Quality Assurance

344

Analysis of business data

SU: 271271968

Table of Contents

Introduction

This report will conduct statistical analysis of several data files. We will use statistical analysis methods to conduct a thorough analysis and uncover underlying trends.

R Libraries

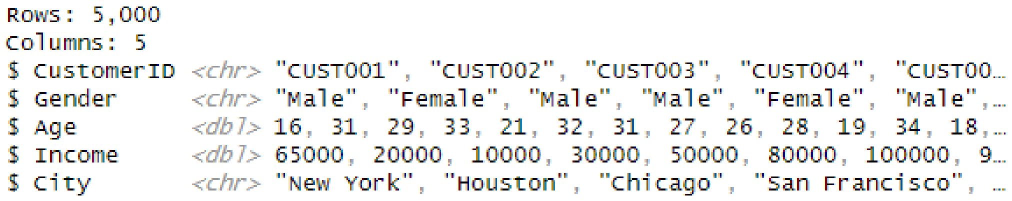
This code uses the tidyverse, ggplot2, dplyr, corrplot, tidyverse and readr libraries.

Part 1

1.2- Data Familiarisation & Descriptive Statistics

We load all datasets into R using the read_csv command. To get a high-level overview of the 4 datasets, we use the glimpse command. This returns the column headers and a statistical summary.

>glimpse(customer_data)



```
Rows: 5,000
Columns: 5
$ CustomerID <chr> "CUST001", "CUST002", "CUST003", "CUST004", "CUST00…
$ Gender     <chr> "Male", "Female", "Male", "Male", "Female", "Male",…
$ Age        <dbl> 16, 31, 29, 33, 21, 32, 31, 27, 26, 28, 19, 34, 18,…
$ Income     <dbl> 65000, 20000, 10000, 30000, 50000, 80000, 100000, 9…
$ City       <chr> "New York", "Houston", "Chicago", "San Francisco", …
```

We learn there are 5 columns in customer_data each customer is given a unique ID and their gender, age, income and city is given. This can be applied to all other datasets to gain a high-level overview. In products_data, which contain the same data as products_Headoffice. We notice discrepancies between the ProductID and Category, for example SOF002 is categorised as Cloud Subscription which is likely to be wrong.

Missing values

Using the functions *sapply* and sum(is.na()) for all 4 files, we learn there are no missing values that need to be dealt with.

Summary Statistics

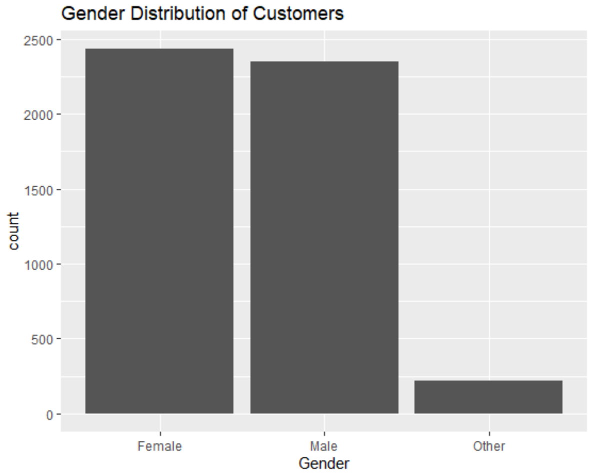
Using functions such as summary(), summarise(), describe(), skimr()
and table(), we can gain insight into the data.

Central Tendency

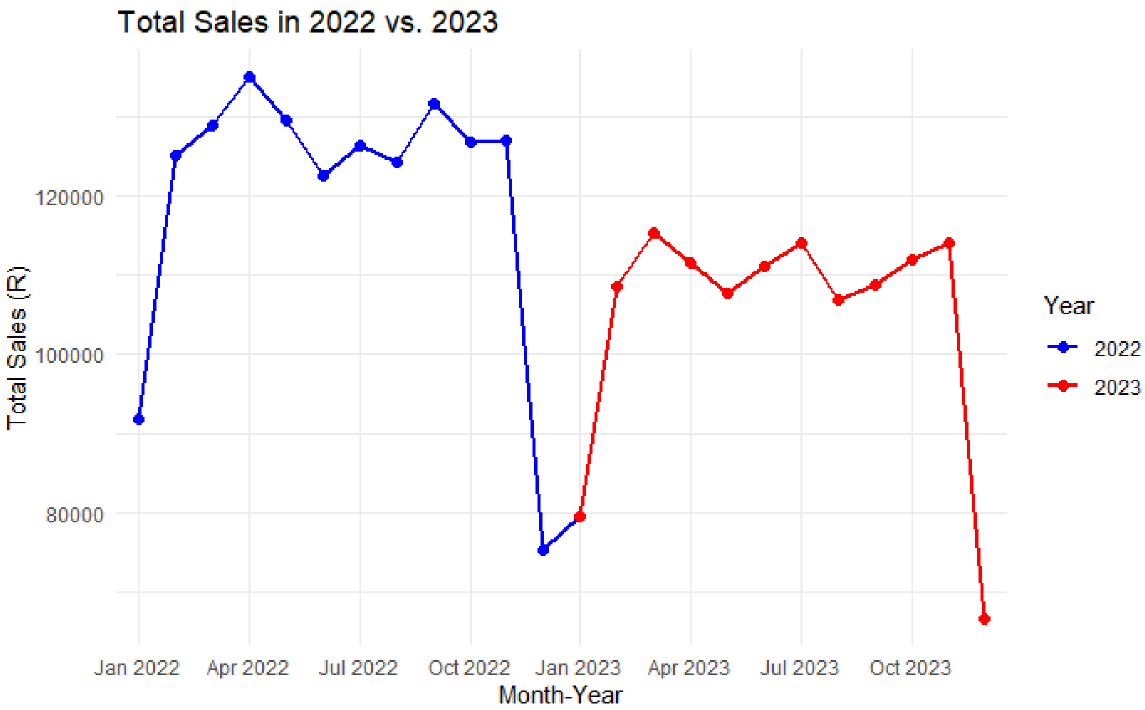
Factor	**Skewed**	**Interpretation**
Age	Symmetric	Data is evenly distributed
Income	Left	Very small values skew the data, median might be a better measure of centre
Quantity	Right	Very large values skew the data, median might be a better measure of centre
orderTime	Left	Very small values skew the data, median might be a better measure of centre
orderDay & orderMonth	Symmetric	Data is evenly distributed
orderYear	Symmetric	Data is evenly distributed
pickingHours	Right	Very large values skew the data, median might be a better measure of centre
deliveryHours	Symmetric	Data is evenly distributed
SellingPrice	Right	Very large values skew the data, median might be a better measure of centre
Markup	Symmetric	Data is evenly distributed

Our central tendency assumptions are backed up by the skew column. Sales, volume, listings and inventory are > 0, indicating the data is skewed right. The kurtosis values measure the distribution, where kurtosis = 3, indicates a normal distribution, <3 means flat and >3 means peaked. Sales, volume, listings and inventory show peaked kurtosis values.

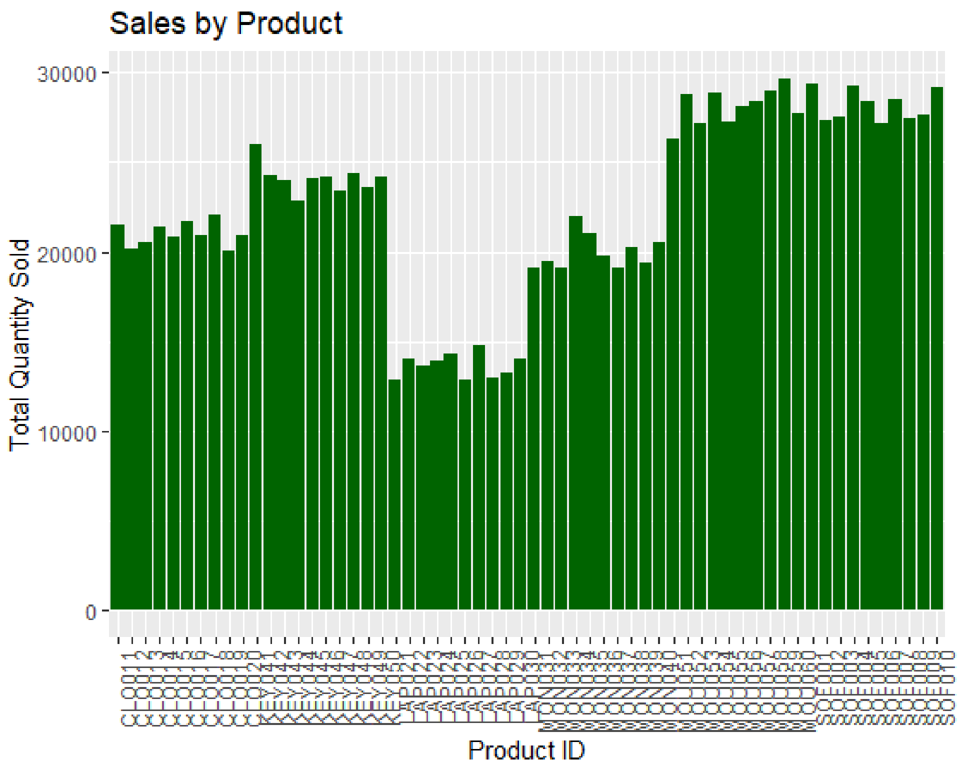
Data Visualisations



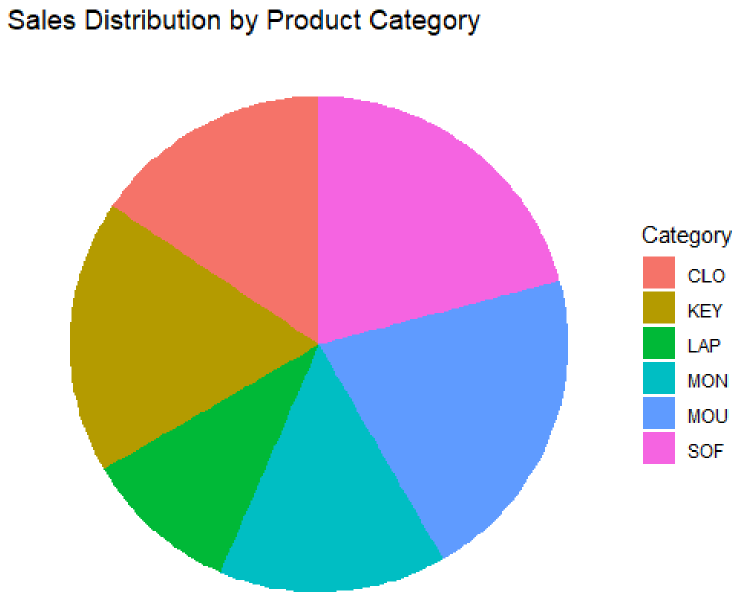
There is a near 50/50 split of male and female customers. No male, female or other gender preference can be derived for the sales.



