–136], [202–205], [237–241]; last three: [774–801], [803–840], [842–864]).
- KEY: 27 windows (first three: [99–102], [112–117], [172–175]; last three: [687–696], [698–724], [726–746]).
- MOU: 25 windows (first three: [194–197], [233–240], [249–252]; last three: [768–775], [777–805], [807–860]).
- MON: 22 windows (first three: [134–137], [171–177], [179–186]; last three: [566–608], [610–613], [615–619]).
- CLO: 14 windows (first three: [122–125], [179–183], [192–200]; last three: [557–602], [604–626], [628–649]).
- LAP: 11 windows (first three: [119–122], [129–140], [153–167]; last three: [348–357], [359–372], [374–425]).

Regular +2σ runs indicate recurrent times when delivery implies a large drift (delay risk). When it comes to root-cause analysis, SOF, KEY, and MOU should be given priority because they exhibit the most persistent upward bias (e.g., demand spikes, staffing

Product <chr>	TotalWindows <int>	First3 <chr>	Last3 <chr>
CLO	14	[122–125], [179–183], [192–200]	[557–602], [604–626], [628–649]
KEY	27	[99–102], [112–117], [172–175]	[687–696], [698–724], [726–746]
LAP	11	[119–122], [129–140], [153–167]	[348–357], [359–372], [374–425]
MON	22	[134–137], [171–177], [179–186]	[566–608], [610–613], [615–619]
MOU	25	[194–197], [233–240], [249–252]	[768–775], [777–805], [807–860]
SOF	27	[133–136], [202–205], [237–241]	[774–801], [803–840], [842–864]

Figure 43: Rule C

shortfalls, supplier/route difficulties). Less mean-shift events are displayed by CLO and LAP, which is consistent with their longer periods of steady variability in Rule B.

4. Risk, Data correction and Optimising

4.1 Type I (Manufacturer's) Error

The table below represents the Type 1 (False-Alarm) probabilities for the Control-Chart decision rules.

Rule	Description	Formulas used	Type 1 Error	Interpretation
A	One sample outside the ±3σ limits	$a = 2 \times [1 - \Phi(3)]$	0.0027 ≈ 0.27%	3 out 1000 samples will trigger a false alarm
B	Run within ±1σ (good control)	$P = P(-1 < Z < 1) = \Phi(1) - \Phi(-1)$ $P = (0.6827)^n$	≈ 0	None, this rule identifies good control and stability
C	Four X-bars beyond the ±2σ same side	$a = [1 - \Phi(2)]^n$ Where n=4	0.000000268 ≈ 0.0000268%	Roughly 3 out of a million run will tigger a false alarm

These probabilities are purely theoretical; they do not depend on sections 3 data. The assumption is that the process is stable and centred at Ho.

4.2 Type II (Consumer's) Error for Bottling Filling Process

A type II error occurs when the chart fails to signal despite this shift. For this bottle-filling process, the specifications are an X-chart centred at 25.05L with limits UCL = 25.089L and LCL= 25.011L. However, the operator is unaware of that the process centre shifted to 25.028L and the standard deviation increases from 0.013 to 0.017. This is seen by the figure below.

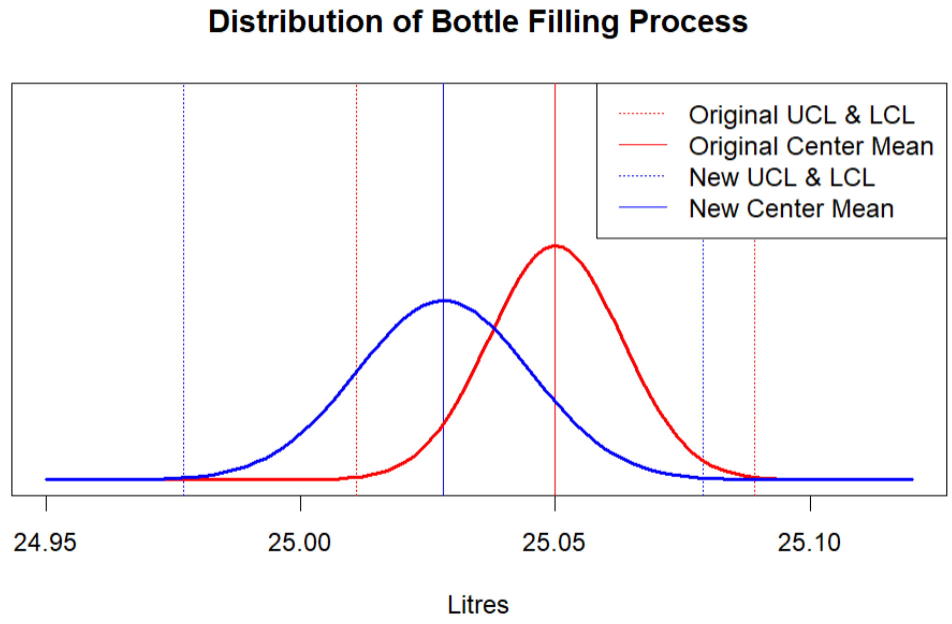


Figure 44: Distribution of original and new bottle filling process

The probability of type II error is calculated using the following formula:

$$\beta = \Phi\left(\frac{UCL - \mu_1}{\sigma^-}\right) - \Phi\left(\frac{LCL - \mu_1}{\sigma^-}\right)$$

Therefore, the probability is then $\beta = \Phi(3.588) - \Phi(-1.00)$ = 0.99983 – 0.15865 = 0.8412. The probability that the chart will detect the shift is then 1 – β = 0.1588 = 15.88%, which indicates that the chart is not very sensitive to small mean shifts.

