Matthew van der Merwe 26080958

QA 344 ECSA Project

Table of Contents

Intermediate Hand-in - Basic Data Analysis

1. Data Loading & Inspection

The *Sales* dataset comprises of 100000 rows and 9 columns, providing information on customer and product ID's, respective product quantities, times of orders as well as picking and delivery hours. The *Customer's* dataset is made up of 5000 rows and 5 columns specifying the customer ID, gender, age and monthly income, as well as the city they reside in. Functions like *colnames()* and *dim()* are used to determine each column name as well as the number of rows and columns in each table.

2. Summary statistics

Used to display statistical information such as central tendency and numerical distribution, functions such as *summary()* can be used to compute sales quantities such as the mean [13.5], median [6], minimum [1] and maximum [50]. Similarly, the same function can be used to determine the mean customer income [ZAR 80.8k] and the mean customer age [52 years]. Information on the timing of orders is also determined, with the average time being around 13h00. With regard to operation times, it can be seen that the mean picking time is just over 14 hours, and the mean delivery time is 17.5 hours.

3. Missing values

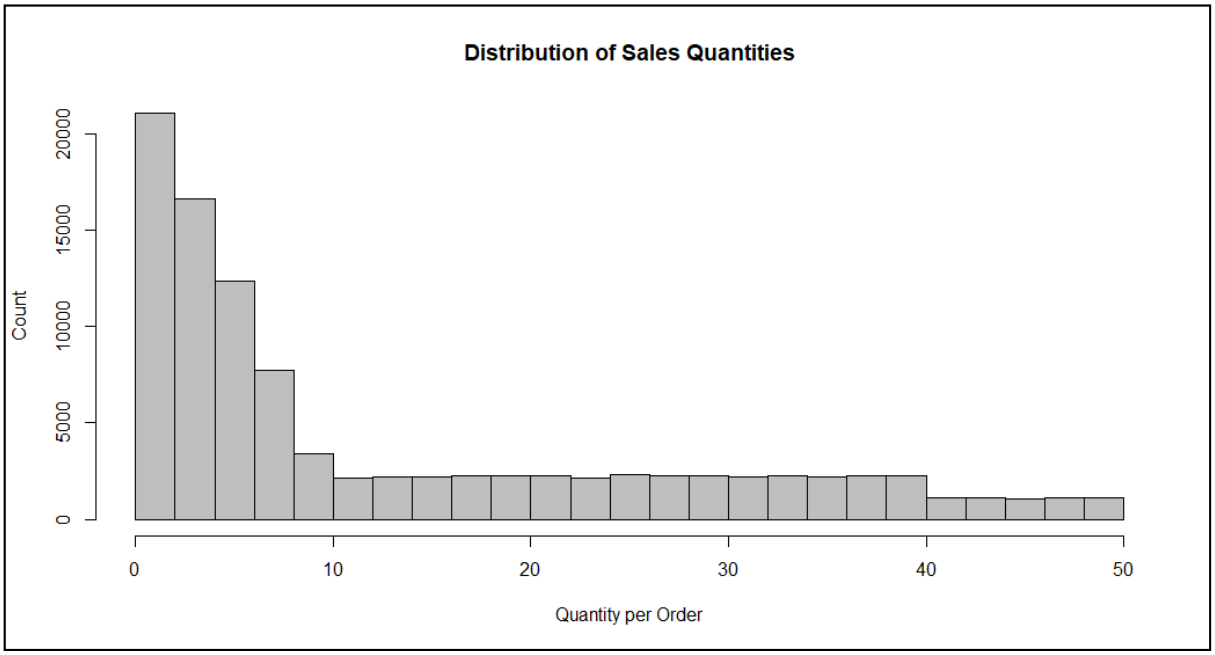
The data given in the csv files is clean and there are no missing values.

4. Data filtering & Subsetting

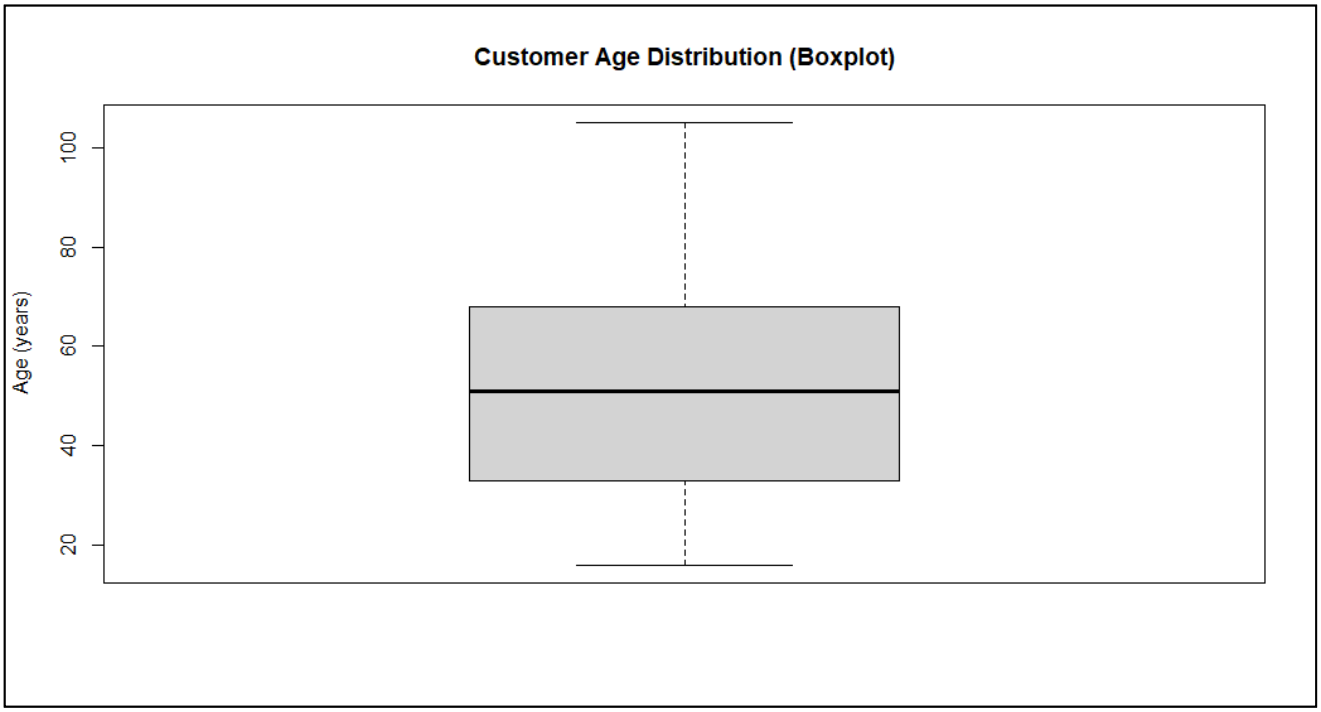
In the data we can see sales in years 2022 and 2023 are being compared, as well as with customers with a high average monthly income versus a lower average income. Customers are also grouped into age classes, being under 30 years, between 30 and 60 years, and older than 60 years.