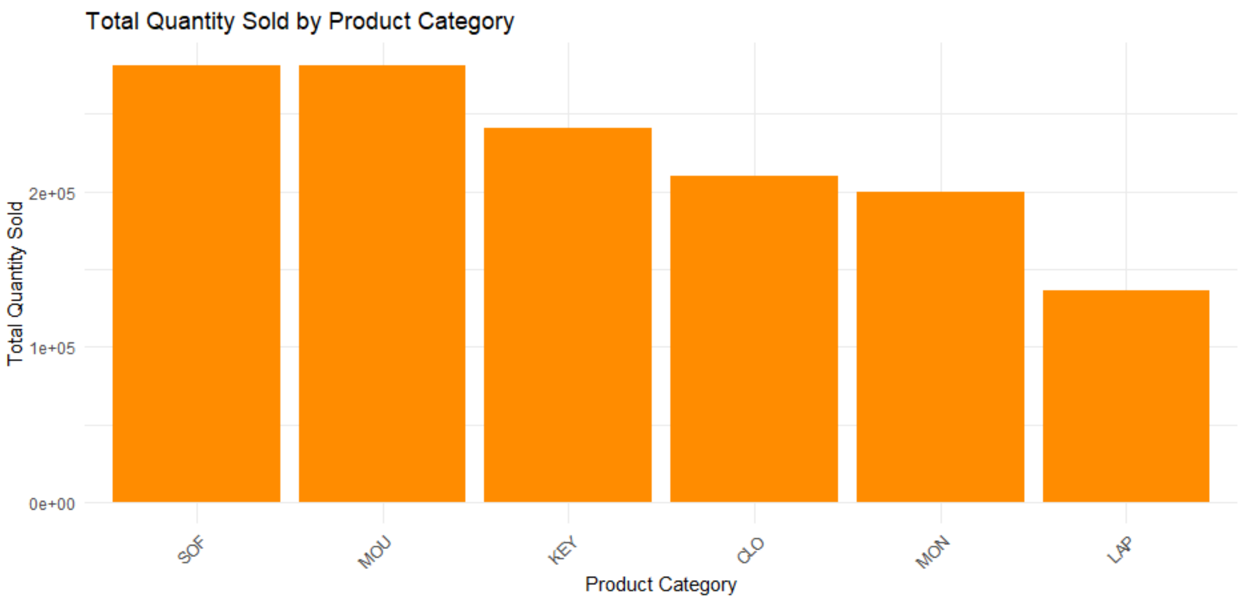
Plotting total sales vs, the months of 2022 and 2023. We noticed for the same month, one year later, a significant decrease in sales. The total sales for 2023 are substantially lower than the sales recorded in 2022. Additionally, there is clear seasonality in total sales. A trough is recorded between the months of December and January, whilst sales tend to be high in the months of April and October.



The bar chart above is not clear. To overcome this, we categorise by product type and extract the first 3 characters of each product ID.

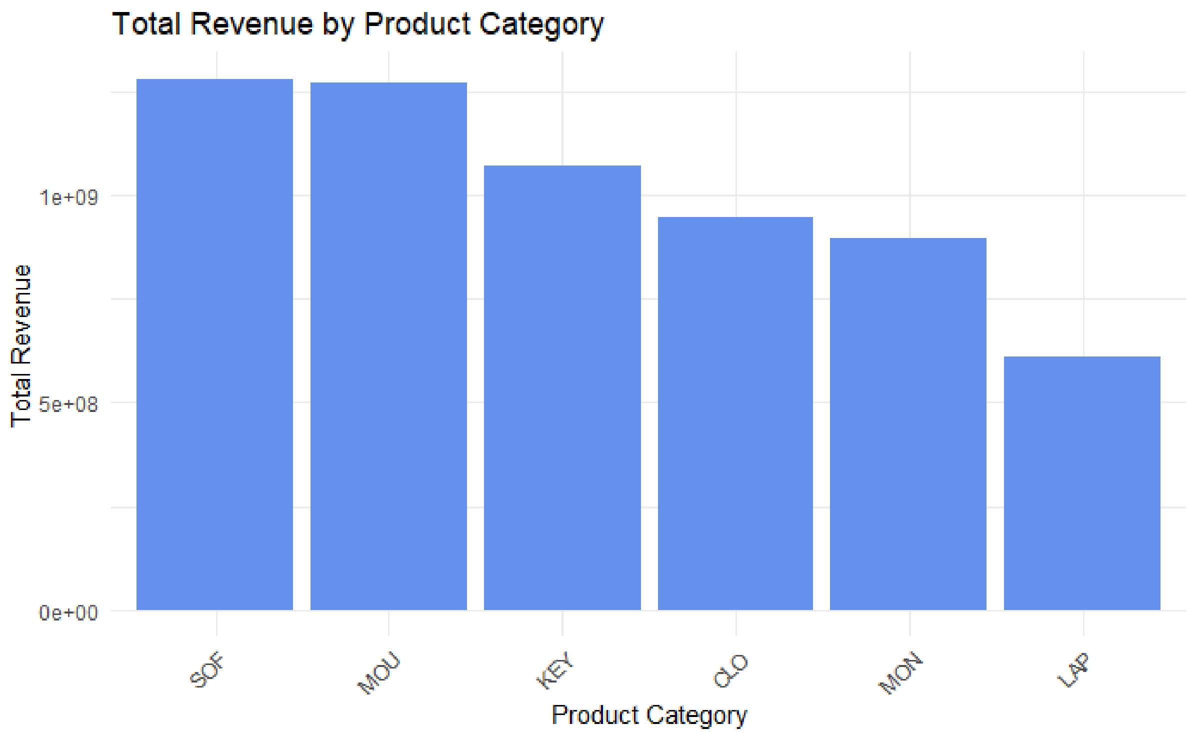


The pie distribution of sales by category (same information as below) but helps us visualise that all the products are more or less equal in their contribution to total sales.

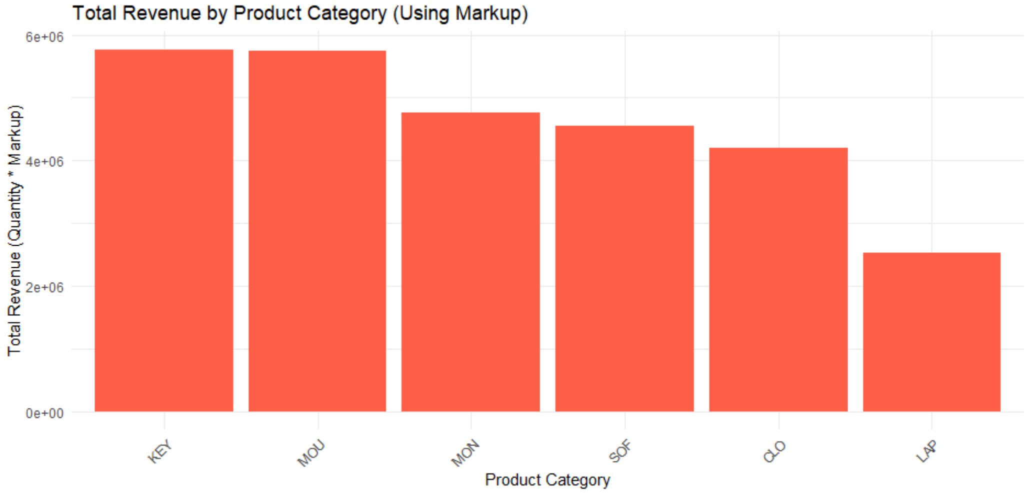


We can now see that SOF and MOU products account for the highest number of sales, whereas LAP accounts for the least number of sales.

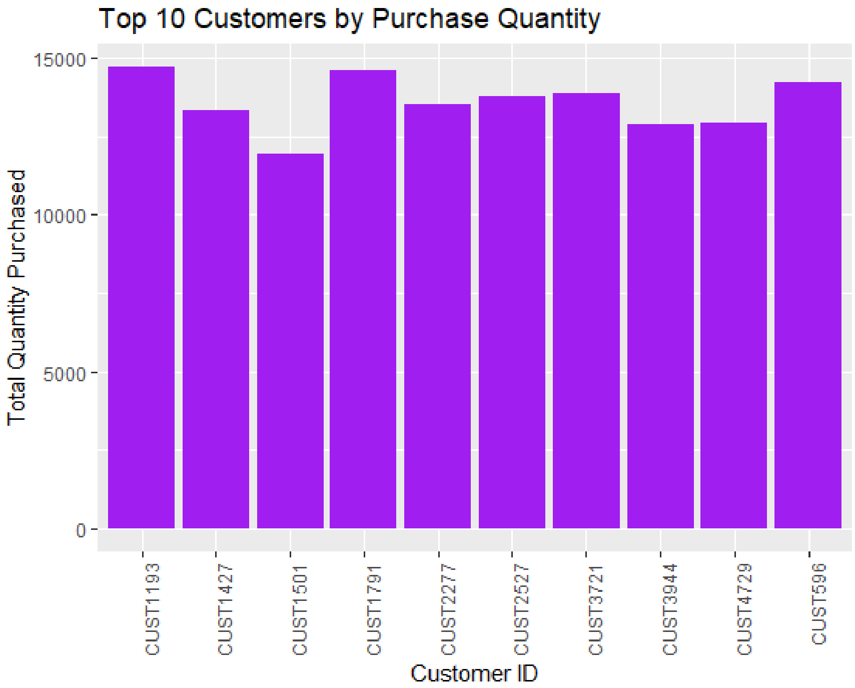
Now, we take a look at the revenue produced by each product category.



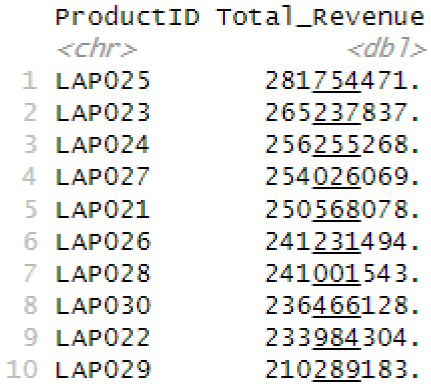
The total revenue is calculated by multiplying the sales price by the total number of sales for the given product category. We can see LAP and MON account for the majority of the company’s revenue.