4.3 Product and Head office Data Correction

Before the product analysis in Q1.2, the data was audited and the mismatch between category and ProductID was identified and corrected to ensure the data was clean prior to any analysis. This was done to ensure that the initial analysis was correctly done. The HeadOffice data was corrected by placing the repeated price and markup patterns to the ID numbers and by fixing the NA in the dataset.

Before the analysis was done, it was checked to determine how it would affect the previous analysis of the sale2022and2023 dataset, however, it was found that there was no overlap between ProductID found only in the HeadOffice dataset and sales dataset making a new analysis that would be identical to the original analysis.

```
ProductID found in Product data: CLO011, CLO012, CLO013, CLO014, CLO015, CLO016, CLO017, CLO018,
CLO019, CLO020, KEY041, KEY042, KEY043, KEY044, KEY045, KEY046, KEY047, KEY048, KEY049, KEY050,
LAP021, LAP022, LAP023, LAP024, LAP025, LAP026, LAP027, LAP028, LAP029, LAP030, MON031, MON032,
MON033, MON034, MON035, MON036, MON037, MON038, MON039, MON040, MOU051, MOU052, MOU053, MOU054,
MOU055, MOU056, MOU057, MOU058, MOU059, MOU060, SOF001, SOF002, SOF003, SOF004, SOF005, SOF006,
SOF007, SOF008, SOF009, SOF010
```

Figure 45: Summary of Product ID found in Product data

```
ProductID only found in Product headoffice data: CLO001, CLO002, CLO003, CLO004, CLO005, CLO006,
CLO007, CLO008, CLO009, CLO010, CLO021, CLO022, CLO023, CLO024, CLO025, CLO026, CLO027, CLO028,
CLO029, CLO030, CLO031, CLO032, CLO033, CLO034, CLO035, CLO036, CLO037, CLO038, CLO039, CLO040,
CLO041, CLO042, CLO043, CLO044, CLO045, CLO046, CLO047, CLO048, CLO049, CLO050, CLO051, CLO052,
CLO053, CLO054, CLO055, CLO056, CLO057, CLO058, CLO059, CLO060, KEY001, KEY002, KEY003, KEY004,
KEY005, KEY006, KEY007, KEY008, KEY009, KEY010, KEY011, KEY012, KEY013, KEY014, KEY015, KEY016,
KEY017, KEY018, KEY019, KEY020, KEY021, KEY022, KEY023, KEY024, KEY025, KEY026, KEY027, KEY028,
KEY029, KEY030, KEY031, KEY032, KEY033, KEY034, KEY035, KEY036, KEY037, KEY038, KEY039, KEY040,
KEY051, KEY052, KEY053, KEY054, KEY055, KEY056, KEY057, KEY058, KEY059, KEY060, LAP001, LAP002,
LAP003, LAP004, LAP005, LAP006, LAP007, LAP008, LAP009, LAP010, LAP011, LAP012, LAP013, LAP014,
LAP015, LAP016, LAP017, LAP018, LAP019, LAP020, LAP031, LAP032, LAP033, LAP034, LAP035, LAP036,
LAP037, LAP038, LAP039, LAP040, LAP041, LAP042, LAP043, LAP044, LAP045, LAP046, LAP047, LAP048,
LAP049, LAP050, LAP051, LAP052, LAP053, LAP054, LAP055, LAP056, LAP057, LAP058, LAP059, LAP060,
MON001, MON002, MON003, MON004, MON005, MON006, MON007, MON008, MON009, MON010, MON011, MON012,
MON013, MON014, MON015, MON016, MON017, MON018, MON019, MON020, MON021, MON022, MON023, MON024,
MON025, MON026, MON027, MON028, MON029, MON030, MON041, MON042, MON043, MON044, MON045, MON046,
MON047, MON048, MON049, MON050, MON051, MON052, MON053, MON054, MON055, MON056, MON057, MON058,
MON059, MON060, MOU001, MOU002, MOU003, MOU004, MOU005, MOU006, MOU007, MOU008, MOU009, MOU010,
MOU011, MOU012, MOU013, MOU014, MOU015, MOU016, MOU017, MOU018, MOU019, MOU020, MOU021, MOU022,
MOU023, MOU024, MOU025, MOU026, MOU027, MOU028, MOU029, MOU030, MOU031, MOU032, MOU033, MOU034,
MOU035, MOU036, MOU037, MOU038, MOU039, MOU040, MOU041, MOU042, MOU043, MOU044, ... <truncated>

Sales per ProductID only found in Headoffice data:
none
```

Figure 46: Summary of ProductID found in HeadOffice dataset

5. Optimising the Company's Profit

5.1 Shop 1 Analysis

The analysis evaluates the number of baristas needed to maximise profit while maintaining reliable service. For each number of baristas, the mean service time was calculated and then converted to a throughput (customers/hour = 3600 / mean service time). This was then further scaled into an 8-hour day, this allowed daily profit to be calculated. The daily profit was calculated by taking R30 per customer minus R1000 per baristas. The net profit increase linearly with staff size which indicates that revenue outweighs the cost of labour. With the tested range of six baristas, it is seen that 6 baristas yield the highest daily profit of R19902.66 (see red marker in figure 47 and 48). This offers the best balance between service capacity and operating cost.