5. Data Visualisation

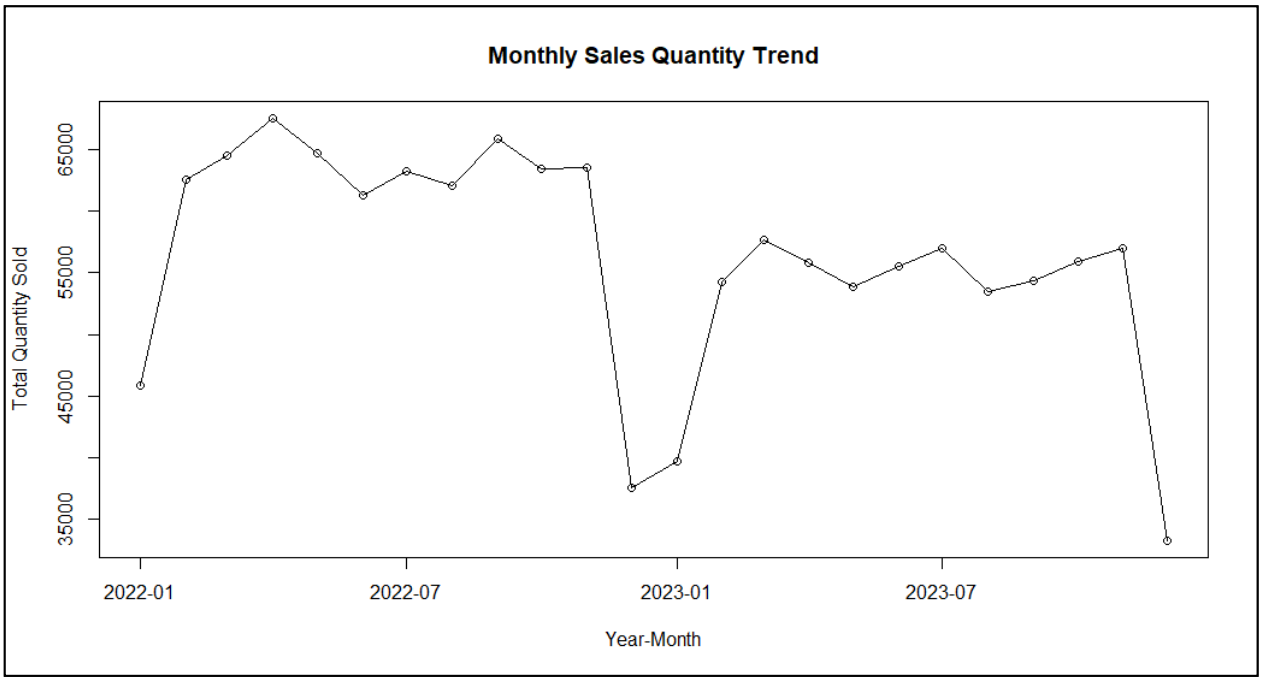
As seen in the histogram below, sales quantity is skewed to the right, with most quantities being smaller than 10. This tells us that the company experiences a higher volume of small orders. The mean order size is around 13.5 units, which is increased by the low number of large orders.



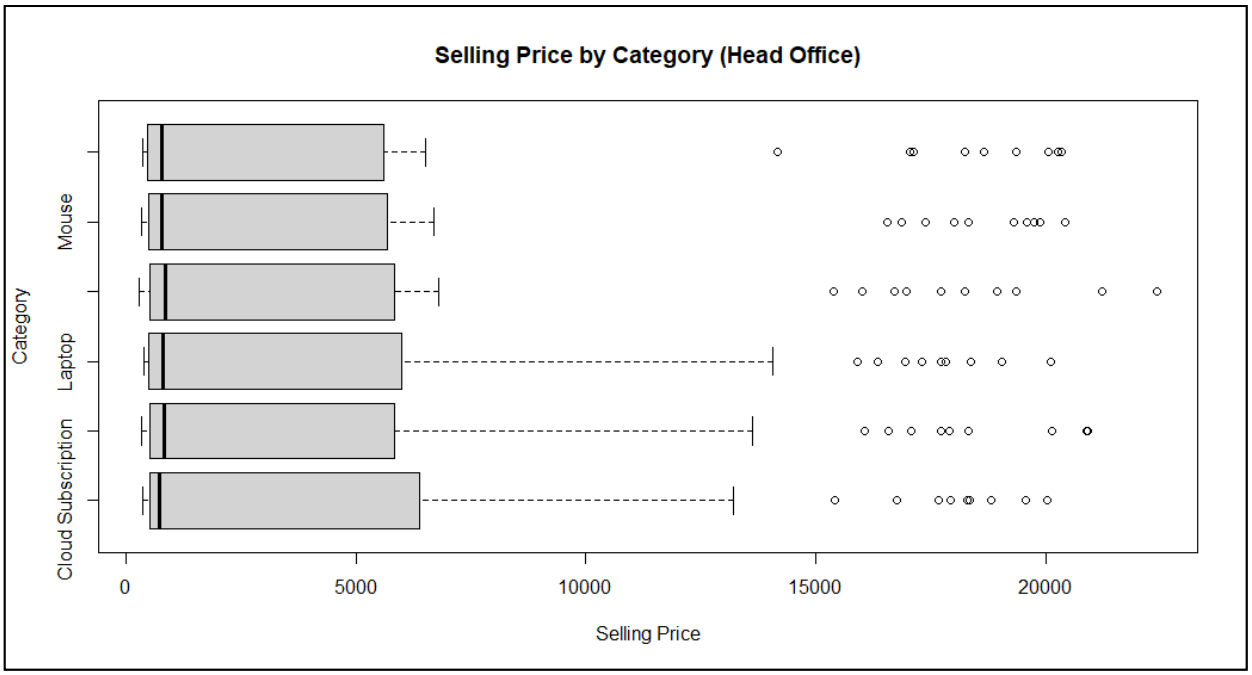
In the boxplot seen below, it can be observed that there is a wide range of customer ages, with the core customer portion being middle-aged. Outliers are seen to be older than 100 years, and some slightly younger than 20.



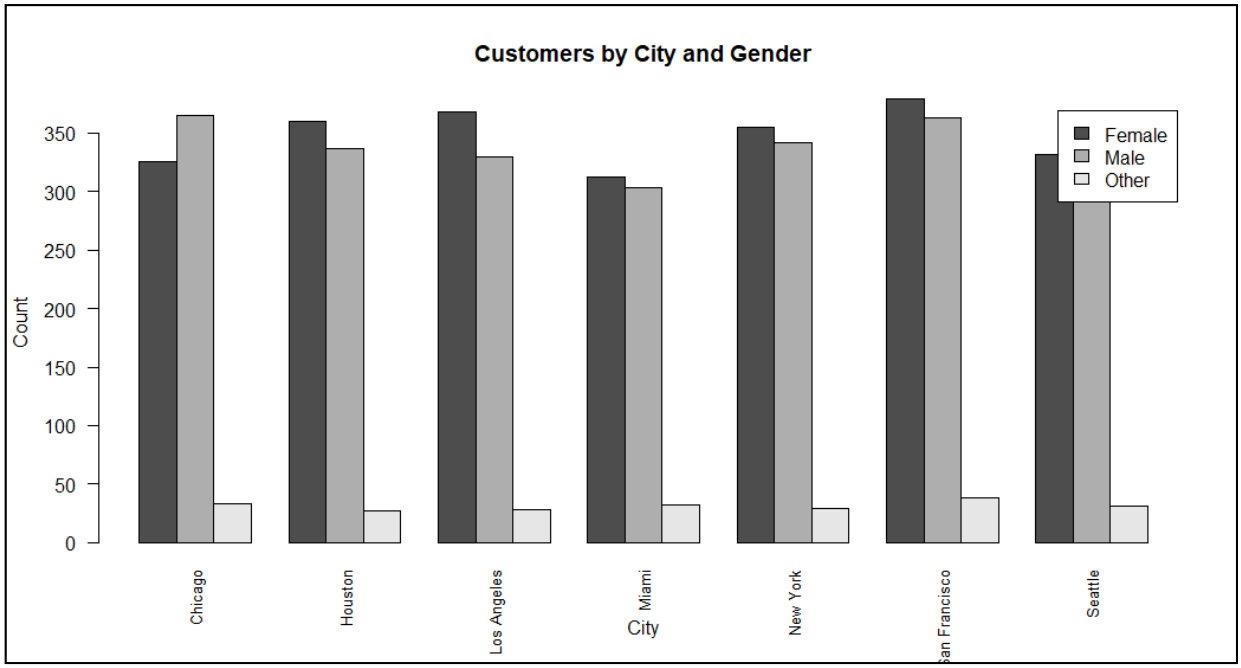
In the line graph displaying monthly sales, seasonality is shown. This indicates that the demand is heavily dependent on seasonal factors such as holidays or industry cycles. Below, it can be seen that the line plot is based on the total quantity sold over the years 2022 and 2023, and a similar trend is shown over both years.



In the second boxplot seen below, where selling price is graphed by category, it is evident that there is a larger price gap with higher-priced categories. In the lower-priced categories, the average price is quite constant, unlike in the higher-priced categories. This is a perfect example of a large price variability, meaning that multiple market segments are catered for.