By comparing the two graphs, we notice that **revenue generated by sales is not correlated to the profit** (markup) of the given product. Products with lowest revenue (KEY and MOU) have the highest markup generated, and products with the highest generated revenue tend to have the lowest total markup generated. This could be down to economies of scale, as MOU and KEY have a relatively high number of sales. It’s an interesting point to learn as it shows quantity and profit are not necessarily directly correlated



The graph shows the quantity of items bought by the top 10 customers. From this we can derive our top customers by sales quantity or revenue. Additional information such as delivery address, order regularity and nature of order could help identify the needs of the business. Does the company require products quickly, or as cheaply and efficiently as possible? Can two orders be merged to reduce shipping costs and complexity? Being able to answer these questions could help improve service level and customer relations with the company's most important customers.

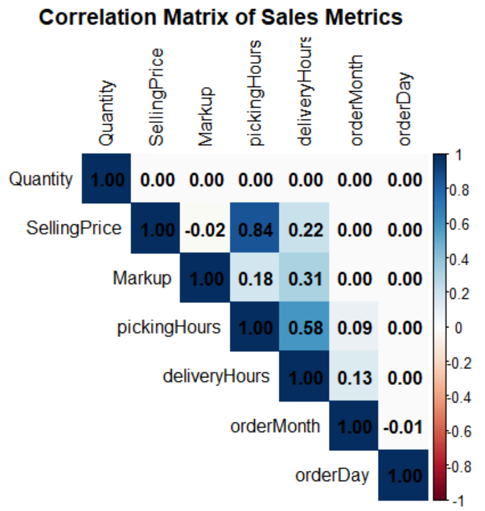


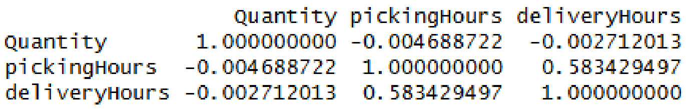
```
   ProductID Total_Revenue
   <chr>             <dbl>
 1 LAP025       281754471.
 2 LAP023       265237837.
 3 LAP024       256255268.
 4 LAP027       254026069.
 5 LAP021       250568078.
 6 LAP026       241231494.
 7 LAP028       241001543.
 8 LAP030       236466128.
 9 LAP022       233984304.
10 LAP029       210289183.
```

Laptops are the top 10 highest revenue producing products for the company.

The table above shows the top 10 ProductIDs in terms of revenue generated. This is computed by multiplying the SellingPrice by the total number of units sold.

An interesting trend emerges: the number of units sold does not predict the revenue. More cheaper items are sold overall, but the fewer, more expensive items account for a bigger portion of the businesses overall income.





```
                  Quantity pickingHours deliveryHours
Quantity       1.000000000 -0.004688722  -0.002713013
pickingHours  -0.004688722  1.000000000   0.583429497
deliveryHours -0.002713013  0.583429497   1.000000000
```

From the correlation matrix above we see the data has a weak positive correlation, especially between SellingPrice and Markup. The strongest correlation is between SellingPrice and pickingHours, as well as pickingHours and deliveryHours, trends which will be explored more in depth later.

Recommendations After Part 1:

In the data analysis we were able to gain both a high level (overview) and low level (in depth) understanding of the businesses offerings. The business has a number of repeat customers which contribute to their earnings. **Additional information** should be gathered to refine their product and service offerings.

High product sales does not dedicate high revenue to the business, but may be critical to getting sales on bigger tickets items. These are often referred to as lead **items** or **traffic drivers**. Exploring more, cheaper items could be useful, such as **tech accessories**. (eg, someone with the intention of buying a monitor could walk out having bought a laptop).

Promotions should be run between the months of **December and Feb** as well as **June to August** to counter the drop in sales. More research is needed to fully understand the drop in sales during the festive season, as one would expect increased sales due to the holiday period and back to work/school period.

Part 3

3.1 - Statistical Process Control (SPC)

The purpose of this step is to initialise X-bar and S control charts for each product type using the earliest available delivery-time data. These charts allow us to monitor whether process variation and average delivery performance remain statistically stable.

We use the first 30 subgroups (each containing 24 observations) to determine the centre lines and control limits (±1σ, ±2σ, ±3σ) for both charts.

Some data preparation is necessary, such as merging sales data, joining the sales data with product and customer tables to gather full context. A DateTime variable is created, combining year, month, day and order time.   We then sort it chronologically to reflect arrival order. For the SPC analysis, deliveryHours is selected as our variable of interest.

SPC Assumptions:

Subgroup size (n): 24
No. of subgroups (k): 30
d_2 constant: 2.223
A_2 constant: 0.223

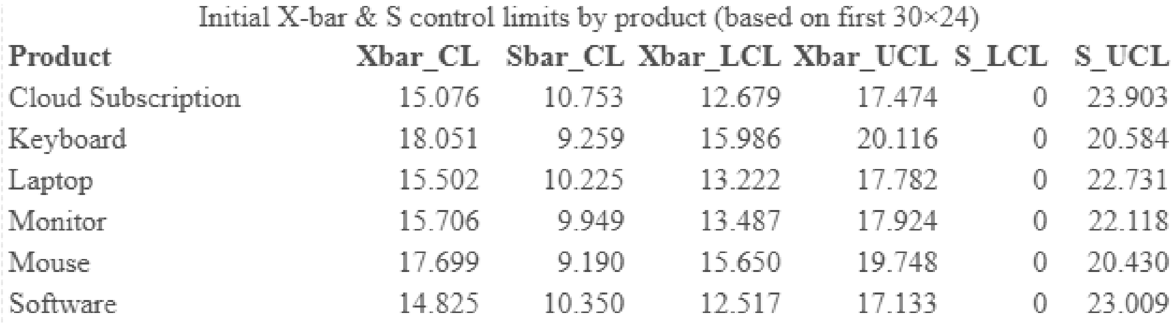
To calculate the mean ($\overline{X}$) and sample std. deviation (s) for each sample:

1. Split the order delivery data into 30 samples, each of 24 values (as per brief)
2. X bear = mean of all X; S bar = mean of all S
3. Calculate the control limits:

$$UCLx = \overline{X} + A2\overline{S} \qquad LCLx = \overline{X} - A2\overline{S}$$

$$UCLs = B4\overline{S} \qquad LCLs = B3\overline{S}$$

The following limits were generated:



Initial X-bar & S control limits by product (based on first 30×24)

Product	Xbar_CL	Sbar_CL	Xbar_LCL	Xbar_UCL	S_LCL	S_UCL
Cloud Subscription	15.076	10.753	12.679	17.474	0	23.903
Keyboard	18.051	9.259	15.986	20.116	0	20.584
Laptop	15.502	10.225	13.222	17.782	0	22.731
Monitor	15.706	9.949	13.487	17.924	0	22.118
Mouse	17.699	9.190	15.650	19.748	0	20.430
Software	14.825	10.350	12.517	17.133	0	23.009

The values calculated above will be plotted on the X and S bar charts and serve as a guide to identify abnormal patterns.

$\overline{X}$ -chart overview: Chart of averages

For all product categories (and thus $\overline{X}$-charts), a similar trend arises. For the first samples, the deliveries are well within the 3σ limits. This is to be expected as the limits are based on the first 30 samples. As the year progresses, however, the mean delivery time tends to increase well beyond the upper 3σ and UCL limit. The process is thus statistically unstable. In the new year, the same trend appears as delivery times start off within limits for roughly the first half of the year, and slip away as the year progresses. There is a slight improvement in 2027, however, it is unclear if this is a direct improvement from changes made within the company, or externally induced via a decrease in demand, among other factors. The charts above don’t give insight into the root cause, but rather signals *when* a change occurred, helping narrow down where attention is needed.

It is important to note the mean times on the y-axis are not all equal. Software has a -3σ = 0.8 and +3σ = 1.1, whereas most other product categories run from -3σ = ≈16 and +3 σ = ≈23.

From the charts it is clear investigation into why delivery times tend to increase into the year is needed. Customer satisfaction is guaranteed to decrease as delivery times increase. The company must make changes to ensure the processes remain statistically stable throughout the year.

s-chart overview: Chart of standard deviation

The s-chart shows the variation within each sample of 24 deliveries. The chart does not follow the same worrying trend as the $\overline{X}$-chart. Variation remains mostly stable with occasional breaches of upper and lower 2σ line and rare breaches of the 3σ line.

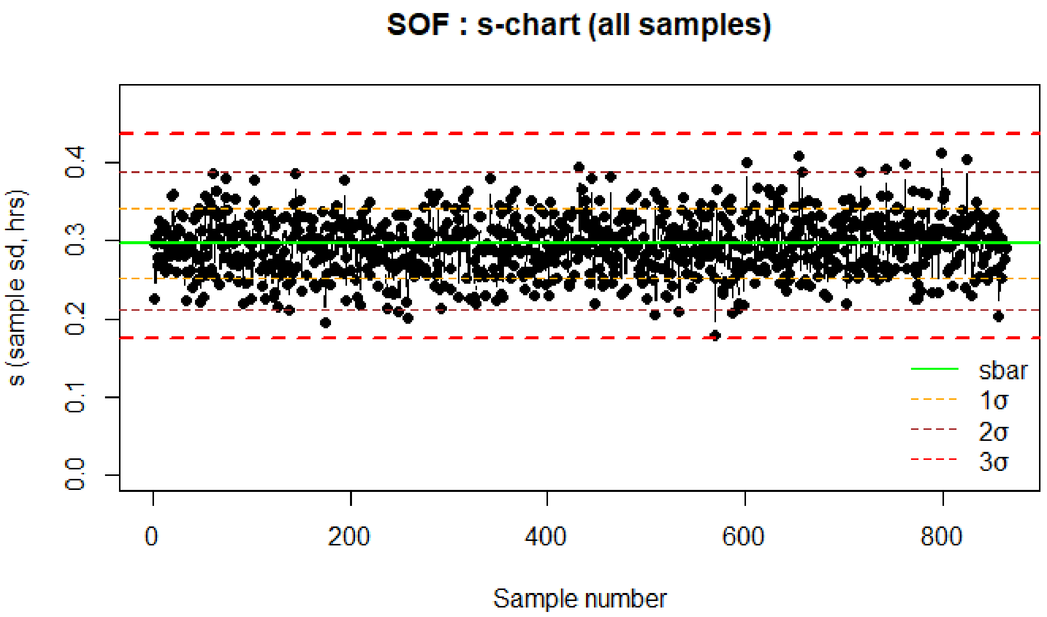
The lower dashed line is the LCL which is calculated as LCL_s*(B_3*s’ or 0 if negative). The upper dashed line is calculated as UCL_s*(B_4*s’) .

The solid horizontal line is the CL_s (mean of sample standard deviations).

The middle-dashed lines are the rule B limits (1σ) .