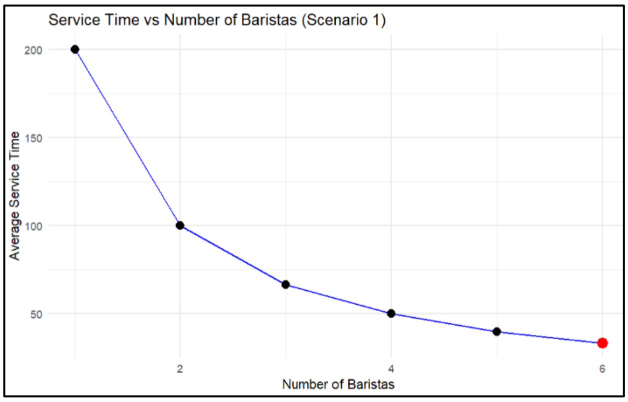


Figure 48: Line graph of service time vs number of Baristas

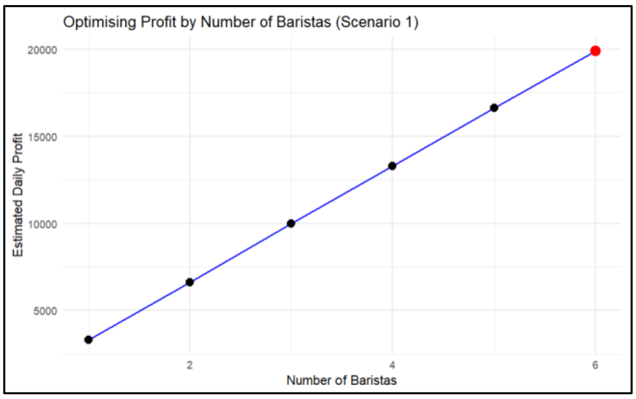


Figure 47: Line Graph optimising profit

5.2 Shop 2 Analysis

The same analysis was applied to shop 2, however, this output different results. For shop 2, the average service time declines as number of baristas increase. Unlike scenario 1, the profit curve peaks at 5 baristas (see the red marker in figure in 49 and 50) and falls slightly at 6 baristas. This indicates that the added labour cost after 5 baristas begins to outweigh the profit. Under the stated assumptions (throughput from mean service time, 8-hour day, R30 per customer, R1,000 per barista), five baristas provide the best balance of capacity and cost.

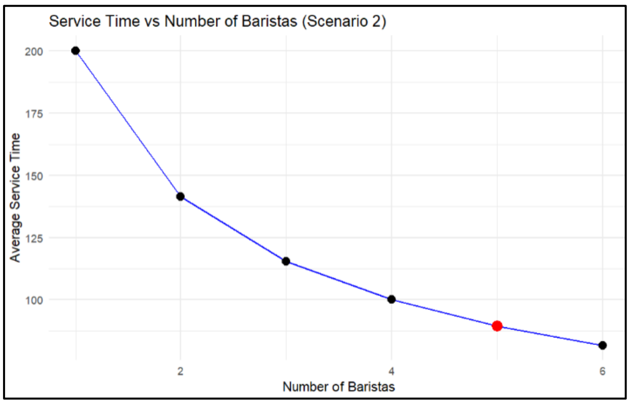


Figure 50: Shop 2 line graph of service time vs number of baristas

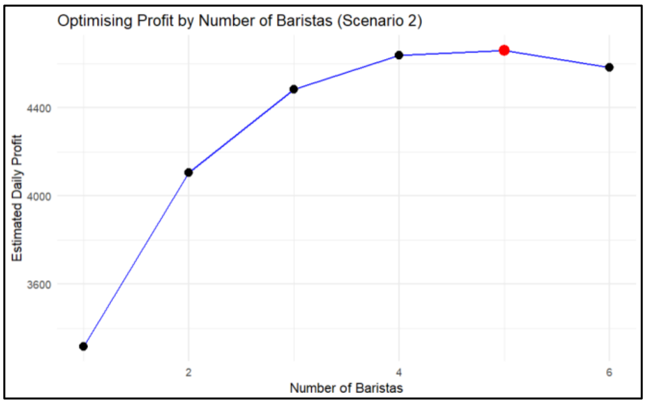


Figure 49: Shop 2 line graph optimising profit

6. DOE and MANOVA or ANOVA

6.1 Approach

The analysis employs in-control Phase-2 data to focus on the SOF product type this was guided by the stability tests in Part 3. DeliveryHours (main) and pickingHours (supporting) were the two responses taken into consideration. Months 1–12 and Year 2022-2023 were fixed factors. At $\alpha = 0.05$, analyses were conducted.

The hypothesises test:

For ANOVA on delivery Hours:

- Ho: No difference by year, no difference by month and no Year vs Month interaction
- H1: At least one factor effect (or interaction) is non-zero

For MANOVA on delivery Hours and Picking Hours

- Ho: The vector of means is equal across Year/Month/their interaction
- H1: at least one multivariate effect differs

6.2 Results and Interpretation

Two-way ANOVA (delivery hours)

- Year: $F \approx 0.00$, $p = 0.996$. This indicates no year-over-year difference.