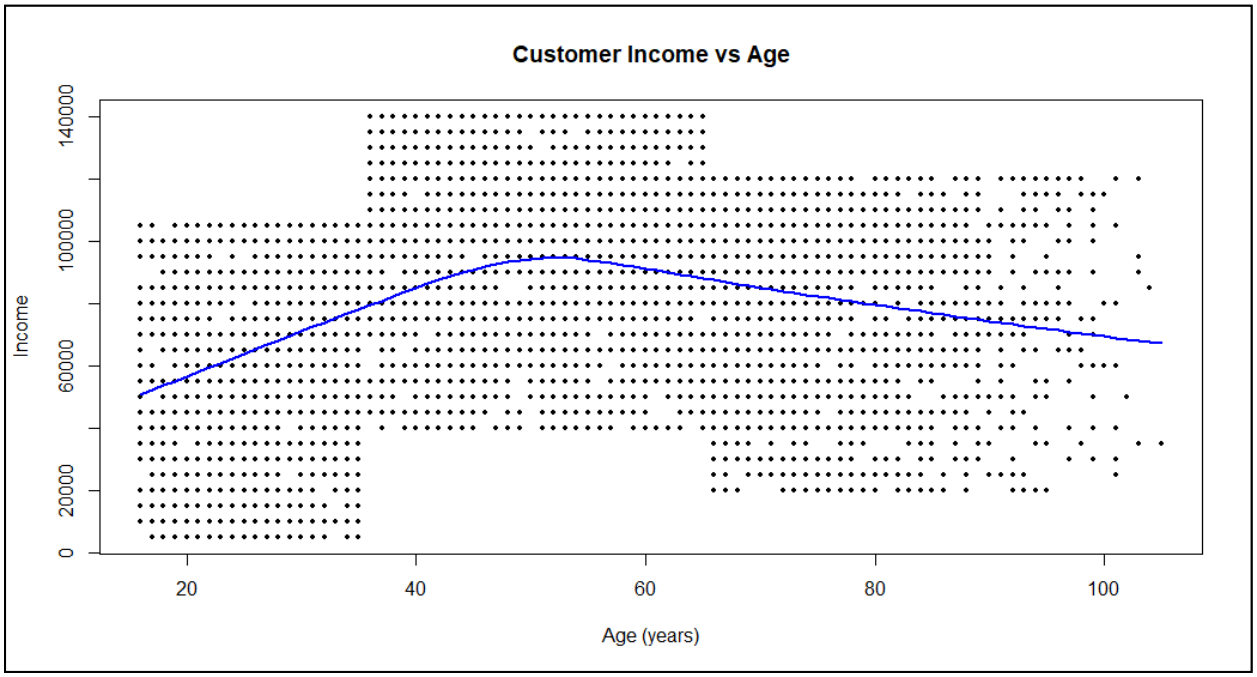
In the bar graph titled “Customers by City and Gender”, the customer base is seen to be distributed over 7 cities. San Francisco is seen to have the largest city with 780 customers. The roughly even split between male and female customers is also seen in each city, showing no customer base is dominated by a specific gender. This is important with regard to marketing strategies, as broad-based strategies would be effective in these cities as opposed to cities with an uneven gender split in customer base.

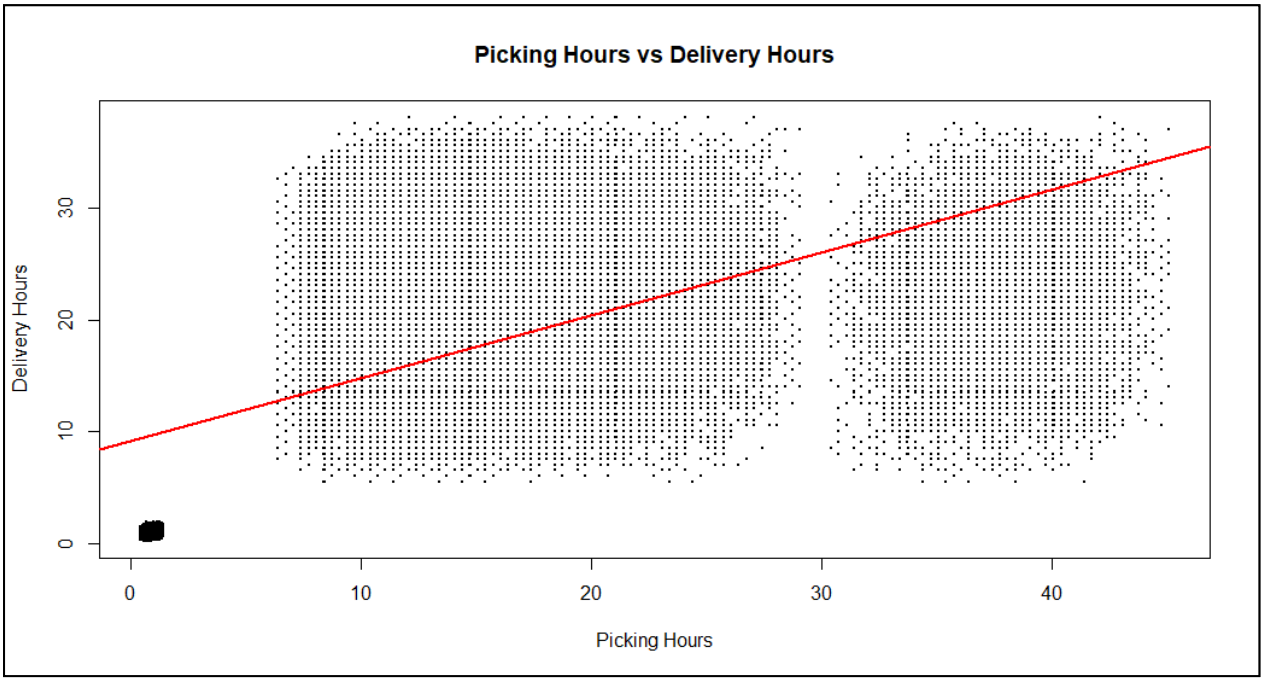


6. Relationships

When graphing customer income against their respective age in the below scatter plot, it can be seen that the older customers earn more on average compared to younger customers. This may mean that customers in an older age bracket are more able and likely to buy premium products, whereas the younger population is more inclined to purchase discounted or cheaper items.

In the second scatter plot seen below, a positive correlation is displayed, showing delivery time increases with picking time. The gradient of the line of regression shows the effect of a delay on delivery.





Part 3 – Statistical Process Control

3.1 Initialisation of $\bar{X}$ & s control charts

To ensure proper chronological sequence for SPC analysis, the delivery time data for each product type was first sorted by *Year*, *Month*, *Day* and *orderTime*. Subsequently, samples of 24 observations were created for each phase, signifying distinct monitoring subgroups. Th initial control limits and centre lines for both the $\bar{X}$ and s-charts were determined by using the first 30 samples (or 720 observations per product type). The upper and lower control limits were worked out to be:

$UCL_{\bar{x}} = CL + A_3 \times \bar{s}$ and $LCL_{\bar{x}} = CL - A_3 \times \bar{s}$ for the $\bar{X}$-chart,
The s-chart limits were established as $UCL_s = B_4 \times \bar{s}$ and $LCL_s = B_3 \times \bar{s}$,
using the constants, $A_3 = 0.619$, $B_3 = 0.555$, and $B_4 = 1.445$ (for n = 24).

In order to determine the degree of deviation from the process mean, the 1σ, 2σ and 3σ zones were also created. The first visual expression of the data such as the ECDF distribution and the *orderTime* boxplot showed a smooth cumulative distribution for all product categories as well as a median of around 13 hours. This suggests that the procedure was stable at first, and appropriate for creating trustworthy baselines for control charts.

3.2 Process Monitoring using new samples

In order to replicate continuous process monitoring, extra samples of 24 delivery times per product were taken successively from sample no. 31 onward after the setup phase. Before analyzing the $\bar{X}$-chart for each product, the s-chart was evaluated to ensure that variation remained within acceptable bounds. Histograms titled *deliveryHours* and *orderTime* contained a small number of extremely low values, and displayed close to normal distributions focused around 20 hours, proving the existence of typical process variation, but no indications of instability. Order frequency was uniformly distributed over time, which indicates that there was no systemic bias associated with the calendar period, according to seasonal plots based on *orderDay* and *orderMonth*. In general, the monitoring phase illustrated that, for all product categories, the process stayed statistically controlled, and continuous sampling did not disclose any notable changes in means or dispersion.