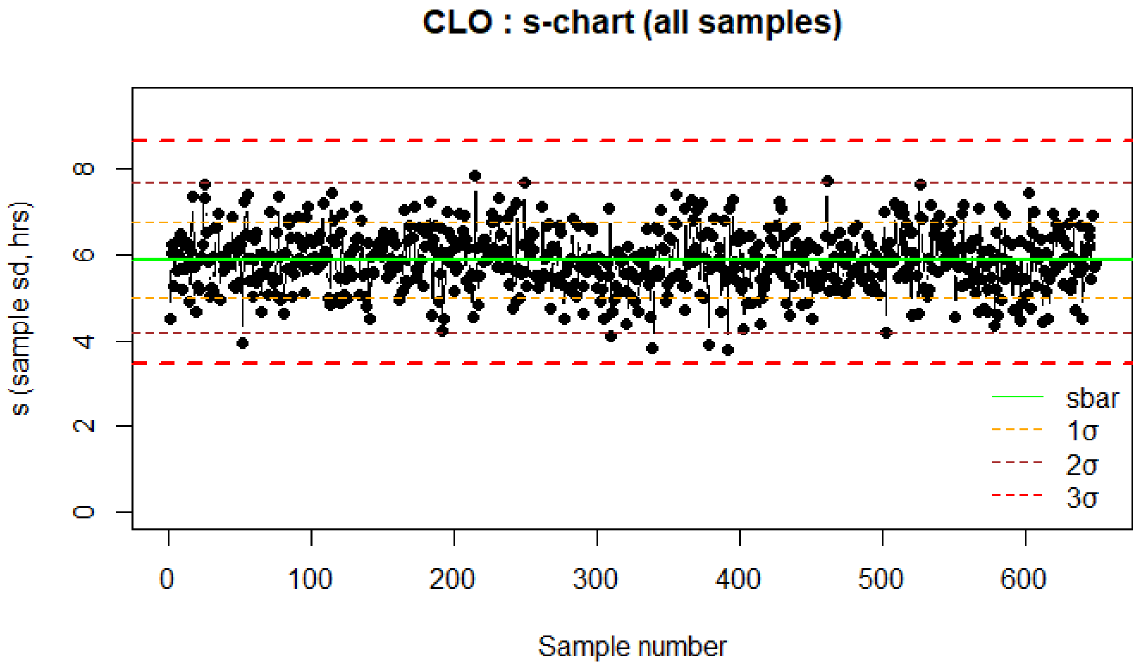
3.2) SPC Charts

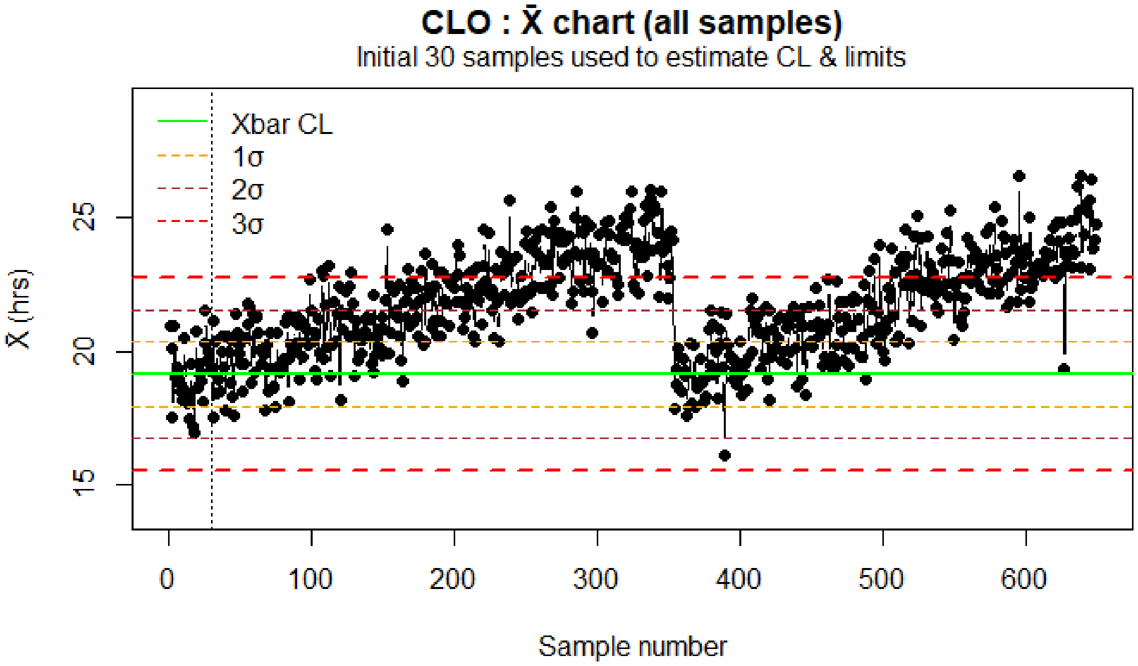
Product Type: Software



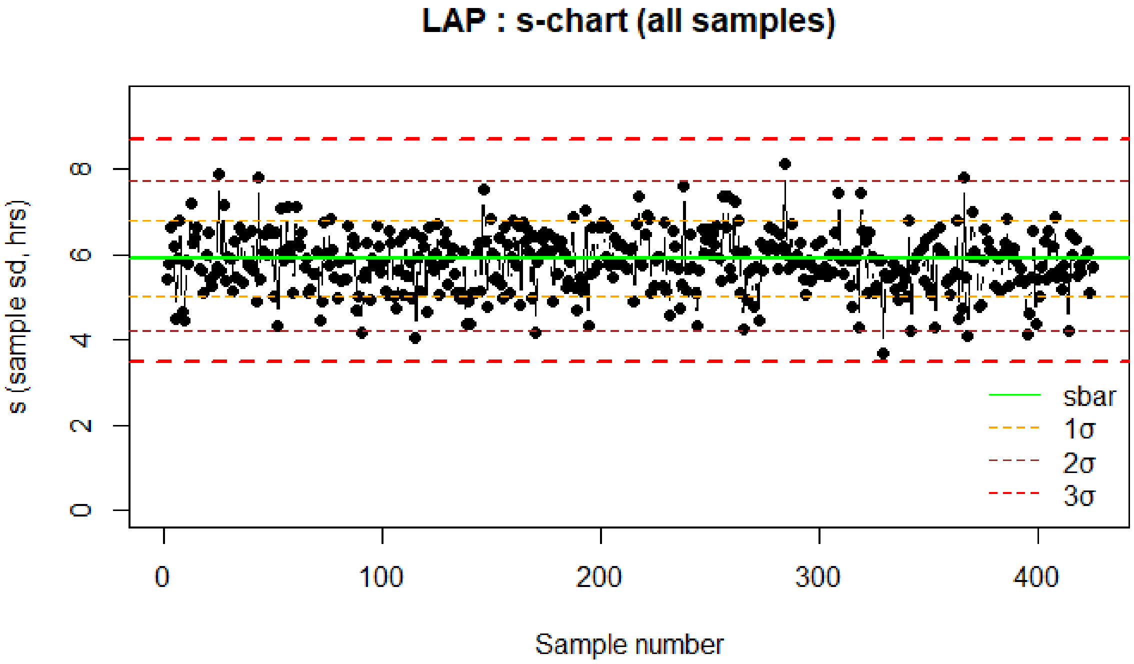
Software has a much faster mean delivery time of just under 1 hour. However the same trend of increasing delivery time as the year progresses is still true.