- Month: $F \approx 21.39$, $p < 2x10^{-16}$. This is a clear seasonal effect.
- Year x Month: $F \approx 0.31$, $p = 0.979$. This is indicating that there is a stable monthly pattern across the years

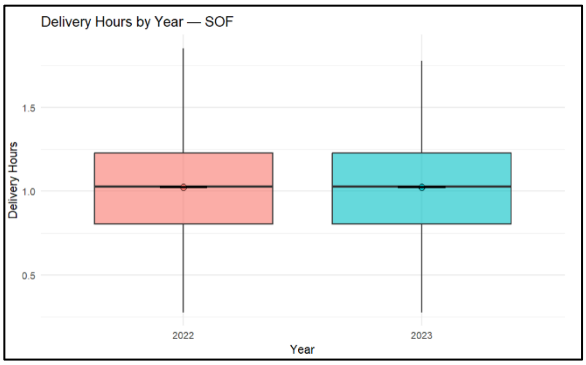


Figure 52: Box Plot of delivery hours per year

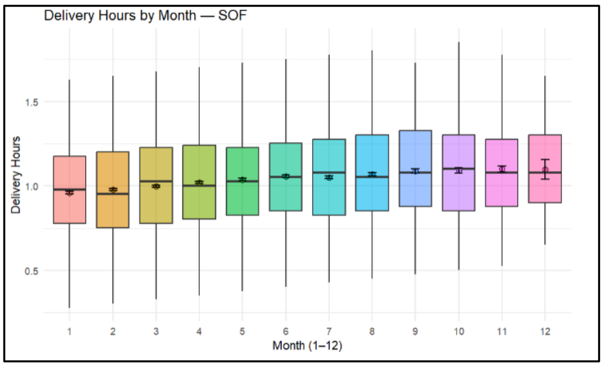


Figure 51: Box Plot of delivery hours per month

The boxplots show that there is a yearly median are virtually identical, while there are gentle increases between monthly medians.

MANOVA (delivery hours and picking hours)

- Year: Wilks $\lambda \approx 1.00$, $p = 0.506$. This indicates there is no multivariate year effect.
- Month: Wilks $\lambda \approx 0.894$, $p < 2x10^{-16}$. This indicates that there is a significant multivariate seasonality.
- Year x Month: Wilks $\lambda \approx 0.999$, $p = 0.794$. This indicates that there is no interaction between the to variates.

```
=== Two-way ANOVA: deliveryHours ~ Year * Month (SOF) ===
Analysis of Variance Table

Response: deliveryHours
             Df  Sum Sq Mean Sq F value Pr(>F)
Year          1    0.00 0.00000  0.0000 0.9964
Month        11   20.52 1.86553 21.3835 <2e-16 ***
Year:Month   10    0.27 0.02706  0.3102 0.9789
Residuals 12745 1111.90 0.08724
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 54: Two-way ANOVA

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

=== MANOVA (deliveryHours & pickingHours) — Wilks' Lambda ===
              Df   Wilks approx F num Df den Df Pr(>F)
Year           1 0.99989    0.681      2  12744  0.506
Month         11 0.89411   66.681     22  25488 <2e-16 ***
Year:Month    10 0.99885    0.734     20  25488  0.794
Residuals  12745
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 53: MANOVA results

7. Reliability of Service

7.1 Determining the Amount of Reliable Service

To determine the optimal number of reliable working days the following binomial was setup, $X \sim \text{Binomial}(n=16, p)$, where p is the attendance probability. Fitting to the 397-day histogram yields a $p=0.9740$ and service is defined as reliable if the is ≥ 15 staff working a day

```
--- Binomial fit ---
Assigned per day (n) = 16
Estimated attendance p̂ = 0.974024