3.3 Process Capability Indices (Cp, Cpu, Cpl & Cpk)

Using the first 1000 delivery observations, process capability indices were calculated for each product category, to determine whether the process satisfied customer requirements. The estimated capability indices are seen below:

$$Cp = (USL - LSL) / (6\sigma) \approx 1.33$$
$$Cpl = (\mu - LSL) / (3\sigma) \approx 1.08, \text{ and}$$
$$Cpu = (USL - \mu) / (3\sigma) \approx 1.58$$

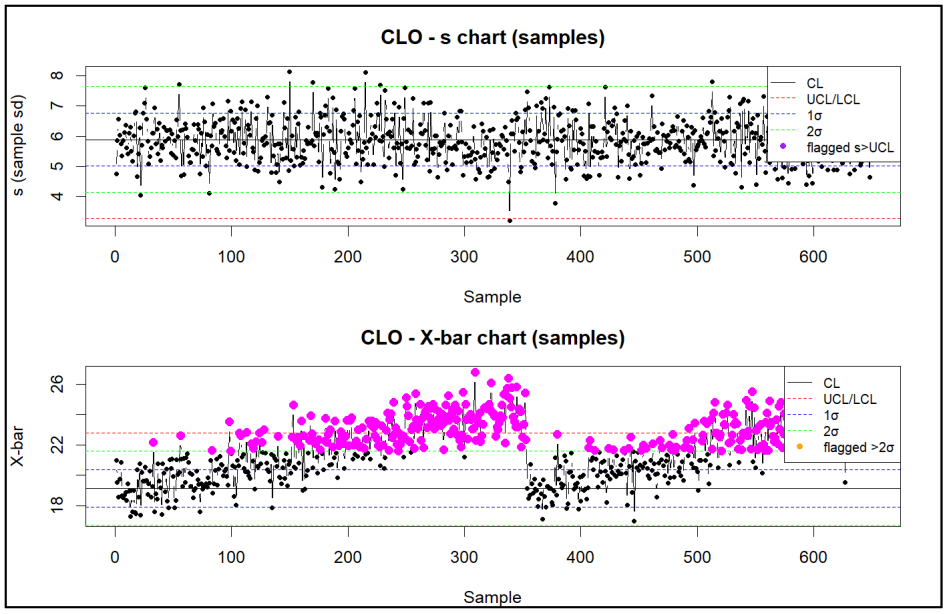
This worked out to yield a Cpk value of 1.08, keeping in mind that the Lower Specification Limit (LSL) and Upper Specification Limit (USL) were 0 and 32 hours respectively, with a standard deviation of around 4 hours. It can be concluded that the Cpk values are higher than the minimum benchmark of 1.0, meaning that the delivery processes for *all* product types satisfy the VOC standards. Despite a slightly off-centre tendency seen in some product types with shorter average delivery times, the overall process variability is comfortably within acceptable limits. Categories such as Software (SOF) and Monitor (MON) showed especially narrow and stable spreads, illustrating more reliable process control performance.

3.4 Identification of Process Control Issues

The usual SPC detection rules were used to take a look into certain process control concerns. In accordance with Rule A, samples with s-values higher than the top +3σ control limit were identified as possible outliers or 'out-of-control' points. Eight cases were flagged, which may indicate sporadic anomalies in the process of concern. The longest series of successive s-values in the +1σ was discovered by Rule B, uncovering 14 samples, namely samples 56-69 for product "KEY", which suggests a prolonged period of exceptional process stability. Rule C identified instances where four or more consecutive $\bar{X}$ samples went over the +2σ upper limit. Five occurrences met the abovementioned criteria; the oldest occurring between samples 38-41 for "LAP" and the latest occurring between samples 114-117 for "CLO". These trends show brief but detectable changes in process means. To conclude, even though the majority of samples remained within the specified control limits, Rules A and C showed that there are minor assignable causes. This deduces that further stability could potentially be achieved, and the current high process capability can be maintained with targeted investigations or small operational adjustments.

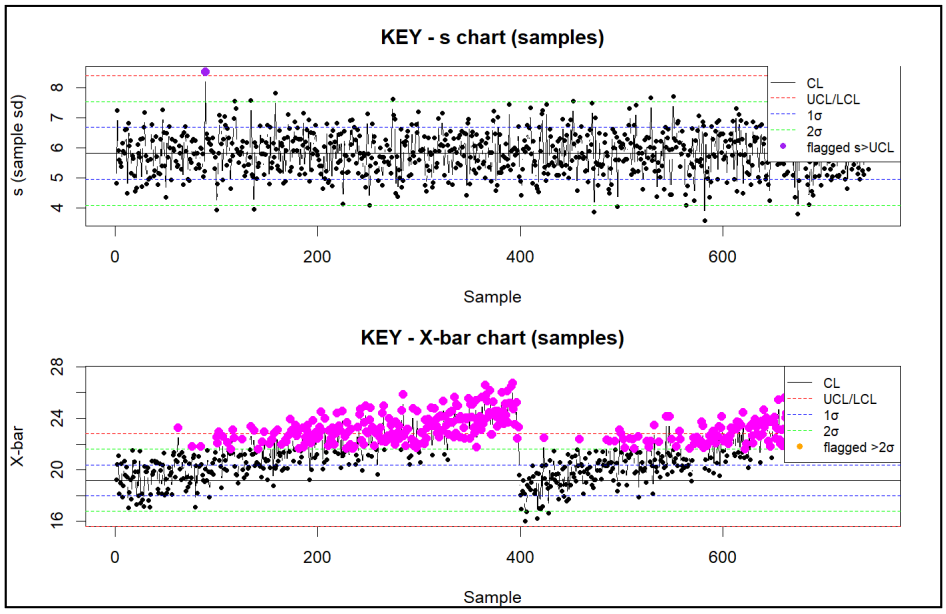
1. CLO

- s-chart: UCL = 8.2; LCL = 4.1
- $\bar{X}$-chart: UCL = 23.8; LCL = 18.0
- s-chart shows consistent variability within the limits
- $\bar{X}$-chart shows many points above the Upper Control Limit (indicated in pink), indicating the process mean is not stable over time



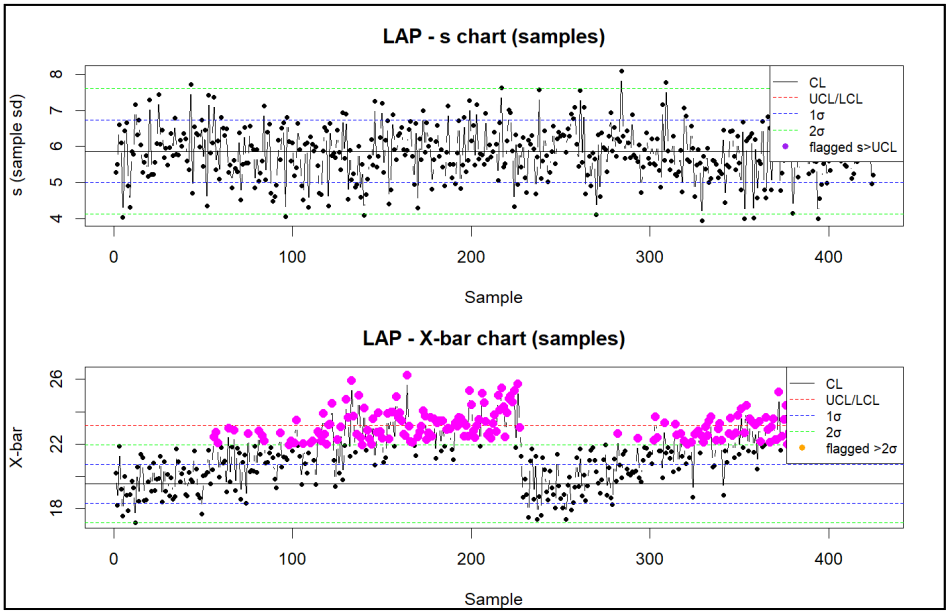
2. KEY

- s-chart: UCL = 8.2; LCL = 4.1
- $\bar{X}$-chart: UCL = 24.8; LCL = 17.5
- s-chart within control limits, showing consistent variation
- $\bar{X}$-chart shows many points breaching the UCL, suggesting an unstable process mean