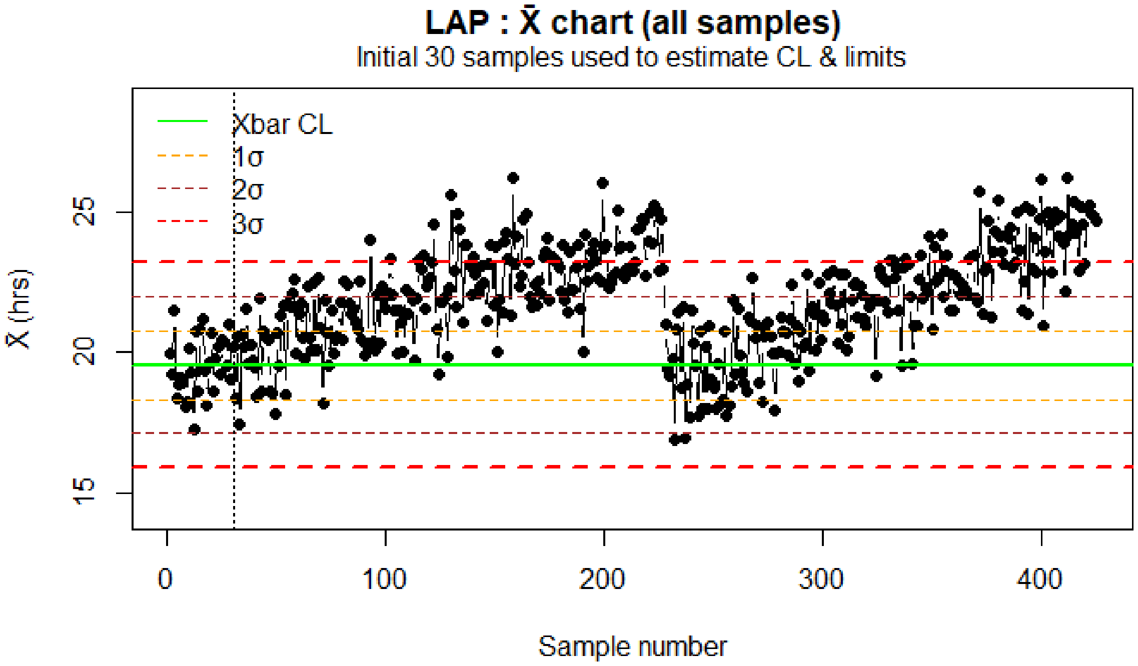
Product Type: Cloud Software





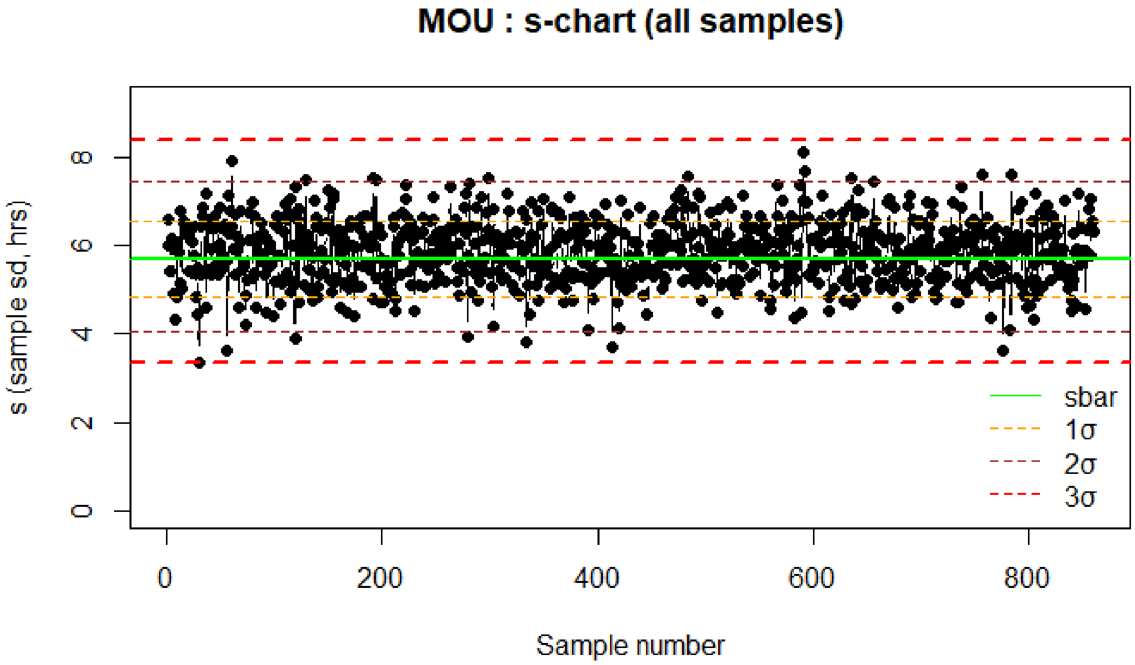
Product Type: Laptop

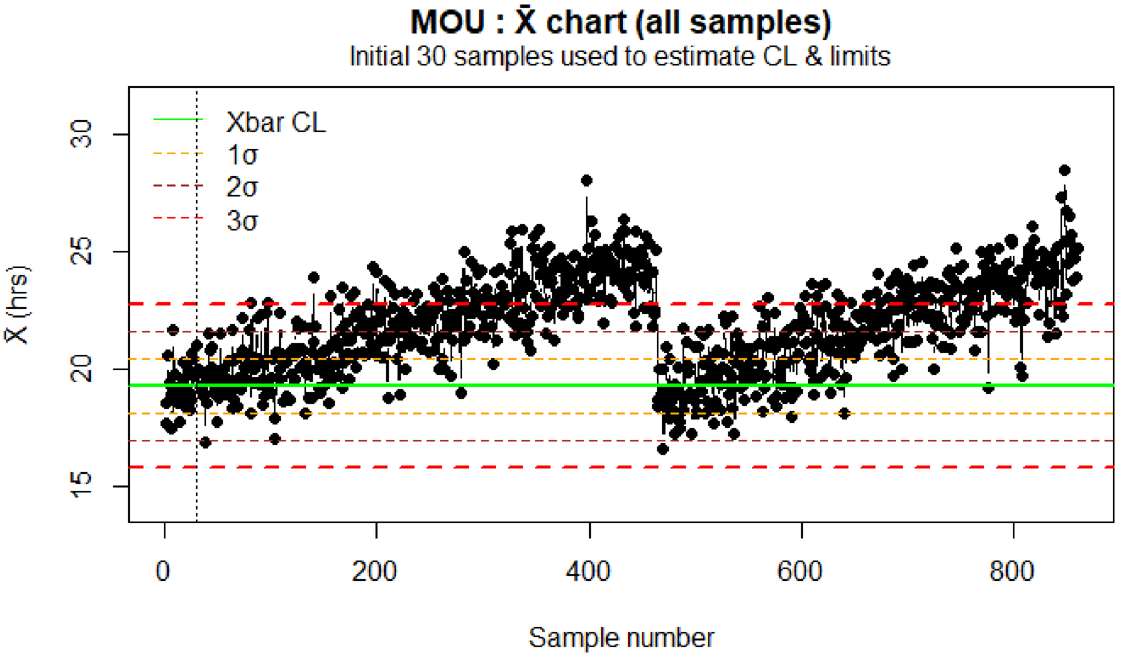




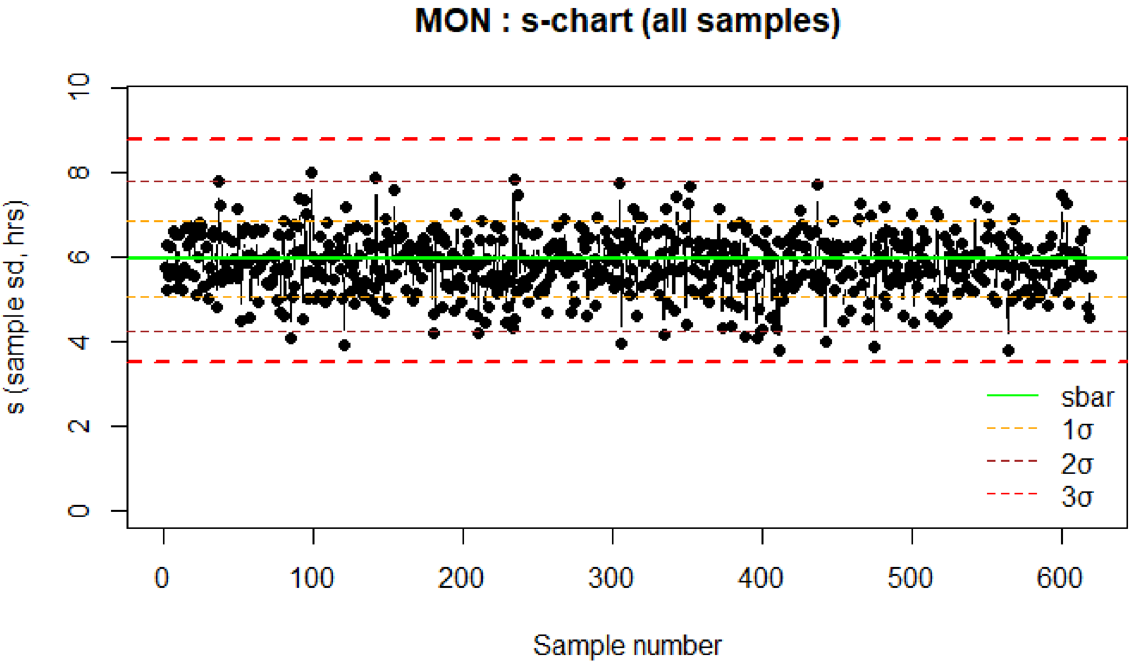
LAP shows the most stable process variability, with the longest run within the 1 ± 1σ zone – with potential to be used as a benchmark group as an indication of consistent process speed. MOU shows subgroups exceeding rule B, with a sharp decline in the last 5 samples.