--- Q1: Reliability with n = 16 ---
P(reliable) = P(X ≥ 15) = 0.936
Expected reliable days/year = 341.8
Expected problem days/year  = 23.2
```

Figure 55: Reliabilty calculations

Results: P (reliable | n= 16, p) = P (X≥15) = 0.936. This means that over a 365-day year there are 341.8 reliable days and 23.2 problem days

7.1 Optimising profit for the Company

To balance the cost and optimise profit, the attendance was modelled as X ~ Binomial (N, p) with a p = 0.9740 and reliability defined as ≥15 staff present. The expected annual total cost was calculated by R300000xN plus the expected sales loss for "problem days" by R200000 x 365 x P (X<15).

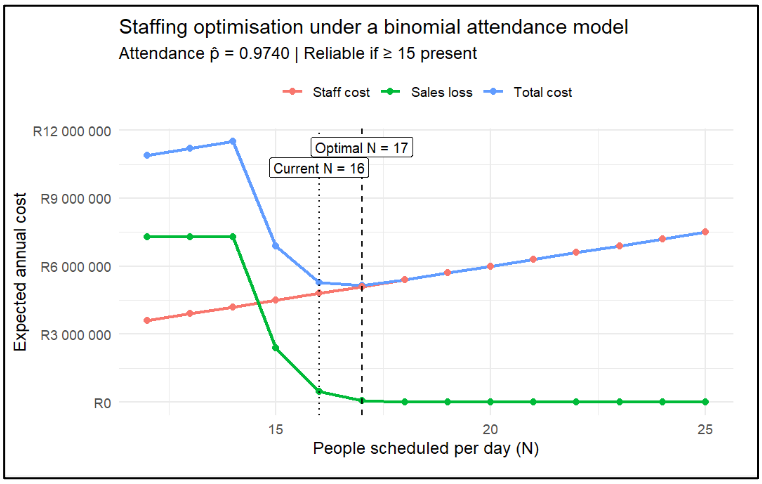


Figure 56: Staffing Optimisation

By plotting the graphs of the staff cost, sales lost and total cost, one can see that the optimal number of staff members is 17 staff members (see figure 56). This optimal number of staff members gives us the lowest cost which will in theory provide the highest profit.