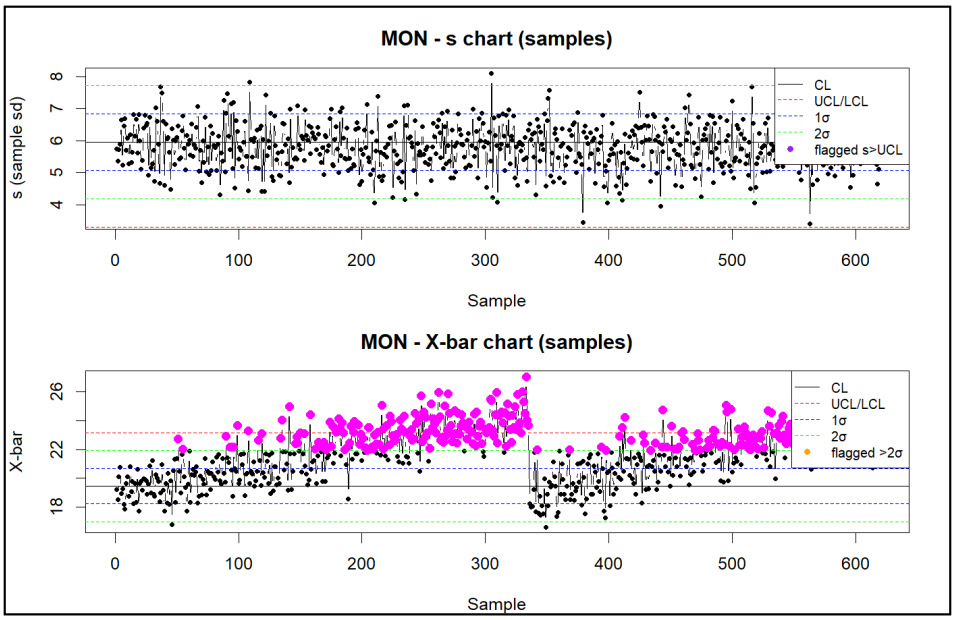
3. LAP

- s-chart: UCL = 7.5; LCL = 4.3
- $\bar{X}$-chart: UCL = 23.8; LCL = 17.8
- s-chart is stable
- $\bar{X}$-chart has frequent upper limit breaches and a clear upward trend



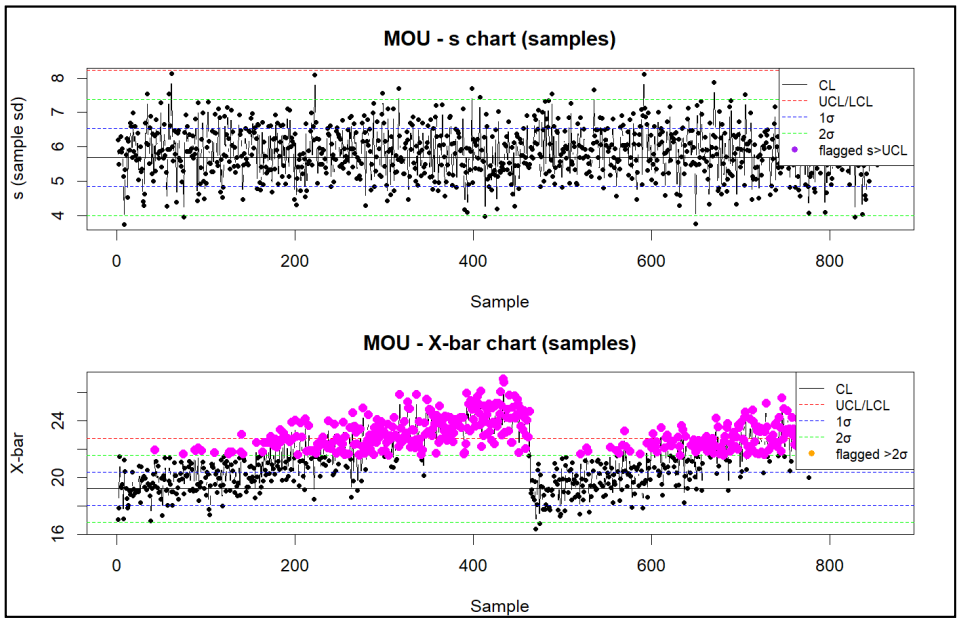
4. MON

- s-chart: UCL = 7.8; LCL = 4.0
- $\bar{X}$-chart: UCL = 24.5; LCL = 18.0
- s-chart is stable throughout, showing a controlled process variation
- $\bar{X}$-chart displays a significant mean upward shift with many points above the UCL, indicating non-random variation in mean output



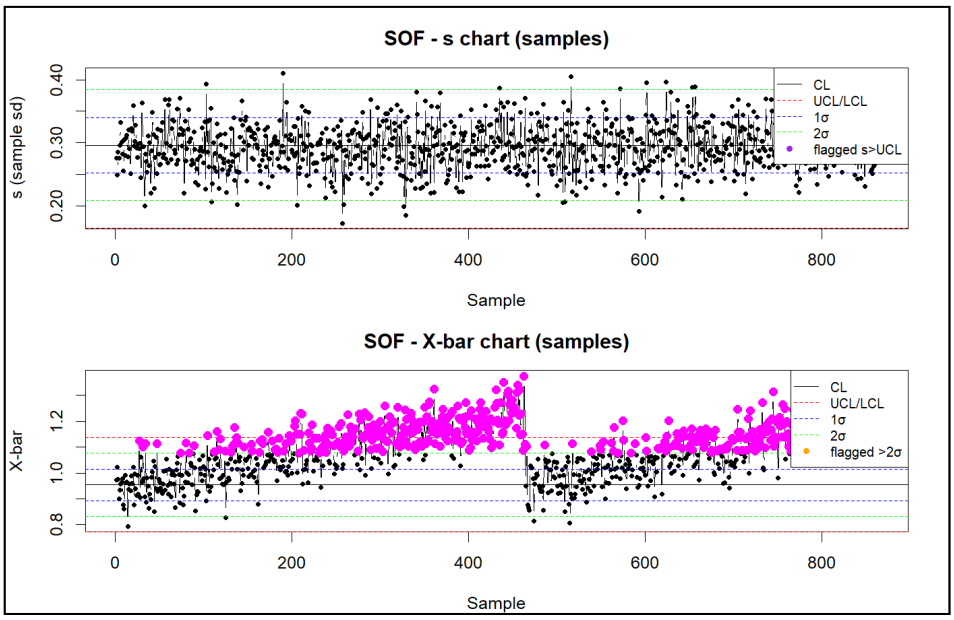
5. MOU

- s-chart: UCL = 8.1; LCL = 4.2
- $\bar{X}$-chart: UCL = 24.2; LCL = 17.0
- s-chart is within bounds, showing low variability
- $\bar{X}$-chart shows process mean above the Upper Control Limit, then a downward correction



6. SOF

- s-chart: UCL = 0.38; LCL = 0.20
- $\bar{X}$-chart: UCL = 1.15; LCL = 0.85
- s-chart shows stable variation with a consistent spread
- $\bar{X}$-chart has several points above the UCL, showing an instable process mean which suggests possible calibration or process drift



Part 4 - Risk, data correction & Optimising for maximum profit

4.1 Type I (Manufacturer's Error)

The likelihood of a false alarm, where a process is deemed out of control when it is actually functioning normally, is represented by Type I errors. This likelihood varies according to the sensitivity of the usual control chart rules. When a single point exceeds the +3σ control limits under Rule A, the chance of a false signal is determined as follows:

$$\alpha = 2P\ (Z > 3) = 0.0027,\ \text{expressed as } 0.27\% \text{ per point tested}$$

This indicates that there would be one false signal for every 370 depicted dots on average.

With p = P (Z > 2) = 0.0228 for Rule B, which requires two of the three consecutive points must fall beyond +2σ on the same side of the centreline, the Type I error probability is:

$$\alpha = 3p^2\ (1 - p) + p^3$$

This results in a false alarm rate of 0.00153, or 0.153%. The probability would double if both sides of the centreline were taken into account, however the overall conclusion would remain the same – this rule is more sensitive to subtle changes but generates slightly more false alarms than Rule A.

In accordance to Rule C, which states that four of five points must be on the same side of the centreline and fall beyond +1σ,

$$p = P\ (Z > 1) = 0.1587$$

The Type I error is roughly 0.00277 (0.277% per five-point window) using the same binomial calculation. Additionally, the chance would double if the rule applied to both sides. Finally, the probability that such a sequence would occur by chance is

$$P = 2\ (0.5\text{^}7) = 1.56\%$$

if the rule of seven points in a row on one side of the centreline is applied. Altogether, these findings demonstrate how adding more "run" rules augments sensitivity while simultaneously increasing the possibility of false alarms (α).