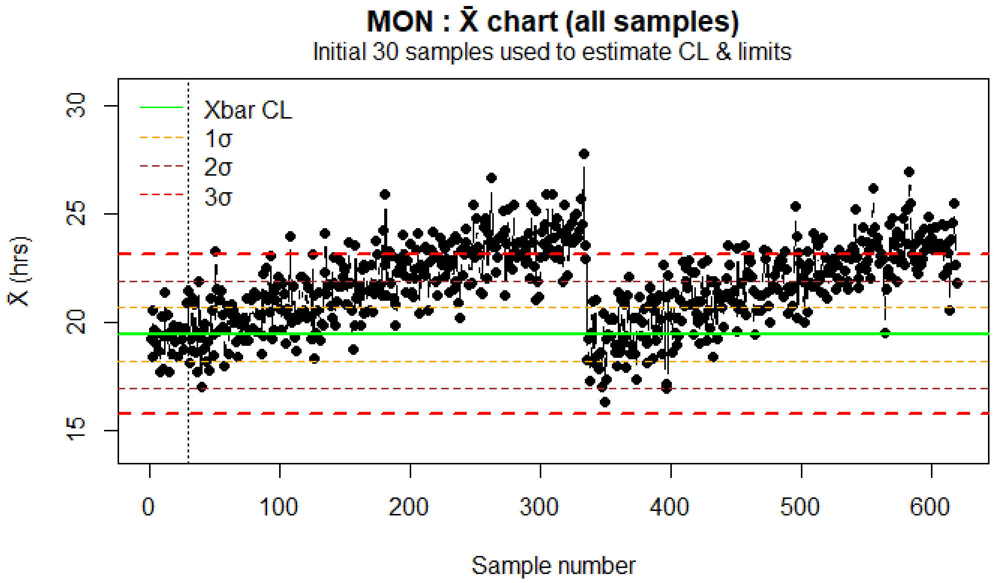
Product Type: Mouse



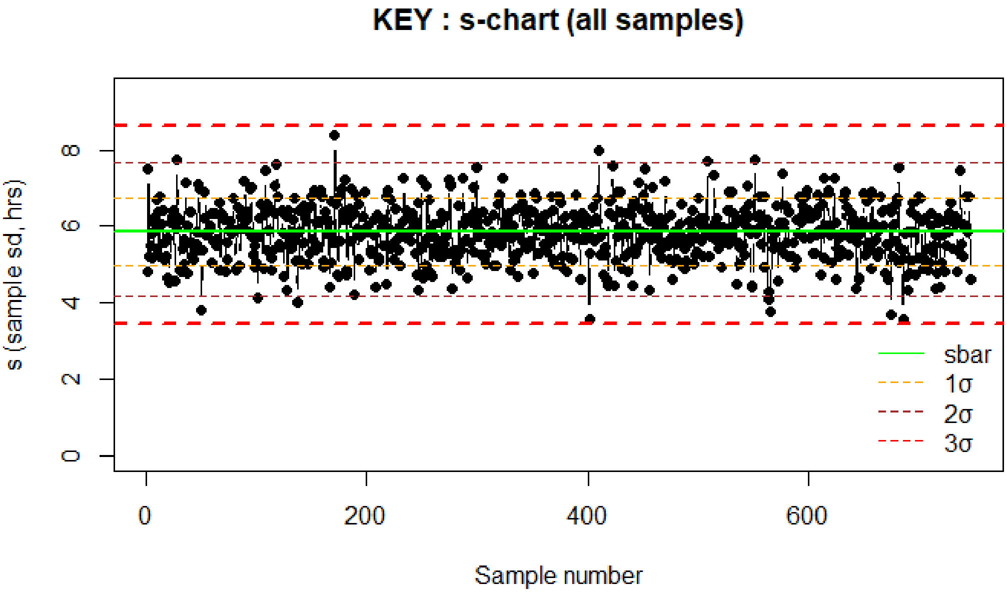


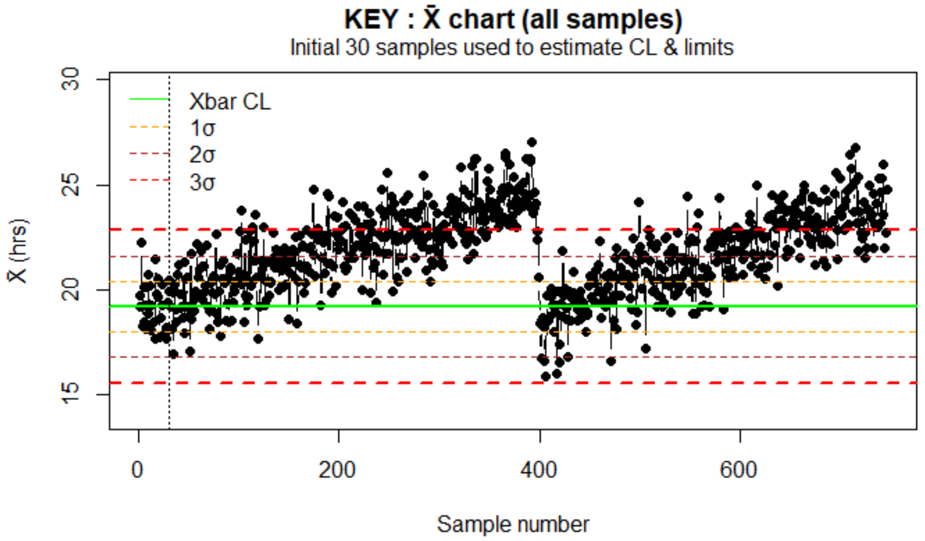
Product Type: Monitor





Product Type: Keyboard





3.3 - Capability Indices Calculations

We will now calculate the Process Capability Indices (Cp, Cpi, Cpl, and Cpk)

- **Cp:** Process Capability Index - Measures the potential process capability assuming the process is centred

$$Cp \ = \ \frac{USL - LSL}{6\sigma}$$

- **Cpl:** Lower Process Capability Index - Measures how well the process meets lower specification limit

$$Cpl \ = \ \frac{USL - \mu}{3\sigma}$$

- **Cpu:** Upper Process Capability Index - Measures how well the process meets the upper specification limit

$$Cpu \ = \ \frac{\mu - LSL}{3\sigma}$$

- **Cpk:** Process Capability Index - Measures the actual process performance, taking into account the process mean μ

$$Cpk \ = \ min(Cpl, Cpu)$$

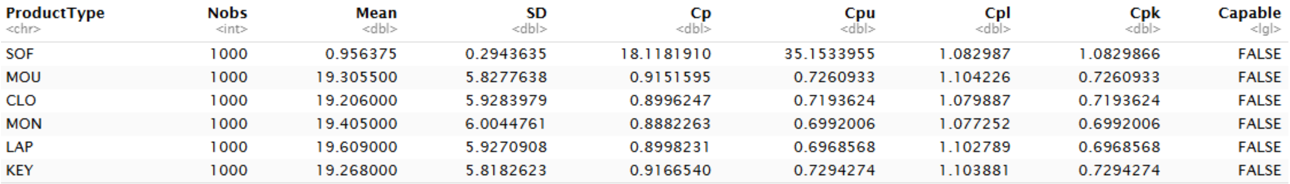
Calculations:

USL = 32 hours
LSL = 0 hours
$\sigma \ = $ sample standard deviation

μ = sample mean

Interpretation:

To determine if a product is able to meet the voice of the customer (VOC), there are general guidelines. Cp and Cpk > 1.33 indicate a process is capable. Product types with Cp or Cpk <1.33 are not capable.



ProductType	Nobs	Mean	SD	Cp	Cpu	Cpl	Cpk	Capable
<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<lgl>
SOF	1000	0.956375	0.2943635	18.1181910	35.1533955	1.082987	1.0829866	FALSE
MOU	1000	19.305500	5.8277638	0.9151595	0.7260933	1.104226	0.7260933	FALSE
CLO	1000	19.206000	5.9283979	0.8996247	0.7193624	1.079887	0.7193624	FALSE
MON	1000	19.405000	6.0044761	0.8882263	0.6992006	1.077252	0.6992006	FALSE
LAP	1000	19.609000	5.9270908	0.8998231	0.6968568	1.102789	0.6968568	FALSE
KEY	1000	19.268000	5.8182623	0.9166540	0.7294274	1.103881	0.7294274	FALSE

According to the valuations, no products are able to meet the VOC.

3.4 - Control Issues (Rules A,B,C)

Rule A – Points above the UCL_s

For the s charts, all the samples are within the 3σ control limits, despite large fluctuations. Some products such as keyboards, mice, monitors and laptops still show large variations and flirt with the extremities of the UCL.

Rule B – Longest run within ±1σ of CL_s

This rule shows periods of maintained process control.

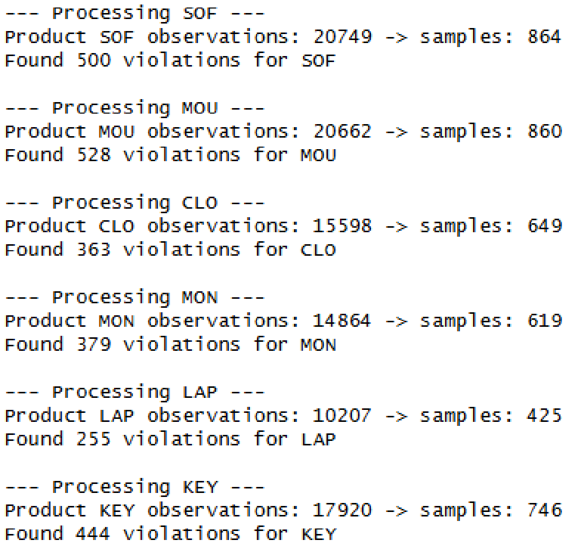
Software shows the most stable process variability, with the longest run within the $\pm\ 1\sigma$ zone – with potential to be used as a benchmark group as an indication of consistent process speed.

Rule C – 4 consecutive samples above the upper 2-sigma line.

All product types trigger rule C in both the s chart. By analysing the graphs we can determine when each product type exceeded the rule and when more detailed analysis needs to take place to determine why the rule was triggered.

Product	Total Violations	Max Consecutive Violations	First 3 violations	Last 3 violations
Software	152	1	200,202,203	850,852,854

Mouse	152	9	171,172,173	862,863,864
Cloud Software	80	1	179,182,188	610,651,662
Monitor	106	6	195,196,197	612,613,614
Laptop	75	1	100,110,130	412,423,434
Keyboard	87	16	171,172,173	737,747,767



```
--- Processing SOF ---
Product SOF observations: 20749 -> samples: 864
Found 500 violations for SOF

--- Processing MOU ---
Product MOU observations: 20662 -> samples: 860
Found 528 violations for MOU

--- Processing CLO ---
Product CLO observations: 15598 -> samples: 649
Found 363 violations for CLO

--- Processing MON ---
Product MON observations: 14864 -> samples: 619
Found 379 violations for MON

--- Processing LAP ---
Product LAP observations: 10207 -> samples: 425
Found 255 violations for LAP

--- Processing KEY ---
Product KEY observations: 17920 -> samples: 746
Found 444 violations for KEY
```

Part 4

4.1 – Type I Error (α): False Alarm

In the context of SPC, a type I error occurs when we incorrectly reject the null hypothesis (H_0), and assume there has been a shift or increase in variation (H_a). For the given case, that would occur when the sample falls outside of the UCL or LCL when the process is in fact in control.

The probability of finding 1 sample above the centerline is 50% (0.5) because the data is assumed to be in control and centered on the centerline with a normal distribution.

7 consecutive samples above the centreline in an X-bar chart is an indication of a shift or systematic issue in the process, the process must thus be investigated.

$$P(7\ consecutive\ samples) = 0.5^7 = 0.0078125 = 0.78\% \approx \frac{8}{1000}$$

Rule A – 1 point beyond the +3σ limits

Applicable to both X-bar and s charts.

Two-sided (LCL or UCL): α= ±0.0027

Upper side only (s>LCL): α = ±0.00135

4.2 – Type II Error (β): “Misses Shift”

A type II error occurs when we fail to reject the null hypothesis (H_0) when the alternative hypothesis (H_a) is true. In SPC, that would mean the process has changed, but our tests fail to detect this change because the sample statistics still fall within the control limits.

For the give case:

H_0 = Process centred at 25.05L (desired target)
H_a = Process has shifter, the new average is 25.028L
UCL = 25.089L
LCL = 25.011L
μ under H_a = 25.028
σ under H_a = 0.017L
σ under H_0 = 0.013L

$$zU = \frac{UCL - \mu 1}{\sigma \bar{x}}\ ; zL = \frac{LCL - \mu 1}{\sigma \bar{x}}$$

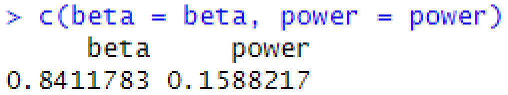
$$\beta = \Phi(zU) - \Phi(zL)$$

$$Power = 1 - \beta$$

$$\beta = \phi(\frac{25.089 - 25.028}{0.017}) - \phi(\frac{25.011 - 25.028}{0.017})$$

$$\beta = 0.841$$

$$Power = 0.159$$



```
> c(beta = beta, power = power)
     beta     power
0.8411783 0.1588217
```

Rule B – 7 consecutive above centre line should be investigated:

Under H_0, each subgroup has a 50% chance to be above the CL

AlphaB = $(0.5)^7$ = 0.0078125

A group with m subgroups has (m-6) distinct windows of length 7.

Chance of at least one violation per week:

$\alpha B = 1 - (1 - 0.5^7)^{m-6} = (m - 6) * 0.5^7$

<u>Rule C – 4 consecutive points above the upper 2-sigma line:</u>

Under H_0, the standardised X-bar: $Z \sim \aleph(0,1)$

Upper 2σ line corresponds to Z > 2

$p = P(Z > 2) = 1 - \Phi(2) \approx 0.0228$

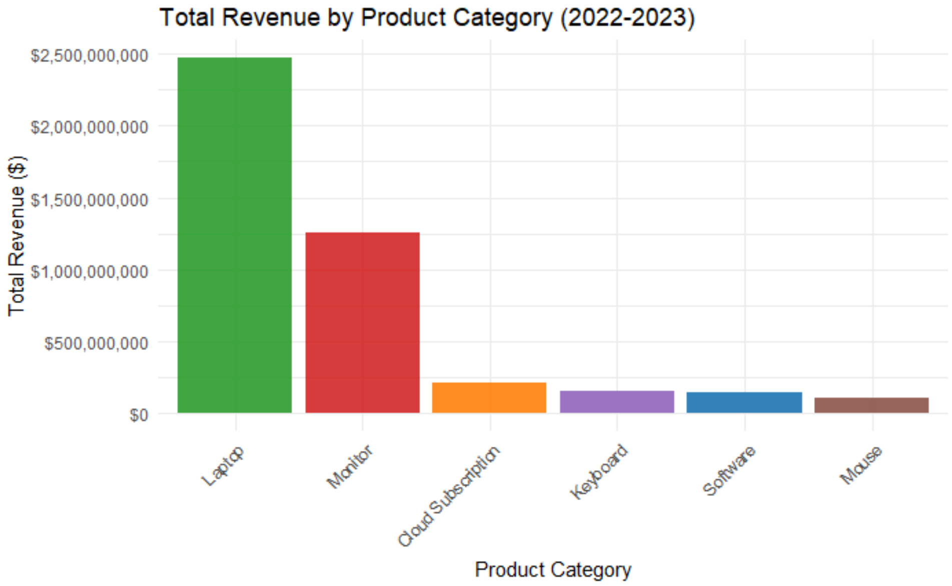
4.3 - Data Correction

We need to update lines 11-60 in the products_Headoffice.csv and products_data.csv where the "NA" prefix should be updated to the correct product type.

From the email we know products 1,11,21,31…etc and 2,12,22,32..etc share the same product details. The code can thus determine the pattern in the first 10 entries, and then apply that to rows 11 - 60.

The programme will extract the prefix of the product group and apply it to the category. From the first 3 characters we can deduce the category (eg. LAP -> Laptop)

Fixing products_Headoffice results in the following Revenue by product category bar graph, with laptop and monitor now accounting for significantly more relative to the other products. Other data analysis stays mostly the same.

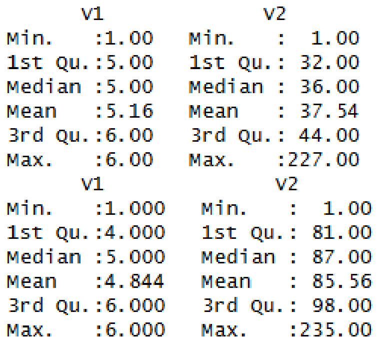


Part 5

5 - Profit Optimisation with timeToServe.csv & timeToServe2.csv

The following analysis will analyse two datasets with the goal of finding the optimal number of baristas a shop should employ. We will data showing service time and number of baristas to determine an answer analytically.

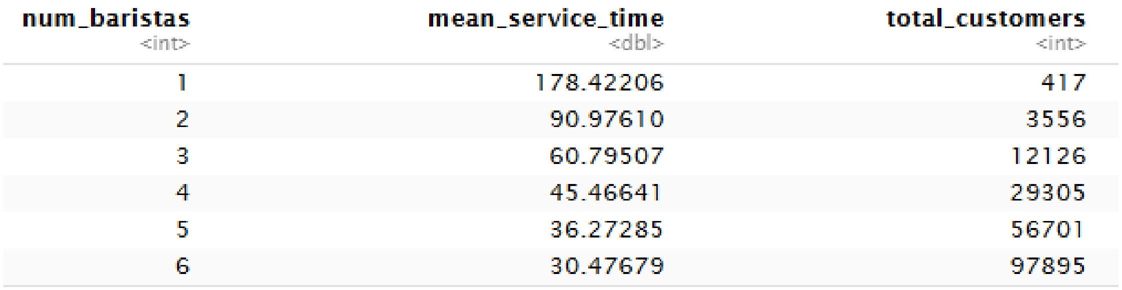
Summary statistics for timeToServe and timeToServe2



```
       V1              V2
 Min.   :1.00   Min.   :  1.00
 1st Qu.:5.00   1st Qu.: 32.00
 Median :5.00   Median : 36.00
 Mean   :5.16   Mean   : 37.54
 3rd Qu.:6.00   3rd Qu.: 44.00
 Max.   :6.00   Max.   :227.00
       V1               V2
 Min.   :1.000   Min.   :  1.00
 1st Qu.:4.000   1st Qu.: 81.00
 Median :5.000   Median : 87.00
 Mean   :4.844   Mean   : 85.56
 3rd Qu.:6.000   3rd Qu.: 98.00
 Max.   :6.000   Max.   :235.00
```

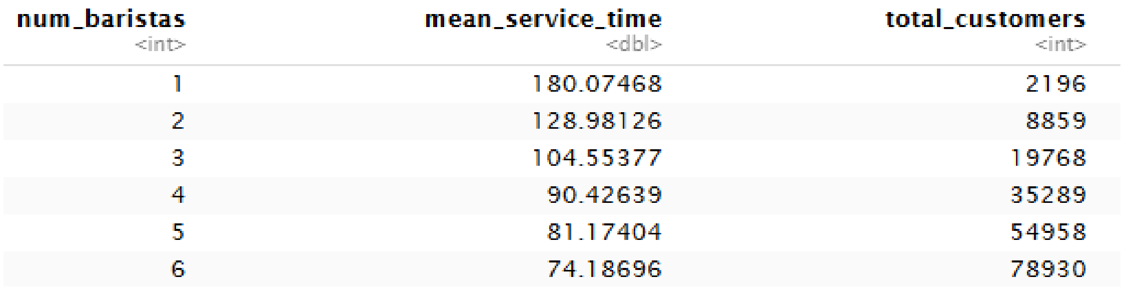
Reliable Service

We set the service time threshold to 5 minutes (300 seconds)



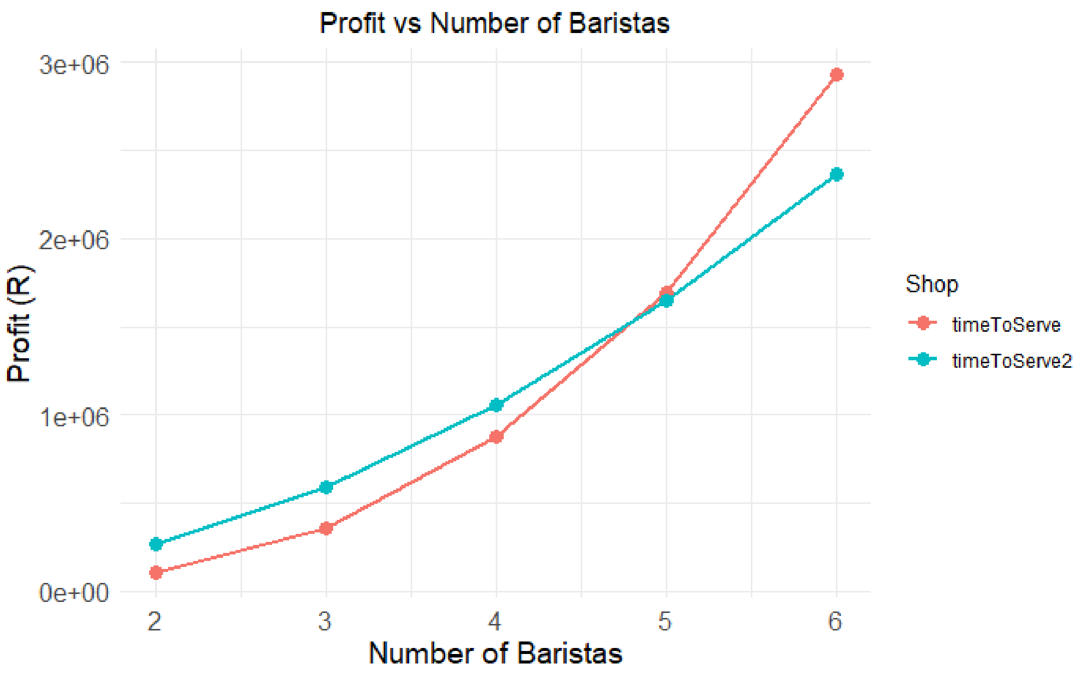
num_baristas	mean_service_time	total_customers
<int>	<dbl>	<int>
1	178.42206	417
2	90.97610	3556
3	60.79507	12126
4	45.46641	29305
5	36.27285	56701
6	30.47679	97895

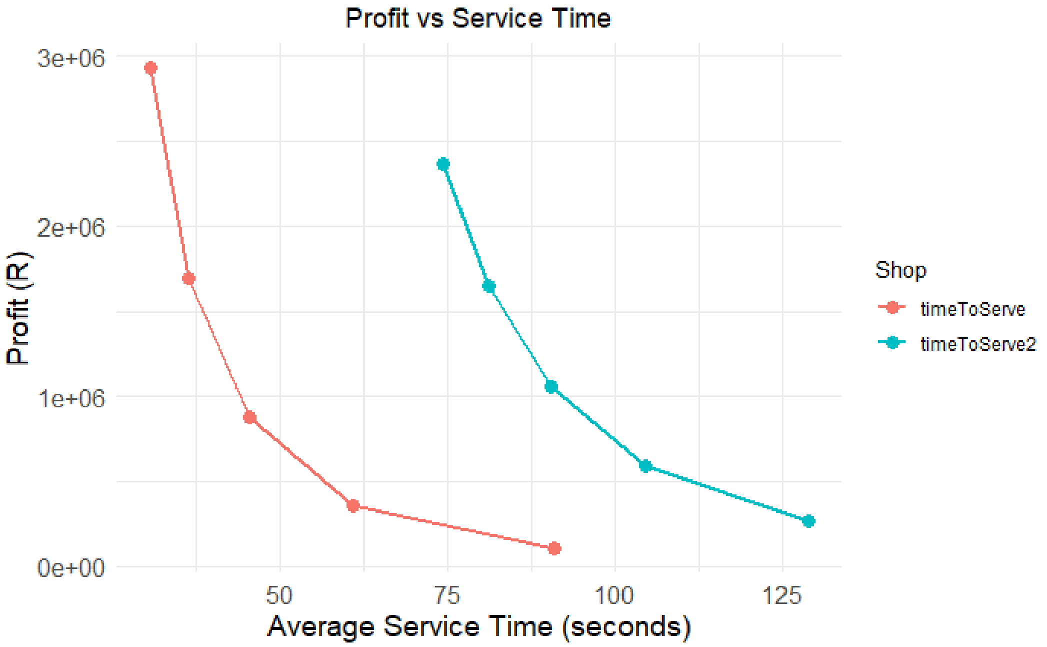
timeToServe - 100% of customers can expect reliable service. More baristas reduces the service time, although the service improvements per additional barista decrease exponentially.



num_baristas	mean_service_time	total_customers
<int>	<dbl>	<int>
1	180.07468	2196
2	128.98126	8859
3	104.55377	19768
4	90.42639	35289
5	81.17404	54958
6	74.18696	78930

timeToServe2 - 100% of customers can expect reliable service. More baristas reduces the service time, although the service improvements per additional barista decrease exponentially.





There is a clear inverse relationship between service time and profit - that being the faster the service (lower service time) the higher the profit. The function is similar to the **Taguchi loss function** in that a minimum loss is achieved. However the trend seems to increase beyond the limits set out, thus we could assume increasing baristas beyond 6 would continue to increase profits.

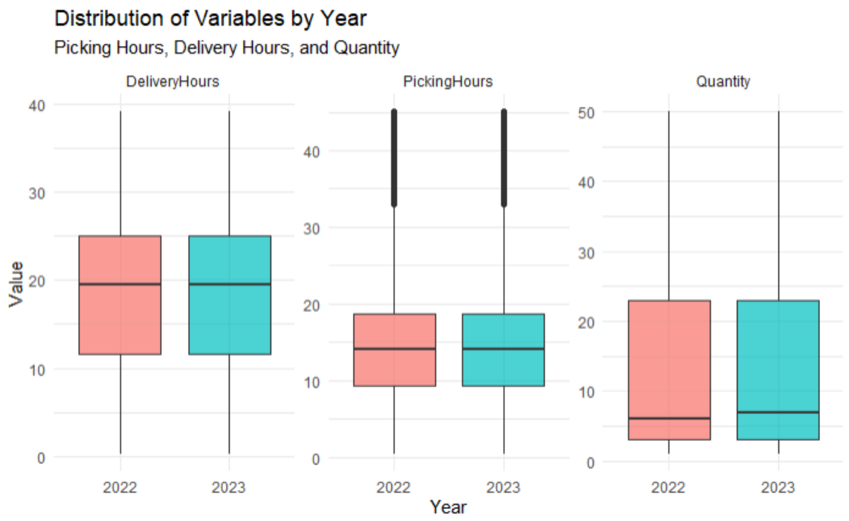
Part 6

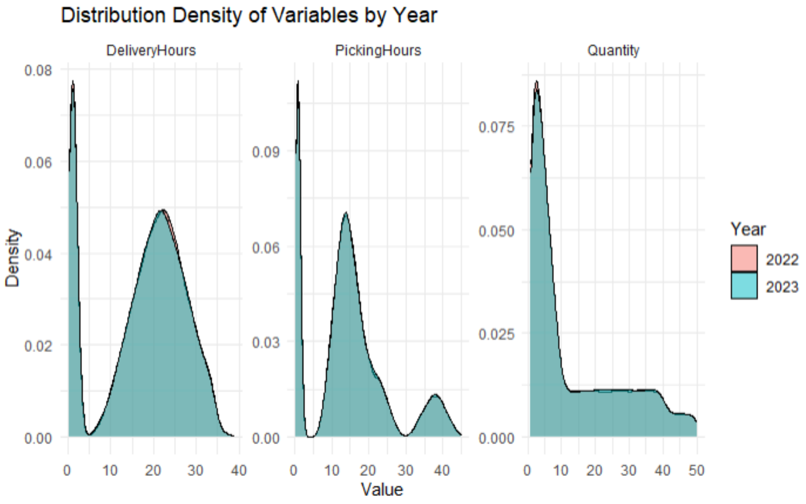
6 - ANOVA Analysis

We will use a MANOVA analysis to test whether the sales data varies dramatically between 2026 and 2027.

<u>Hypothesis:</u> There is significant change in PickingHours, DeliveryHours or Quantity between 2022 and 2023.

Through the R code we are able to generate the following box plot and distribution dentisties.





Note - the 2023 data is overlaid, with small bits of 2022 extruding, we can conclude the data is very similar. In both dataset there is no significant change in the distribution or mean of DeliveryHours, PickingHours or Quantity, thus the hypothesis is rejected.

Part 7

7.1- Estimate the Reliable Service Days

Given that problems are experienced when less than 15 employees are on duty, reliable service is to be expected when there are at leat 15 employees on duty.

No. days with 15 employees on duty: 96

No. days with 16 employees on duty: 270

Days per year with reliable service $= \frac{(270+96)}{397} \times 365 = 336.4987 \approx 336\ days$

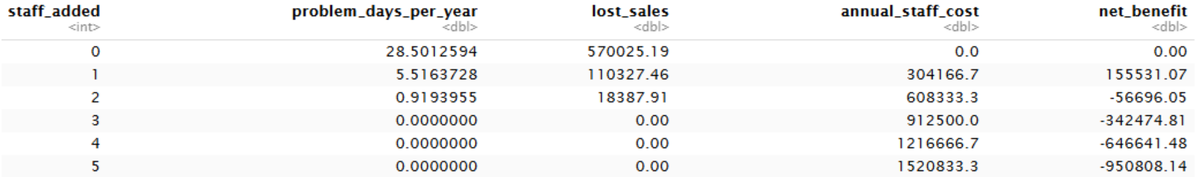
7.2 - Profit Optimization

Revenue lost with < 15 employees on duty: R20,000/day

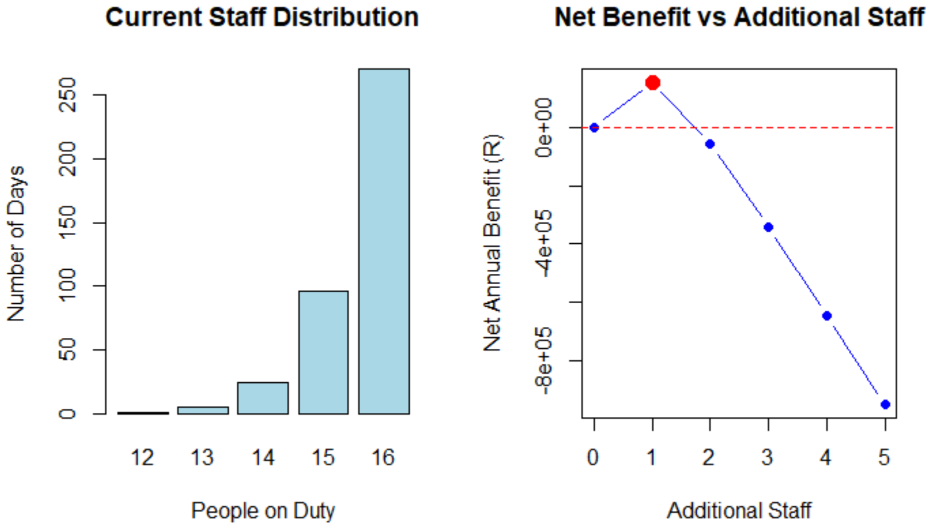
Cost per employee: R25,000/month

The program will find a solution to minimize lost sales as well as hiring costs. In essence, maximize revenue with minimum staff.

1) It will calculate the probability that fewer than 15 staff are present, and thus the lost revenue.
2) We then find how many additional workers are needed per category (3 when 12 are currently present, 2 when 13 are currently present and 1 when 14 are currently present) and the days in which the additional staff are required.. We are then able to calculate the total hiring costs
3) The code then finds when the additional cost of an extra employee outweighs the benefit of the sales revenue, which for this case would mean hiring 1 additional employee.



staff_added	problem_days_per_year	lost_sales	annual_staff_cost	net_benefit
<int>	<dbl>	<dbl>	<dbl>	<dbl>
0	28.5012594	570025.19	0.0	0.00
1	5.5163728	110327.46	304166.7	155531.07
2	0.9193955	18387.91	608333.3	-56696.05
3	0.0000000	0.00	912500.0	-342474.81
4	0.0000000	0.00	1216666.7	-646641.48
5	0.0000000	0.00	1520833.3	-950808.14



The figure above shows the current number of staff on duty which matches what we were given (a confirmation we have imputed the data from the question correctly). To the right is the net benefit calculated per additional staff member employed. It becomes clear that to maximise profit the company should hire one additional staff member.

Conclusion & Final Recommendations

An extensive analysis of the given data we were able to highlight some important trends and considerations of the dataset. The company saw a decrease in sales between 2022 and 2023. Their sales are seasonal, dropping between December and February and again between June and August. Laptops and monitor sales generate the majority of the company's revenue. There is a trend for delivery times to increase beyond control limits throughout the year, and this needs to be addressed. Through ANOVA analysis of the 2022 and 2023 data we conclude there is no significant change in the values studied. This core issue the company must deal with is their failing delivery time, more research into this is needed and should be the top priority for the company to address.

References

Statistical Process Control (SPC) & Control Charts

Montgomery, D.C. (2019) Introduction to Statistical Quality Control. 8th ed. Hoboken, NJ: John Wiley & Sons.

Oakland, J.S. (2018) Statistical Process Control. 7th ed. London: Routledge.

Wheeler, D.J. and Chambers, D.S. (2010) Understanding Statistical Process Control. 3rd ed. Knoxville, TN: SPC Press.

R Programming & qcc Package

Scrucca, L. (2004) 'qcc: an R package for quality control charting and statistical process control', R News, 4(1), pp. 11-17.

R Core Team (2023) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/

Wickham, H. and Grolemund, G. (2016) R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Sebastopol, CA: O'Reilly Media.

MANOVA & ANOVA Analysis

Tabachnick, B.G. and Fidell, L.S. (2019) Using Multivariate Statistics. 7th ed. Boston: Pearson Education.

Field, A., Miles, J. and Field, Z. (2012) Discovering Statistics Using R. London: Sage Publications.

Johnson, R.A. and Wichern, D.W. (2018) Applied Multivariate Statistical Analysis. 6th ed. Upper Saddle River, NJ: Pearson.

Process Capability Analysis

Kotz, S. and Lovelace, C.R. (1998) Process Capability Indices in Theory and Practice. London: Arnold.

Bothe, D.R. (1997) Measuring Process Capability. New York: McGraw-Hill.

Data Visualization

Healy, K. (2018) Data Visualization: A Practical Introduction. Princeton, NJ: Princeton University Press.

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. 2nd ed. New York: Springer-Verlag.

Experimental Design & DOE

Box, G.E.P., Hunter, J.S. and Hunter, W.G. (2005) Statistics for Experimenters: Design, Innovation, and Discovery. 2nd ed. Hoboken, NJ: Wiley.

Montgomery, D.C. (2017) Design and Analysis of Experiments. 9th ed. Hoboken, NJ: John Wiley & Sons.

Online Resources & Documentation

RStudio Team (2023) RStudio: Integrated Development Environment for R. Boston, MA: RStudio, PBC. Available at: http://www.rstudio.com/

STHDA (2023) STHDA: Statistical tools for high-throughput data analysis. Available at: http://www.sthda.com/english/

CRAN (2023) The Comprehensive R Archive Network. Available at: https://cran.r-project.org/

Additional Methodological References

Kassambara, A. (2017) Practical Statistics in R for Comparing Groups: Numerical Variables. STHDA.

Mangiafico, S.S. (2015) An R Companion for the Handbook of Biological Statistics. Available at: rcompanion.org/handbook/