Conclusion

A thorough assessment of process performance and quality across several datasets was offered by the ECSA GA4 analysis. The descriptive statistics showed that Laptops were clearly the most profitable due to high selling price, and it was also clear that clients were evenly distributed across demographic and geographic segments. There was only one product capable of meeting the VOC (Software), where on the other hand the rest were not. The results unequivocally show the importance of routine process monitoring and preventative actions. Through the baristas problem, it was shown how one should use data-driven decision-making to optimize their process to see increases service operations' profitability. Overall, the study showed how to handle challenging quality assurance issues by applying statistical reasoning, computational effectiveness, and engineering judgement in order to improve internal operation within multiple different operations such as sales, profit, and lastly deliveries.

References:

Banerjee, A., Chitnis, U., Jadhav, S., Bhawalkar, J. & Chaudhury, S. 2009. Hypothesis testing, Type I and Type II Errors. *Industrial Psychiatry Journal*. 18(2):127. DOI: https://doi.org/10.4103/0972-6748.62274.

Buderer, N.M. & Brannan, G.D. 2024. *Comparing the Means of Independent Groups: ANOVA, ANCOVA, MANOVA, and MANCOVA*. StatPearls Publishing. Available: https://www.ncbi.nlm.nih.gov/books/NBK606084/.

ChatGPT. 2015. Available: https://chatgpt.com/g/g-p-68da8030252481918c849f2a8923ff4e-ecsa-report/c/68f4de0b-cb20-8327-94f4-909d2c2a2c0d [2025, October 23].

Clinical Excellence Commission. 2024. *Pareto Charts & 80-20 Rule - Clinical Excellence Commission*. Available: https://www.cec.health.nsw.gov.au/CEC-Academy/quality-improvement-tools/pareto-charts.