4.2 Type II (Consumer's Error)

A Type II error is the likelihood of whether the process is truly out of control, but the chart does not indicate a problem. With the $\bar{X}$-chart's centreline at 25.050 L, the upper control limit at 25.089 L, and the lower limit at 25.011 L, the process standard deviation has slightly increased from 0.013 L to 0.017 L, with the true process mean also shifting to 25.028 L. Calculating the likelihood that a sample mean falls between the two control limits given the new distribution yields the Type II probability. This is calculated below as

$$\beta = P\ (25.011 \leq \bar{X} \leq 25.089 \mid \mu = 25.028,\ \sigma\bar{X} = 0.017)$$

When these limits are converted to z-scores, the result is

$$zL = (25.011 - 25.028) / 0.017 = -1.0zL = (25.011 - 25.028) / 0.017 = -1.0,$$

and

$$zU = (25.089-25.028) / 0.017 = 3.588$$

Consequently,

$$\beta = \Phi\ (3.588) - \Phi\ (-1.0) = 0.841,$$

Meaning that, on any given sample, there is an 84.1% chance of missing this shift. The $\bar{X}$-chart has a one in six chance of detecting the change in process mean and variability, as the corresponding detection power $(1-\beta)$, is only about 15.9%.

The s-chart, which tracks variability, should also be triggered by the variance increase (σ increasing from 0.013 to 0.017). We cannot directly calculate its β without its limits, but if we assume that the charts are independent, this would mean that the combined miss probability would be approximately equal to the product of the two β values. Therefore, the $\bar{X}$-chart's figure of 0.841 can be regarded as a conservative estimate.

Part 5 – Profit Optimisation

Data representation

The CSV files named timeToServe and timeToServe2 contain data for two different shops whose profits need to be optimized. In each file, there is a column containing the number of baristas on shift (V1), and the time (in seconds) they serve (V2). Each file is considered to be a full year of demand, being 200 000 customers per year in each shop.

1. **Service capability per barista (μ)**

   μ = 3600 s

   μ (Shop 1) = 41.22s = 87.34 customers/hr

   μ (Shop 2) = 94.32s = 38.17 customers/hr

2. **Demand/arrivals**

   If each shop trades for 12 hours per day for 360 days per year, it is open for a total of 4320 hours. With 200 000 customers per year,

   Λ = 46.30 customers/hr

3. **Queueing & service level**

   It is recommended to model each shop as an M/M/s system, with M random arrivals, M random services, and s baristas. Erlang-C is used to estimate the probability that any given customer's total time (wait + service) is ≤ 120 seconds.

4. **Profit**

   Using the profit contribution of R30/customer, equating to R1000/day per barista, with between 2 and 6 baristas.

   Profit = 30 x (no. of served customers) – 1000 x s x 360

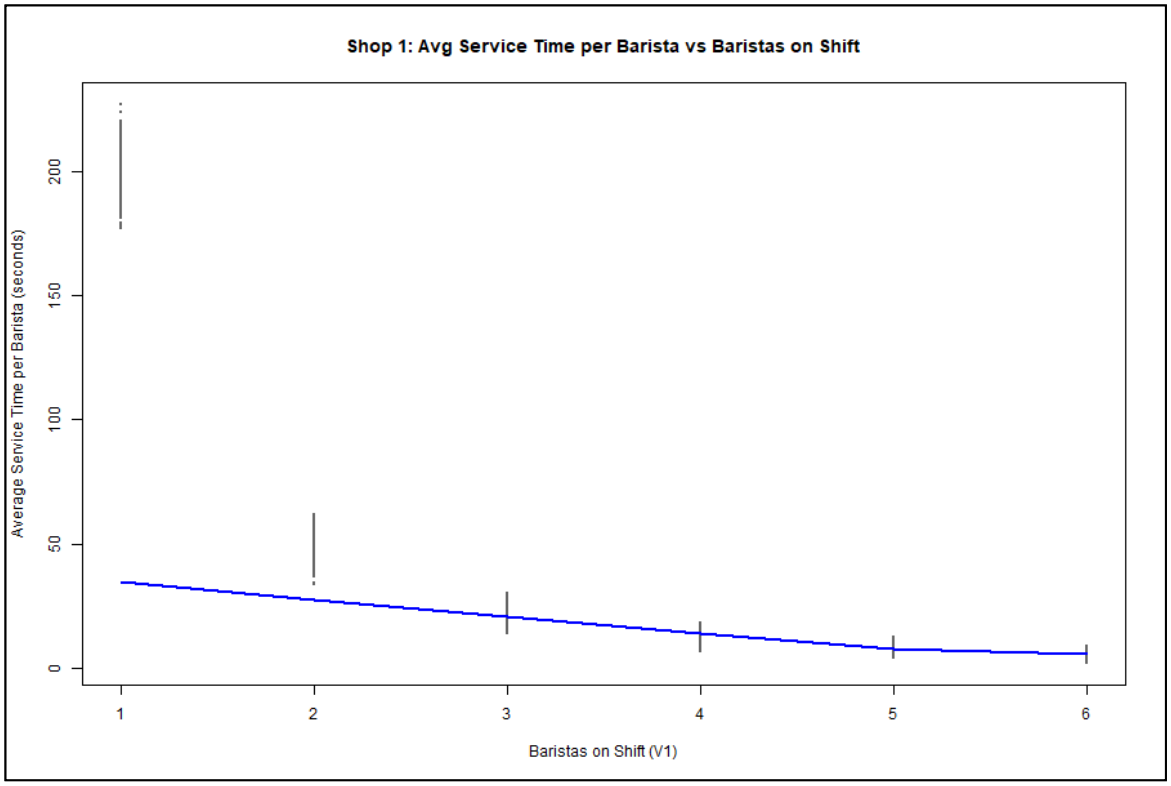
Significant statistics

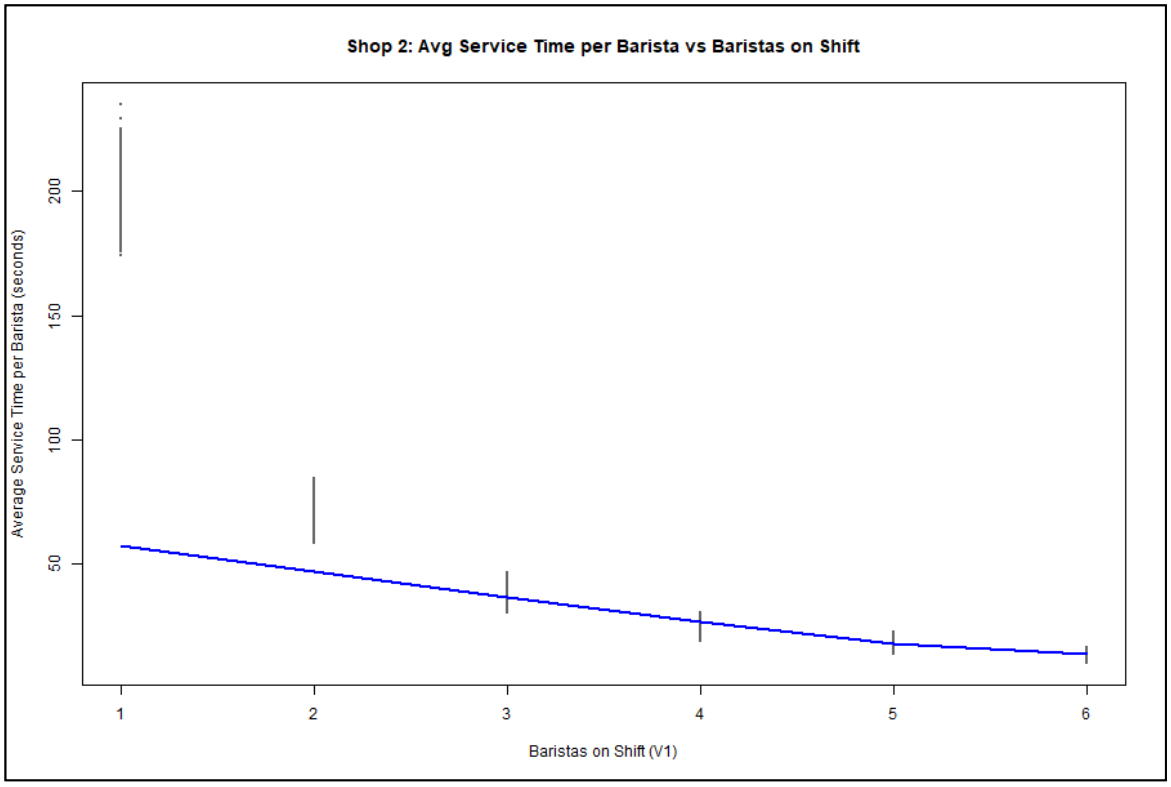
Shop 1 was seen to be the faster of the two with two baristas, achieving a service level within 120 seconds of around 99.33%. Shop 2, however, was observed to be significantly slower, needing at least 3 baristas to obtain a service level of more than ninety percent, within the same time constraint:

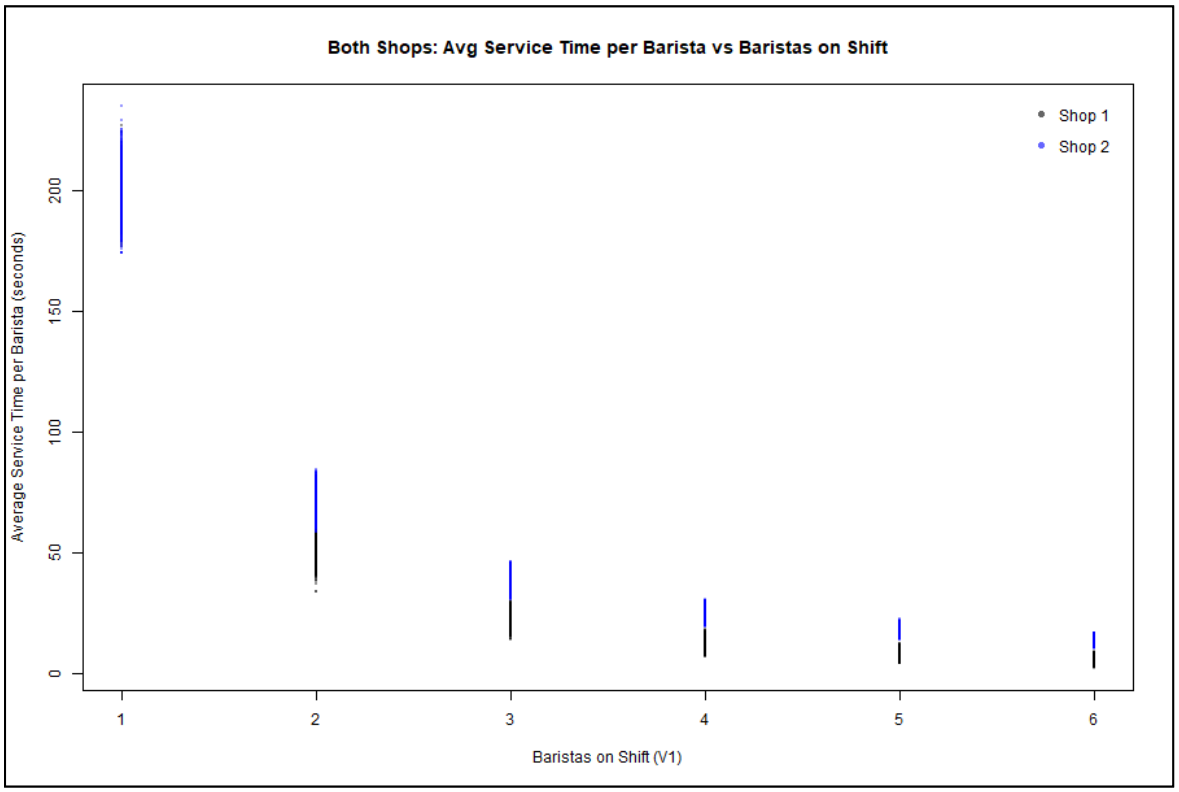
Staff	Shop	MeanServiceSec	ServiceRate_perHr	ArrivalRate_perHr	Served_perYear	Profit_ZAR	ServiceLevel_total<=120s
2	Shop 1	41.22	87.34	46.3	200000	R5 280 000.00	99.33%
2	Shop 2	94.32	38.17	46.3	200000	R5 280 000.00	63.04%
3	Shop 1	41.22	87.34	46.3	200000	R4 920 000.00	99.98%
3	Shop 2	94.32	38.17	46.3	200000	R4 920 000.00	91.10%
4	Shop 1	41.22	87.34	46.3	200000	R4 560 000.00	100.00%
4	Shop 2	94.32	38.17	46.3	200000	R4 560 000.00	98.20%
5	Shop 1	41.22	87.34	46.3	200000	R4 200 000.00	100.00%
5	Shop 2	94.32	38.17	46.3	200000	R4 200 000.00	99.69%
6	Shop 1	41.22	87.34	46.3	200000	R3 840 000.00	100.00%
6	Shop 2	94.32	38.17	46.3	200000	R3 840 000.00	99.96%

As seen in the table above, Shop 1 achieves the largest profit with 2 baristas on duty. Similarly, Shop 2 reaches the target service level with 3 baristas. Profits and percentage reliable service are calculated and displayed in the last two columns in the above table.

Graphical Representation




Shop 2: Avg Service Time per Barista vs Baristas on Shift
Average Service Time per Barista (seconds)
50
100
150
200
1
2
3
4
5
6
Baristas on Shift (V1)


Both Shops: Avg Service Time per Barista vs Baristas on Shift
Shop 1
Shop 2
Average Service Time per Barista (seconds)
0
50
100
150
200
1
2
3
4
5
6
Baristas on Shift (V1)

Part 6 – DOE & MANOVA/ANOVA

Introduction

To decide whether variations in group means are statistically significant, two methods are utilized, namely Design of Experiments (DOE) and Analysis of Variance (ANOVA or MANOVA). ANOVA concentrates on only one variable, and MANOVA expands this to include multiple dependent variables. DOE principles help us to determine whether process variables such as year, month and product type have a significant effect on the performance of the delivery process outlined in Part 3.

In the context of this analysis, delivery time is the main dependent variable, with year (2026 vs 2027) and product type are the independent variables. Testing whether the mean delivery times have changed significantly over time is appropriate due to the fact that the same procedures are used over a number of time periods. Depending on the number of dependent variables taken into consideration, an ANOVA/MANOVA test is conducted after the data has been summarised. By investigating the results, one can determine whether the observed differences in delivery times are the result of chance, or whether they represent significant process variation that may warrant managerial intervention.

Application of ANOVA/MANOVA

To determine whether the average delivery times for each product type in 2026 and 2027 (Years 1 and 2 respectively) differ significantly, an ANOVA test was performed using the corresponding chart findings from Part 3. Based on roughly symmetric histograms, previously obtained from Rstudio, normality was assumed for each delivery observation, which was sorted by product type and year.

While the alternative hypothesis (H_1) argues that at least one mean differs, the null hypothesis (H_0) states that there is no apparent variation between the mean delivery time between Year 1 and 2. The following was the specification for the one-way ANOVA model:

deliveryHours ~ Year + ProductType

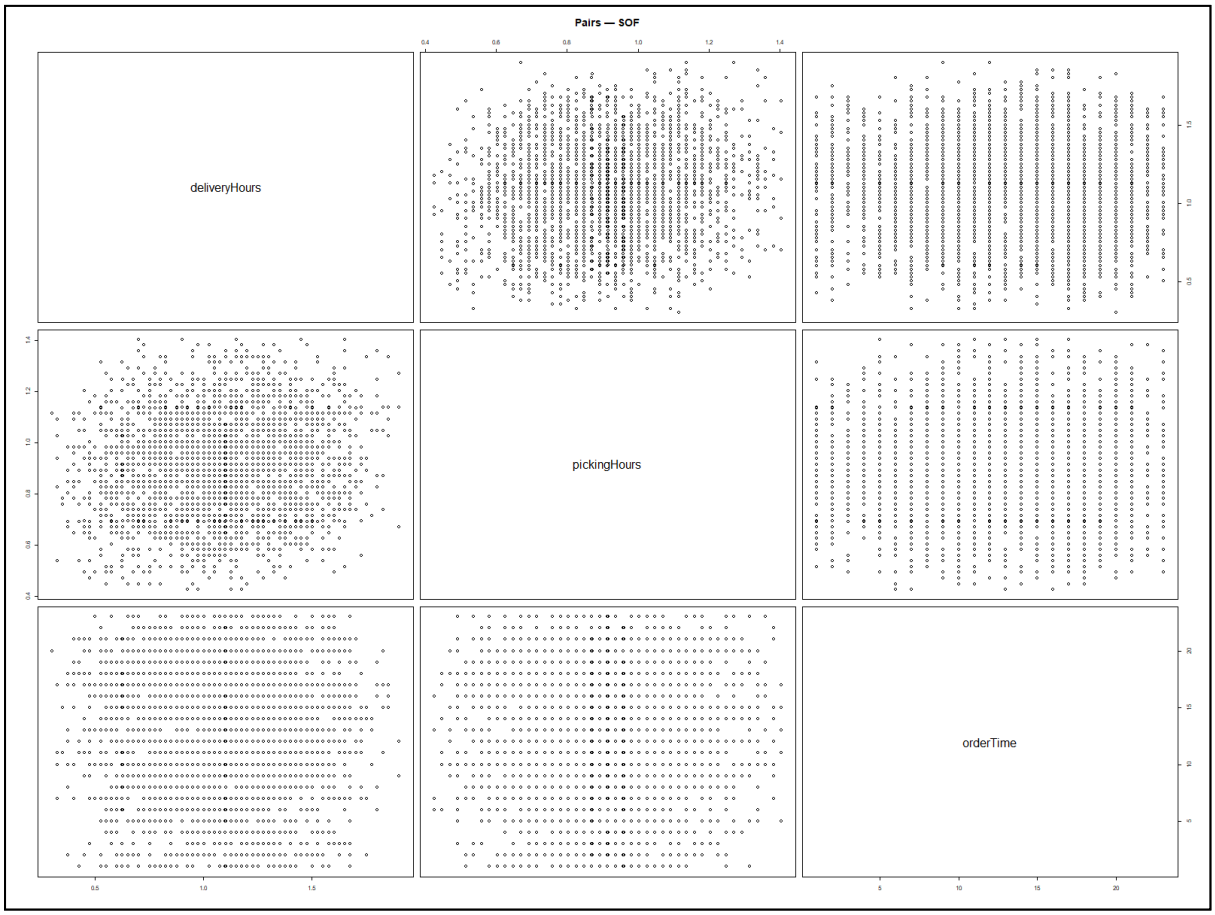
Interpretation

Upon investigation of the ANOVA results, there was no statistically significant difference in mean delivery times between 2026 and 2027 across all 6 major product types. For most products, the p-values were greater than 0.05. According to the stable control charts obtained in Part 3, this implies that the mean delivery procedures remained the same during the two years. In product types such as LAP and CLO, the second year showed slightly longer mean delivery times, although this could be due to operational bottlenecks or temporary shifts previously identified in Rule C from the SPC analysis.

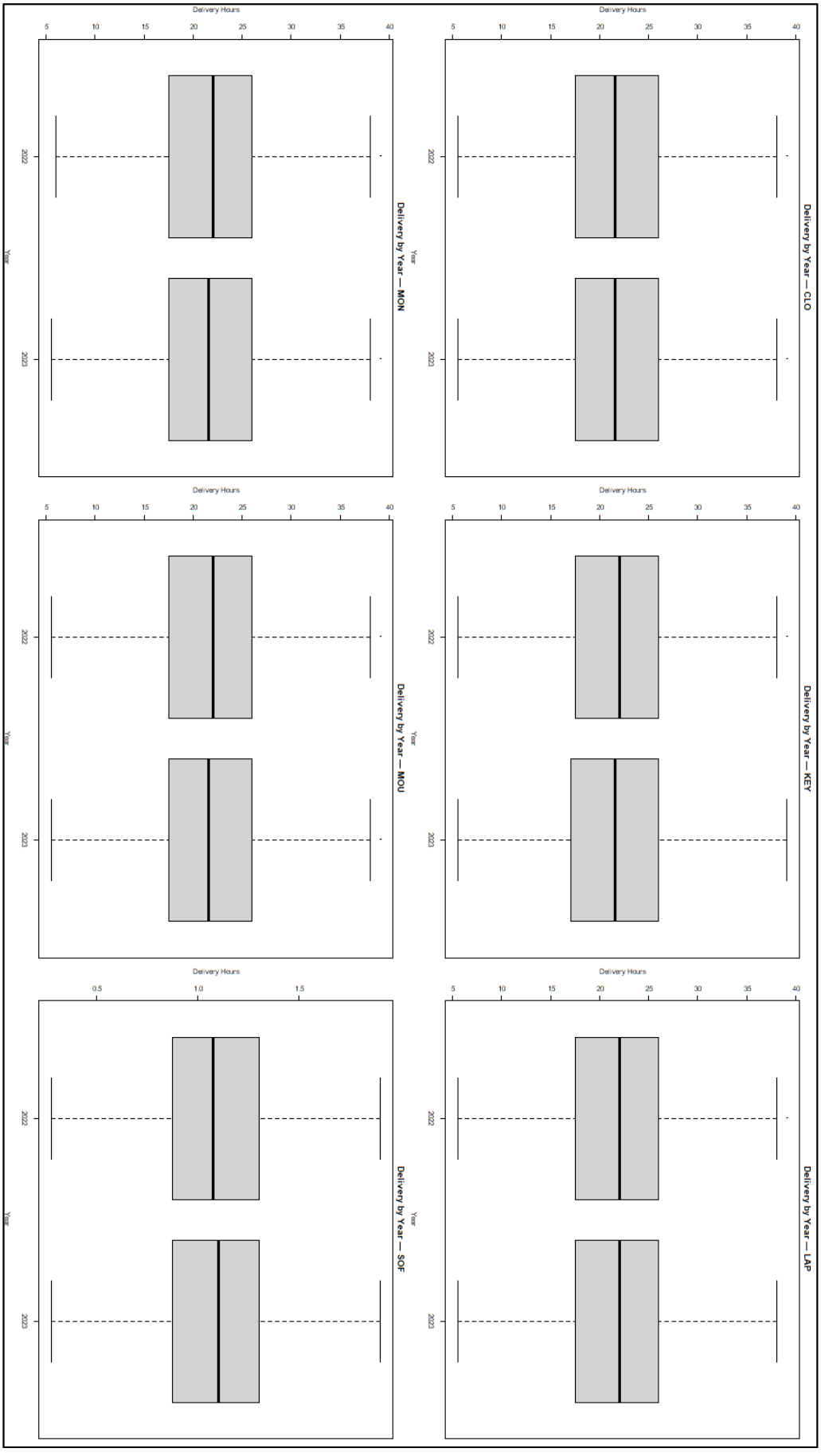
The same model was used to compare delivery times from month 1 to month 12, the results showing that seasonal variation was generally statistically insignificant. These findings indicate that variations in mean delivery times from month to month are most likely as a result of normal process variation rather than structural performance changes.

The graphs seen below support the fact that the process is capable and under control, showing that the mean and spread of delivery times for each product type were constant across both years.

MANOVA Pairs by Type



Product Delivery by Year



Part 7 – Reliability of Service

It is reasonable to assume sixteen workers are put on the roster, with each person independently 'showing up' with probability *p*.

From the histogram, total workers observed is calculated as follows:

$$12(1) + 13(5) + 14(25) + 15(96) + 16(270) = 6187$$

Mean workers present = 6187 / 397 = 15.5844

$$p = \text{mean present} / 16 = 15.5844 / 16 = 0.9740 = 97.40\%$$

no. people on roster *S*	P (X < 15)
16	0.063631
17	0.009071
18	0.00104
19	0.000101
20	0.0000087

Sales shortfall on a problem day = R20 000 [given]

Extra staff cost: R25 000 per person per month [given]

- Assume 30 days/month

$$C(S) = [25000 \times MAX(S - 16.0)] + [20000(30) \times P(X < 15 \mid S, p)]$$

S	Expcted monthly cost C(*S*) [ZAR]
16	R 38 179.00
17	R 30 443.00
18	R 50 624.00
19	R 75 061.00
20	R 100 005.00

My recommendation would be to increase the roster size from 16 to 17, which has a significant impact on expected monthly cost, lowering it by around R7700.

Conclusion

Throughout this report, multiple tests needed to be conducted in order to conclude our findings, such as Statistical Process Control testing towards the beginning, ending with variance analysis tests.

Sales data was cleaned, summarised and visually displayed using histograms, boxplots and trend charts. In the SPC tests, control limits were calculated using commonly-used constants which revealed that most sample means and standard deviations remained within the specified limits, suggesting a well-balanced process with minimal special cause variation.

ANOVA tests were used to determine significant mean differences among product categories and years, while MANOVA tests confirmed that, after expansion, some product groups show measurable differences, however the overall variation between years was statistically modest.

After comprehensive analysis, the company's operational and sales processes are statistically stable, with minor sources of variation needing attention. The utilized statistical methods mentioned previously present a sturdy framework for continuous quality amelioration, to secure process reliability as well as cost efficiency.