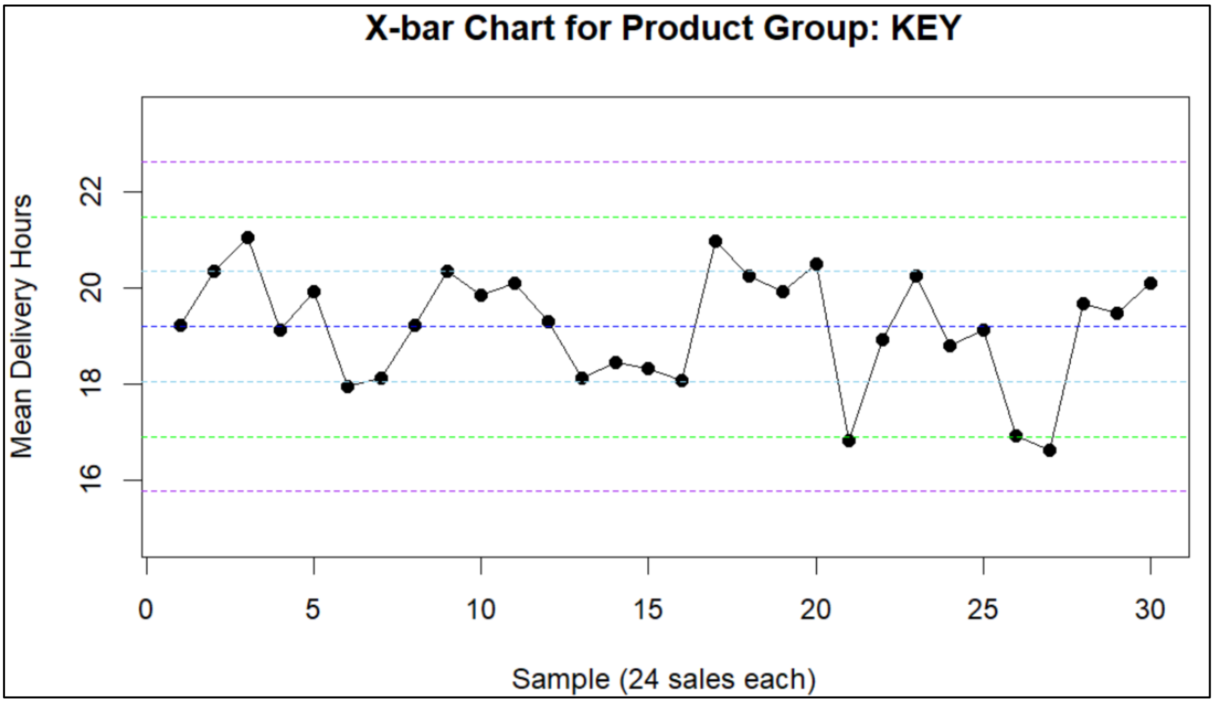
Hessing, T. 2019. *X Bar S Control Chart | Six Sigma Study Guide*. Available: https://sixsigmastudyguide.com/x-bar-s-chart/.

MANOVA Test in R: Multivariate Analysis of Variance - Easy Guides - Wiki - STHDA. 2025. Available: https://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance.

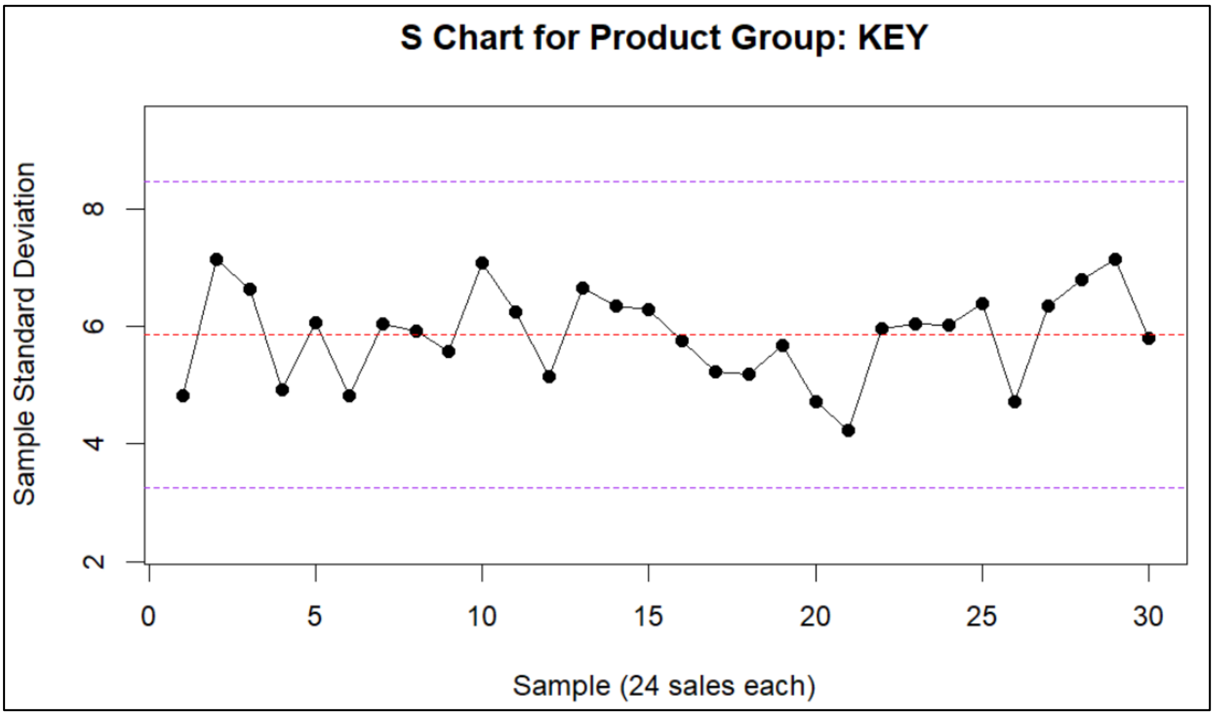
Steynberg, S. 2022. *Subject Guides: Research Guide: Data analysis and findings*. Available: https://library.up.ac.za/c.php?g=485435&p=4425510.

Appendix A:

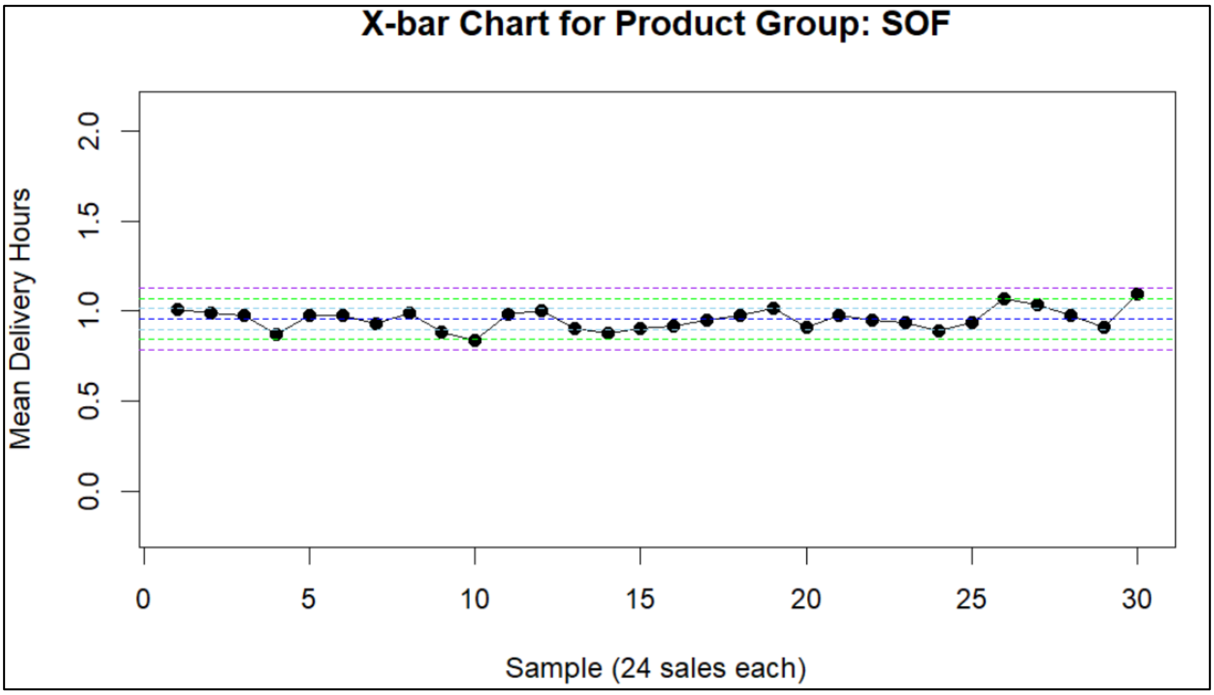
Graph A1:



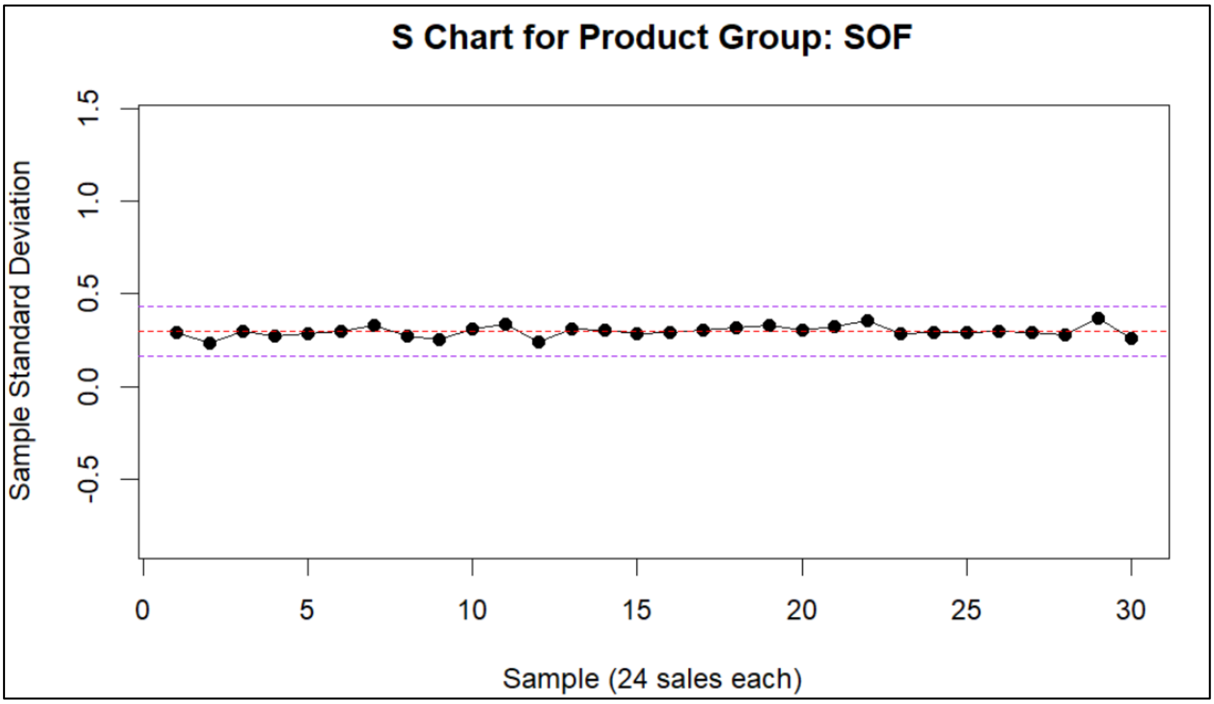
Graph A2



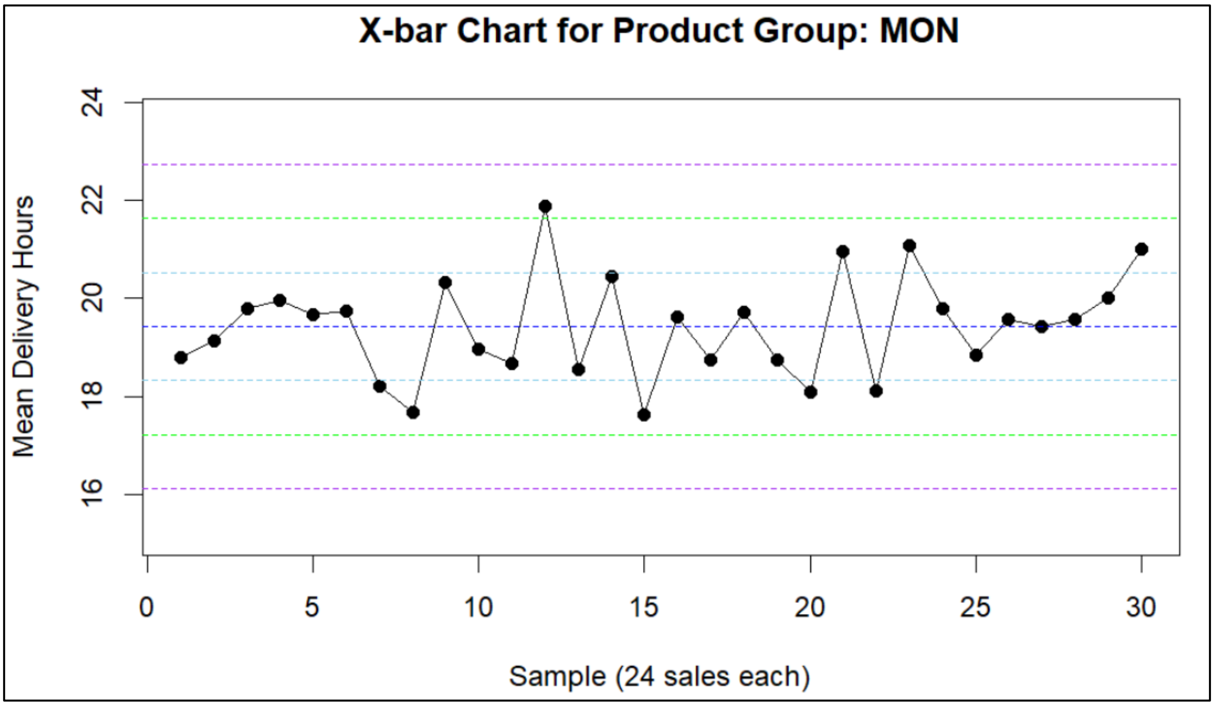
Graph A3



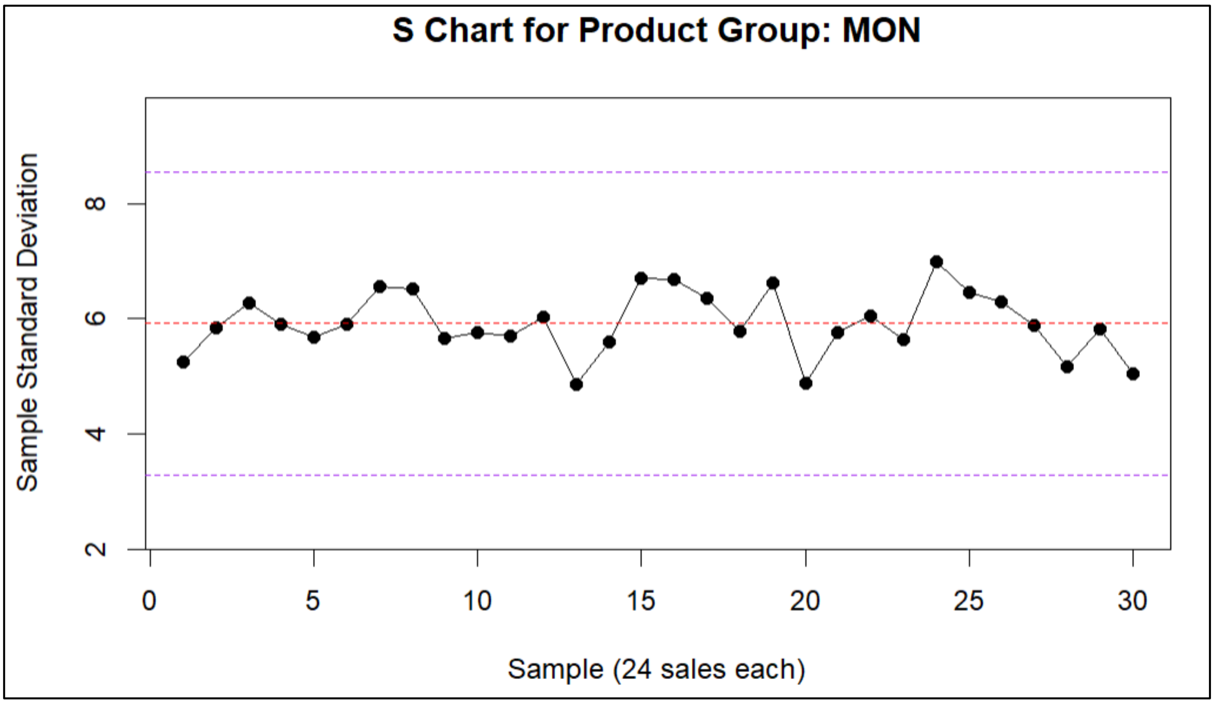
Graph A4



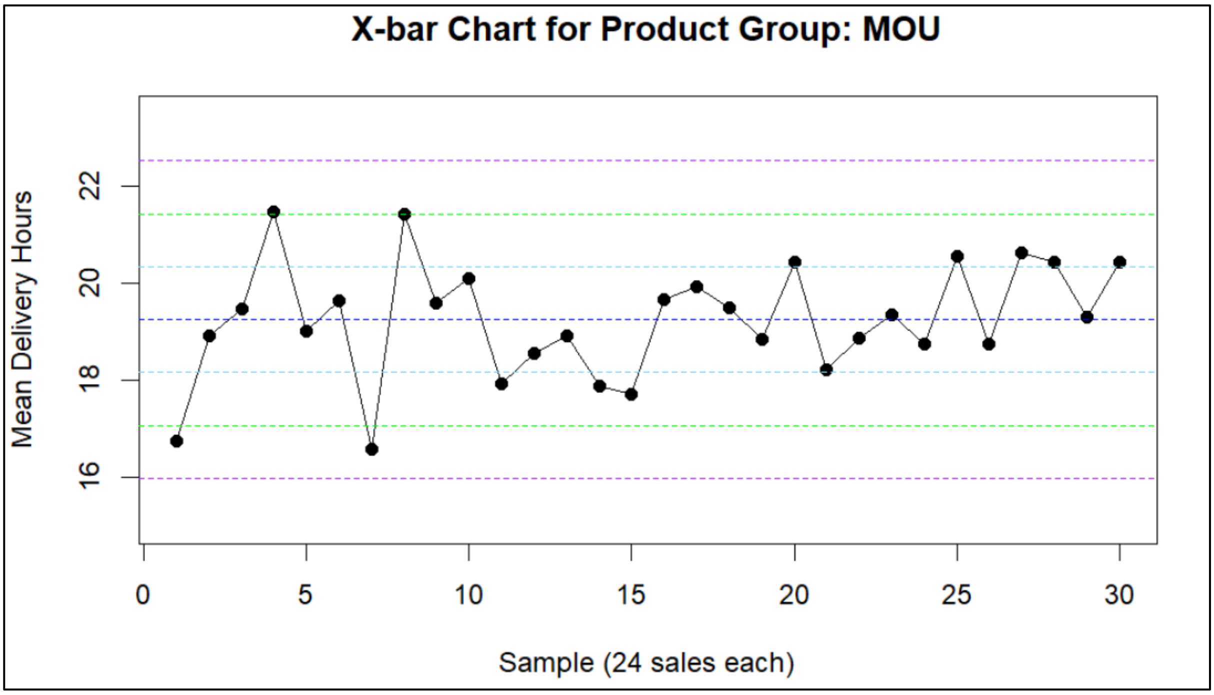
Graph A5:



Graph A6:



Graph A7:



Graph A8: