Quality Assurance

ECSA Project – Final Hand-In

Lecturer: Theuns van Schalkwyk

Chris Laubscher

27116441

Date: 24 October 2025

Table of Contents

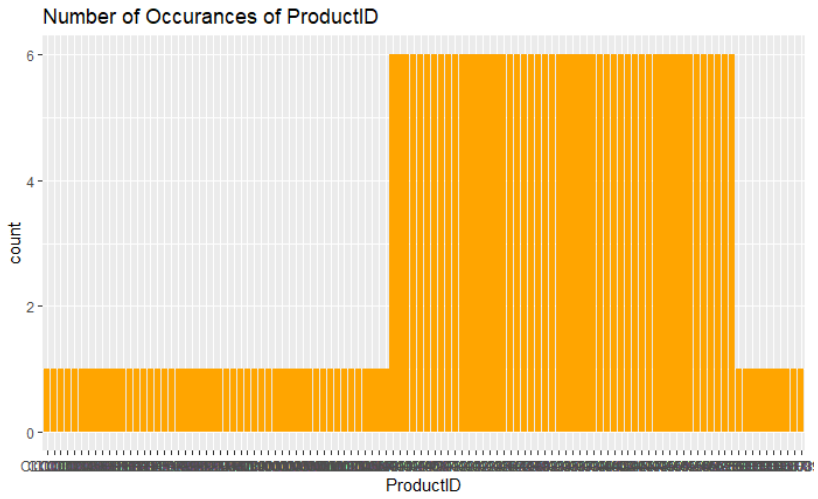
Introduction

The following data explorations consists of visualisations with explanations on conclusions made from them. Four data sets were given to analyse, namely 'Sales', 'Customer_Data', 'Products_Data' and 'Products_HeadOffice'. These data sets were imported into RStudio for analysis and the visualisations were then copied from there.

The data exploration aims to find useful trends and findings that can be used to improve business decisions and move forward as efficiently as possible.

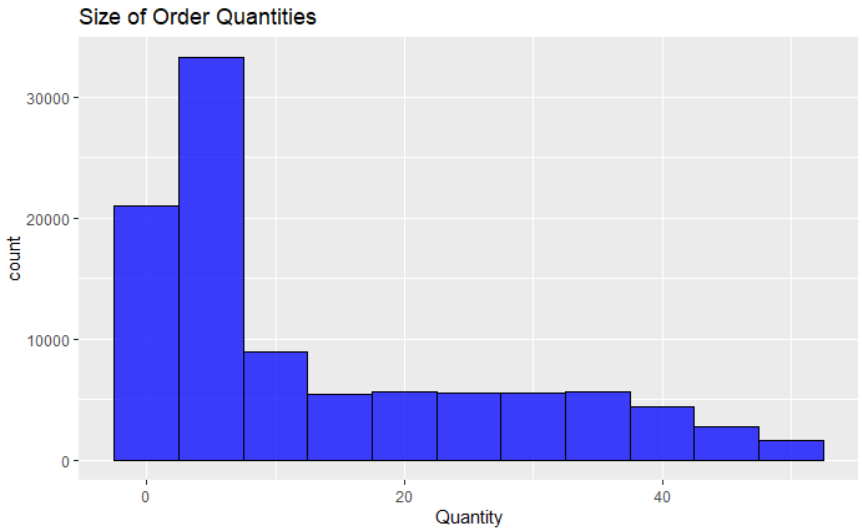
Data Exploration

Data Quality Issues

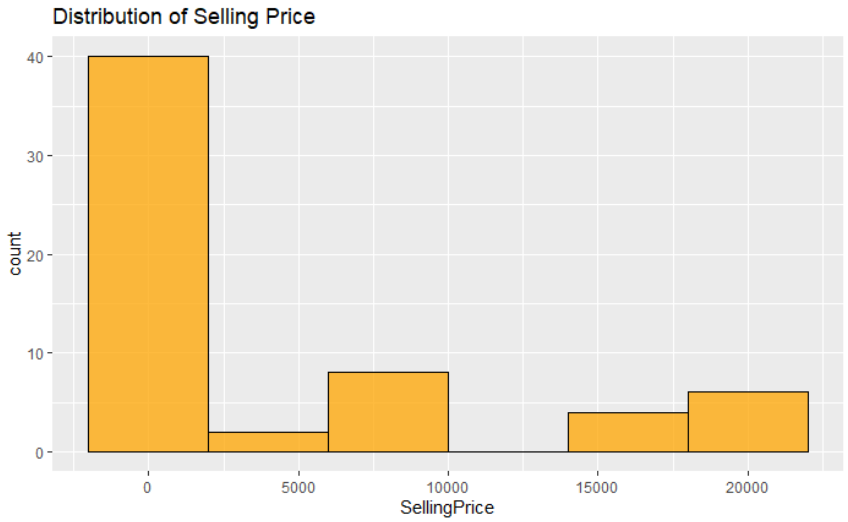


The visualisation on the left shows the number of times each ProductID has occurred in the Products_Data set. Each product should be represented by its own ProductID and thus it is expected that each ProductID only appears once. This can however be seen to not be the case in the data set. This could indicate incorrect data and will have to be audited to ensure correct data analysis on the rest of the data. Potential actions to take include removing the instances who all appear with the same ProductID. This might however cause the data analysis to miss important trends and lose valuable insights into the products. It will be best to investigate those instances further and correct the ProductID codes in the data set.

Sales Patterns

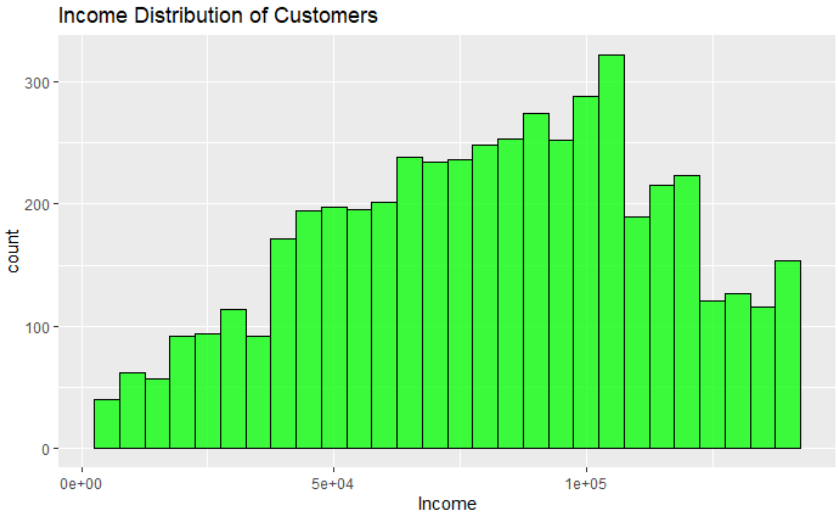


This graph represents the general trend towards the size of the orders. The visualisation shows the count of different order sizes. It was determined and can be roughly seen that the mean order size is 13.5. This value is on the lower side of the order size range compared to order sizes of up to 50.

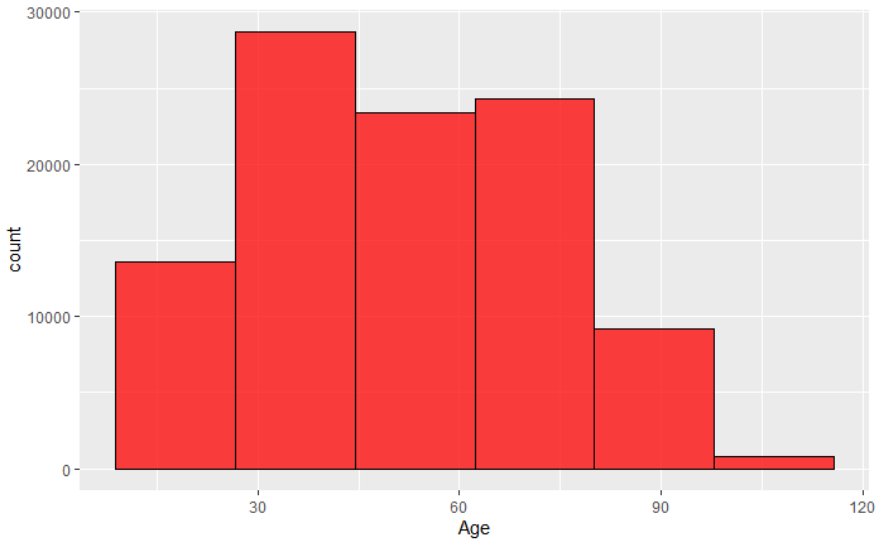


The visualisation on the left shows the number of times products of which value was purchased. There is a very clear peak for selling prices of low value. This implies that the customers most likely prefer to spend less. Investigation will have to be done as to whether customers simply prefer the lower cost items or whether the value aspect of the higher cost items is not clear to the customer. Customers might not see much perceived value in the higher cost items leading to them only purchasing the cheaper options.

Customer Patterns

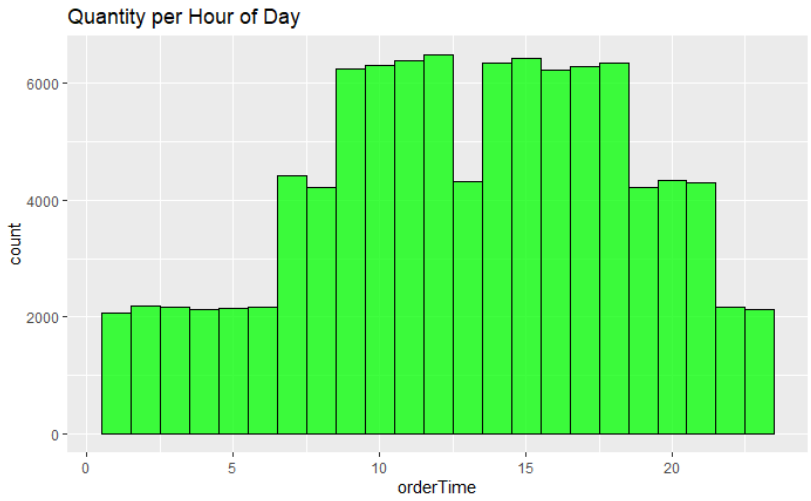


This graph shows the distribution of the income that the customers earn. This will provide insight into the target audience that the company caters for. From the visualisation we cannot see a clear distinction between regular customers' income level and less regular customers' income levels. It is however clear that the middle range income customers purchase the most often. This is emphasised by the graph roughly following a normal distribution. This must be evaluated by management to ensure that the middle-class customer is catered for and perhaps broaden the scope of products to get customers with a larger spread of income levels to purchase more often.

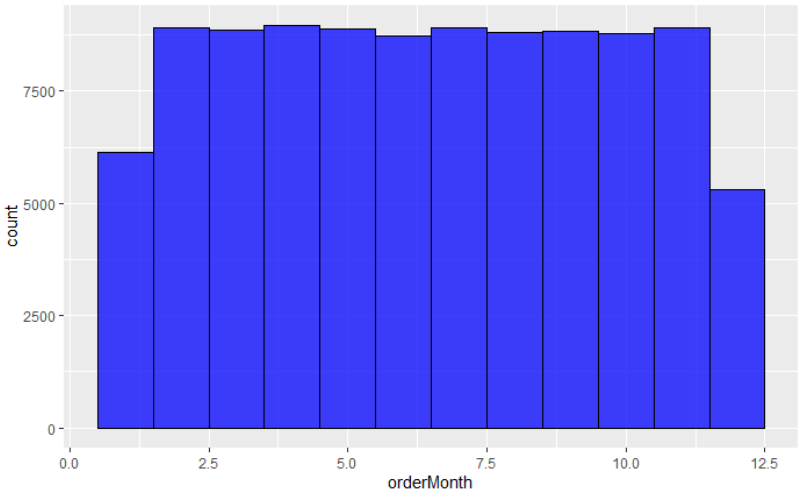


The visualisation on the left shows the distribution in age of the customers. This visualisation is skewed to the right, and therefore we can say that the older ages purchased the least. We see that the most popular age range of customers lies between 30 and 40. After precise calculation, it was determined that the modal age is 31. The distribution is however quite stable implying a good variation in target audience when it comes to age. There are however quite a few customers over the age of 100. These instances might have to audited however, as it might imply incorrect data as not many people reach that age and are not often interested in purchasing technological items.

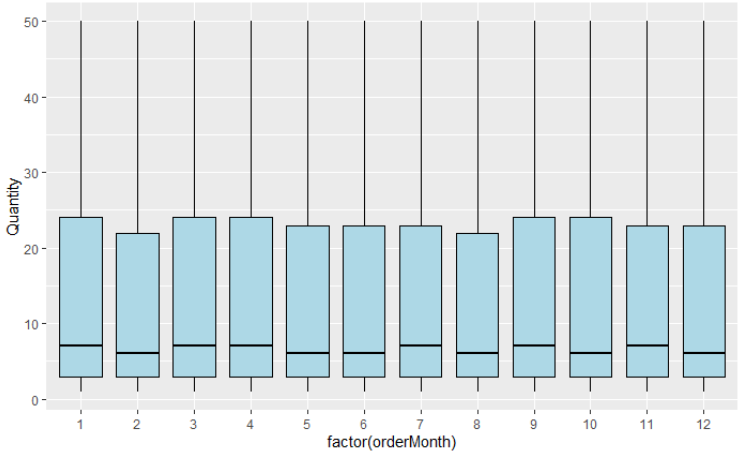
Purchasing Patterns



This visualisation shows the number of sales per time of day which roughly follows a normal distribution. It is clear to see that sales are much lower in the late evening and early morning hours as is to be expected due to most people sleeping during that time. There is also a drop at lunchtime (13h00) which is unexpected due to working people having off time during lunch. There are however no extreme conclusions to be drawn from this visualisation that management should use to drastically change or use to their advantage. The quantities purchased align with what is expected to happen.

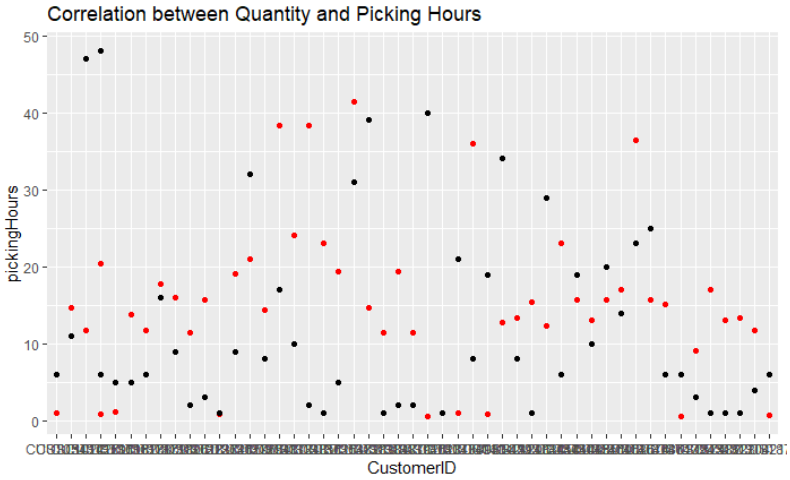


The visualisation on the left shows the distribution of the sales per month. The visualisation follows a uniform distribution, which means that there is not a clear difference between any of the months when it comes to the quantity of orders during that period. It is clear to see that sales are the quietest during the December/January holidays. This is unexpected due to most buying Christmas presents during that time. This data is however useful for demand forecasting to ensure that too much stock is not kept during that time.



The boxplot on the left shows the distribution in order sizes of the different months. There are however no conclusions to be drawn from the boxplot as the quantities appear to be roughly the same per month. The fact that the largest order of the month is 50 for all 12 months indicate that it may be a supplier who consistently orders products.

Other Findings



The visualisation on the left overlays the time spent on picking the orders with the order size (quantity ordered). One would expect a strong correlation since a large order size should relate to a longer picking time. This is however not the case and therefore this might be due to workforce inefficiencies and can be seen by the fact that the visualisation follows a uniform distribution. This is something management must investigate further as a problem like this might lead to overspending on workforce. The precise correlation was determined to be -0.0047. This emphasises that there is absolutely no correlation between Quantity and Picking Hours.

3.1 Control Charts

The following control chart limits were constructed based on the values of the first 30 recorded samples. The following methodology was used:

For x-bar chart → UCL = x_bar_bar + A3*sd_bar

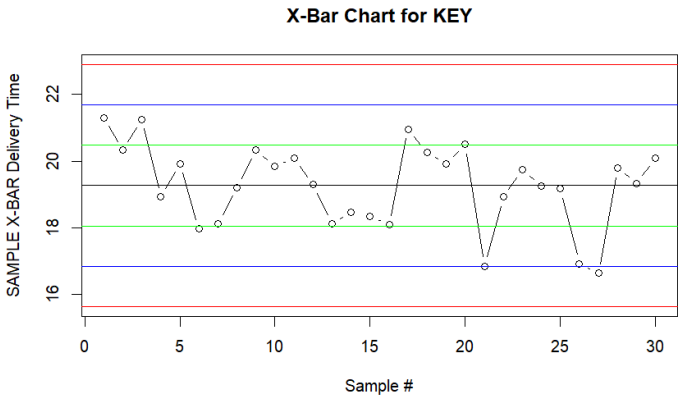
LCL = x_bar_bar - A3*sd_bar

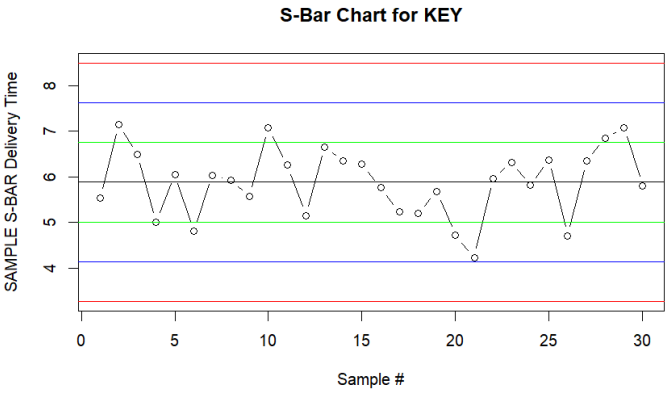
For s-bar chart → UCL = B4*sd_bar

LCL = B3*sd_bar

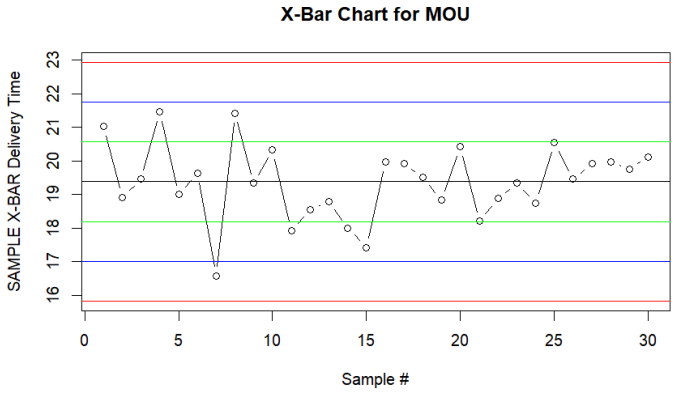
The plotted sample values for the following visualisations also only include the first 30 recorded samples. The delivery process does very well within these sample values as no sample's value exceeds the UCL. These control chart limits will later be used in the assessing of the process in future and help gauge whether the delivery process remains consistent or starts to deviate (QA344 Statistics).

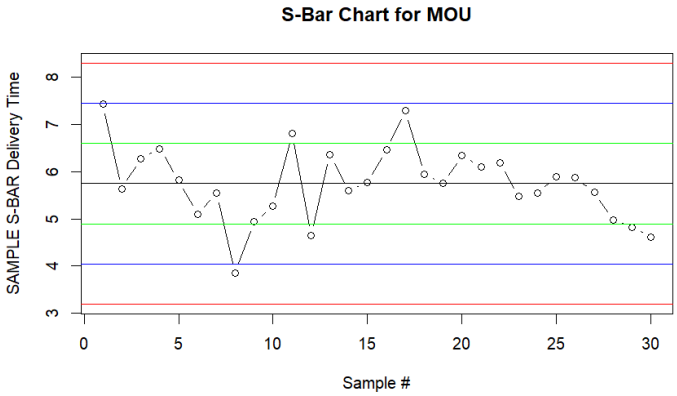
Keyboard



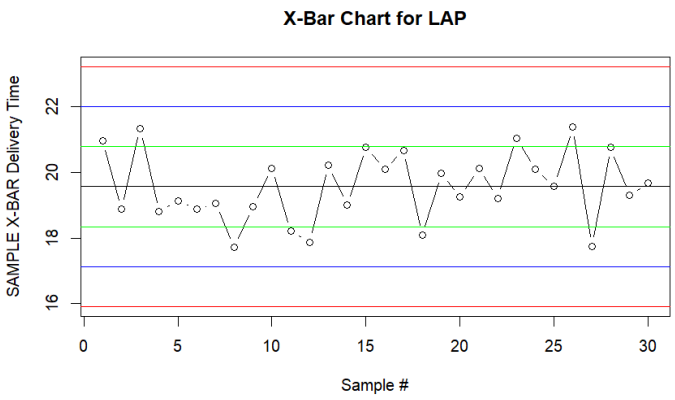


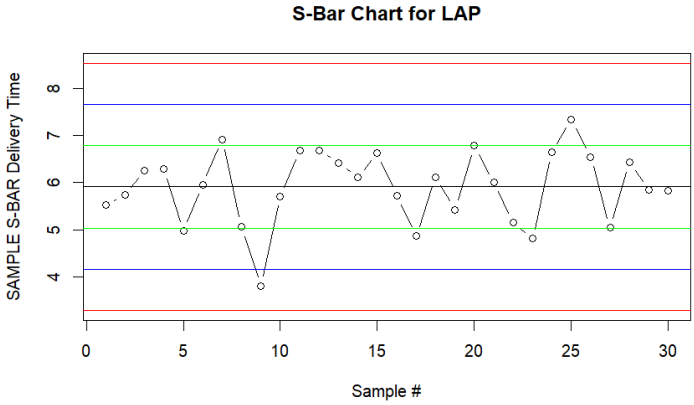
Mouse



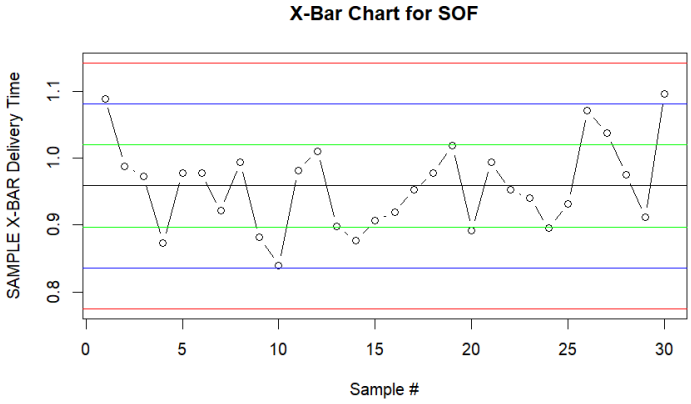


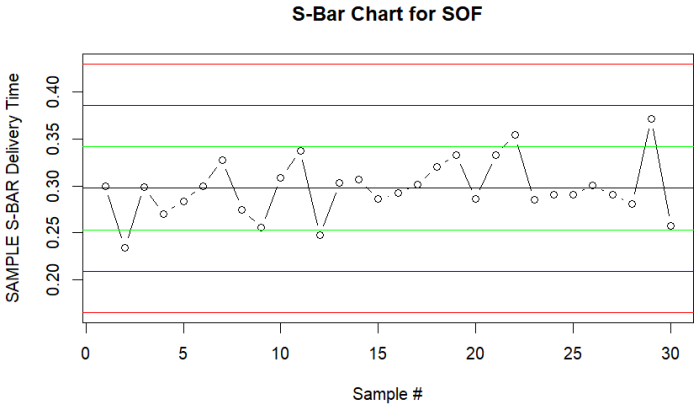
Laptop



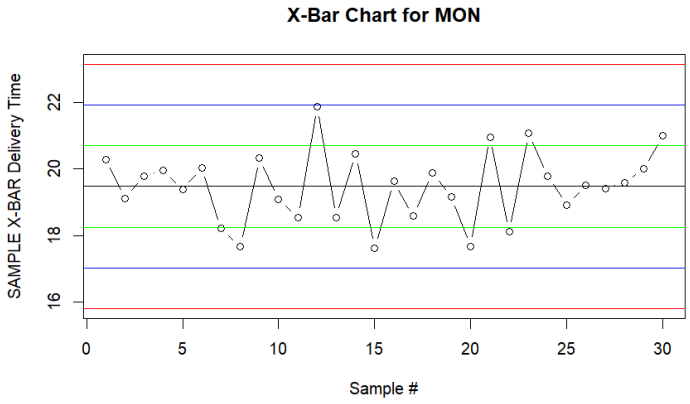


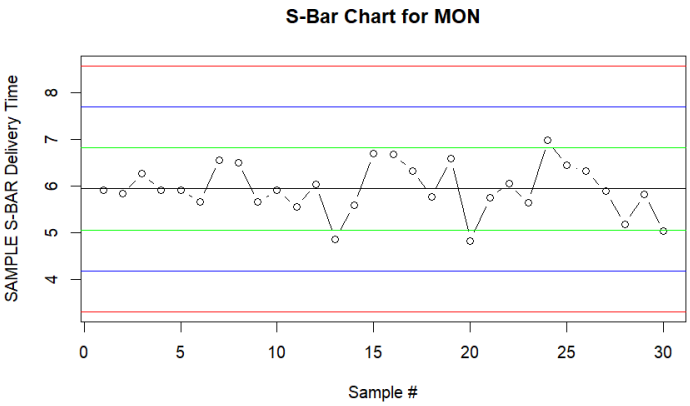
Software



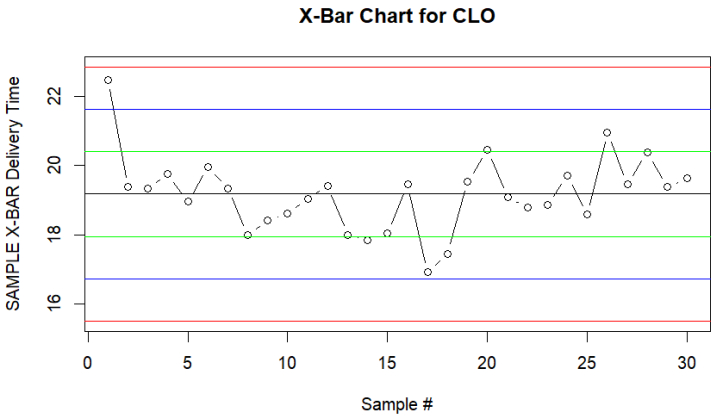


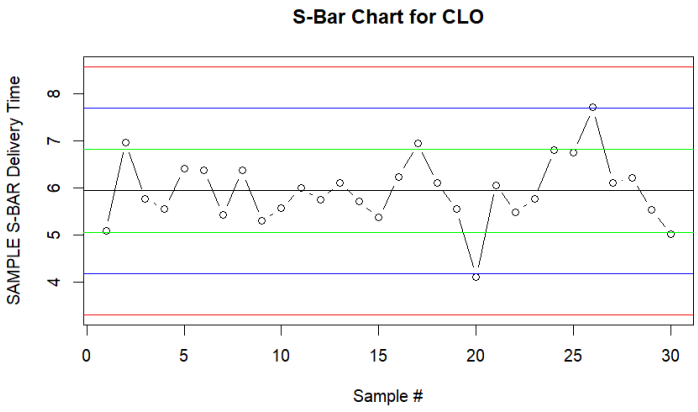
Monitor





Cloud



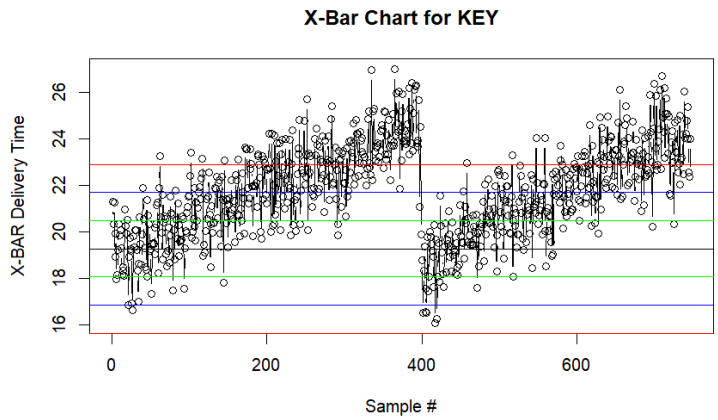


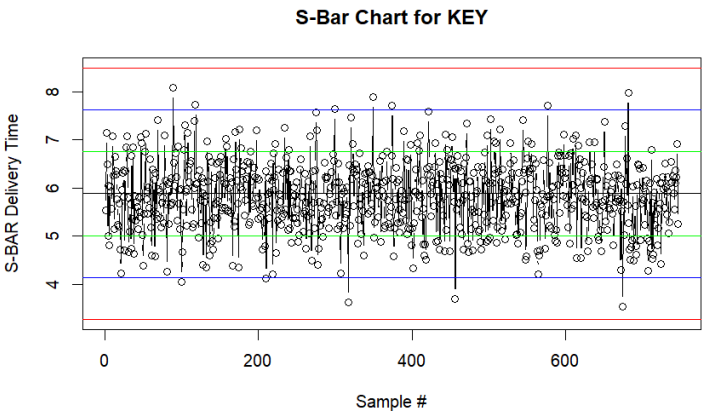
3.2 Process Control

The visualisations below show the average delivery times of samples chosen in sizes of 24. The control chart lines were determined based on the values of the first 30 samples of the recorded delivery times. Each product type has its own x-bar and s-bar chart. The x-bar charts show how the average delivery time per sample changes over time and the s-bar chart shows how the standard deviation value changes per sample over time.

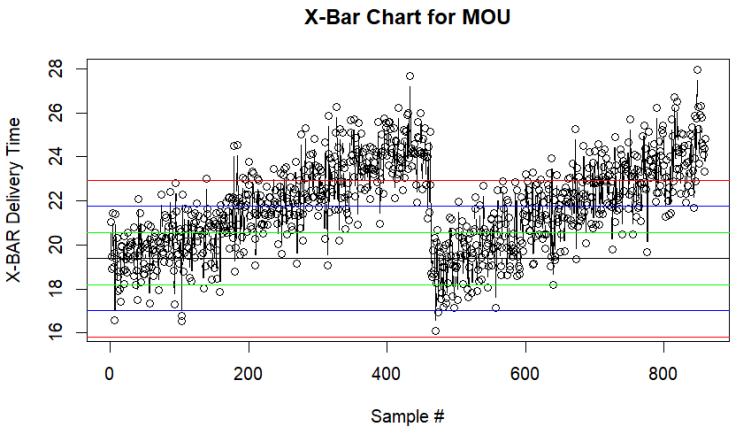
These graphs can then be used to identify problem areas in the delivery process and especially point out at what times the process goes out of specification. This is a crucial piece of information for the principle of continuous improvement (QA344 Statistics).

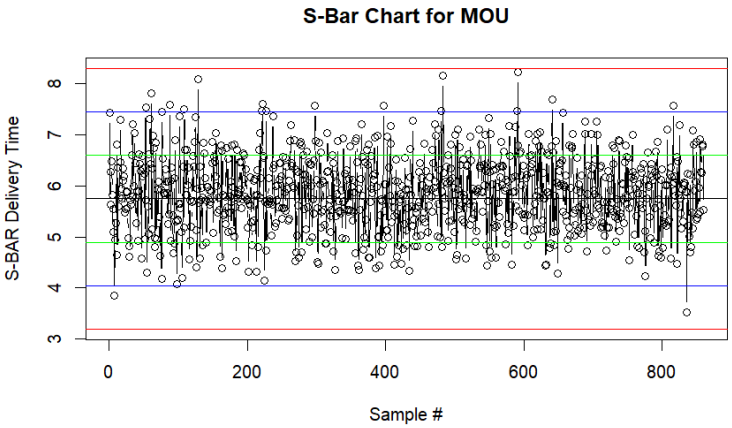
Keyboard



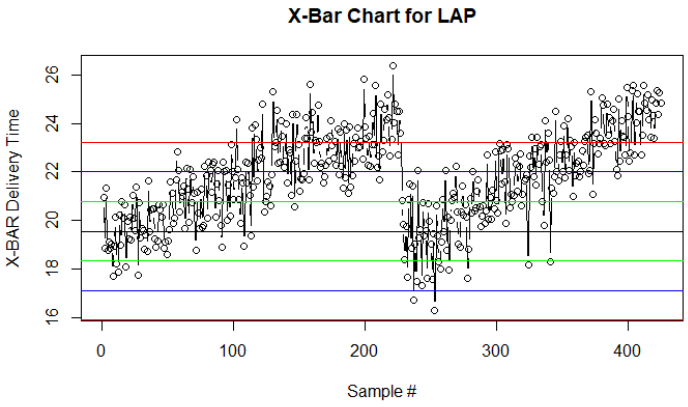


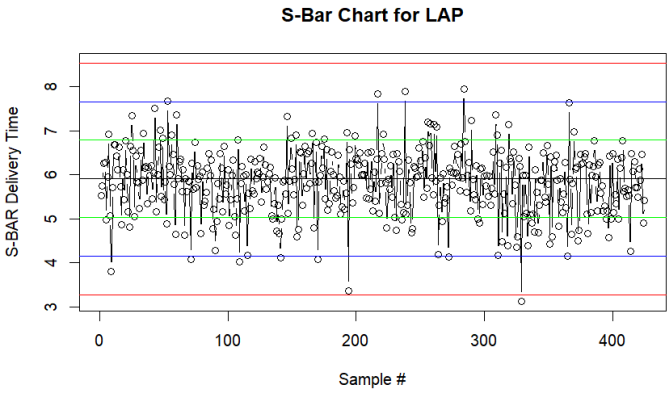
Mouse



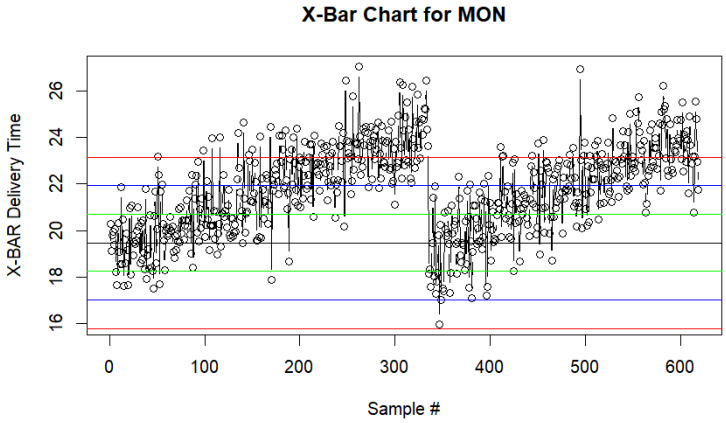


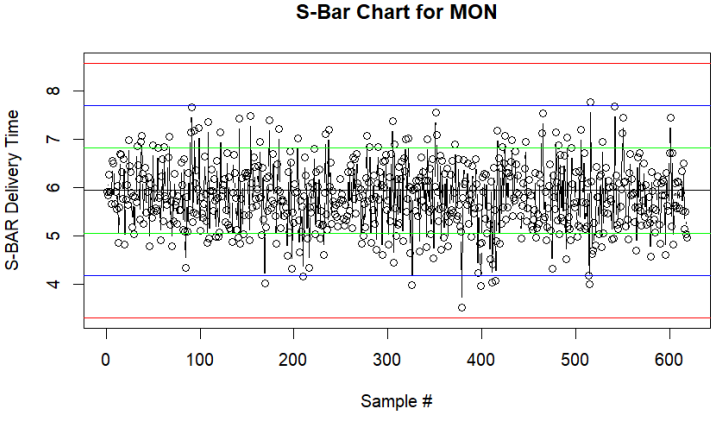
Laptop



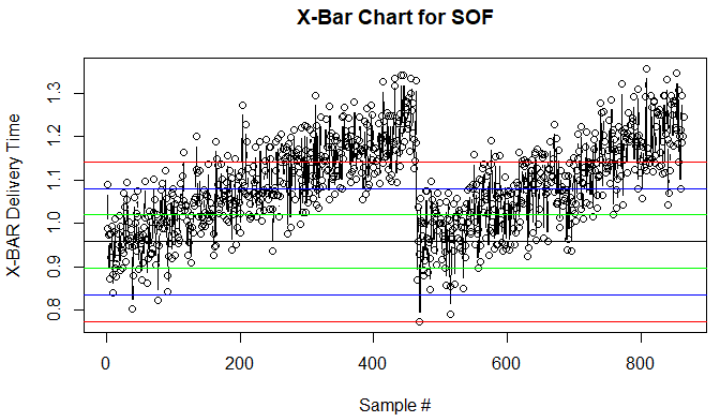


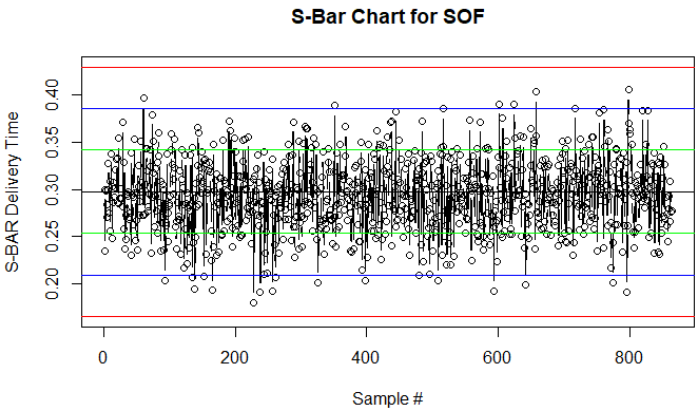
Monitor



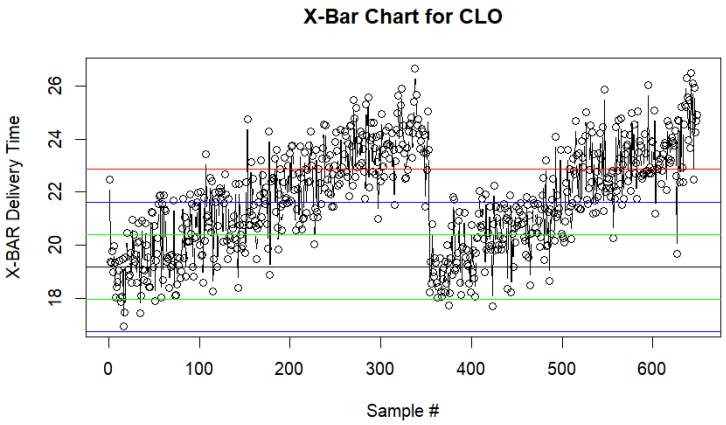


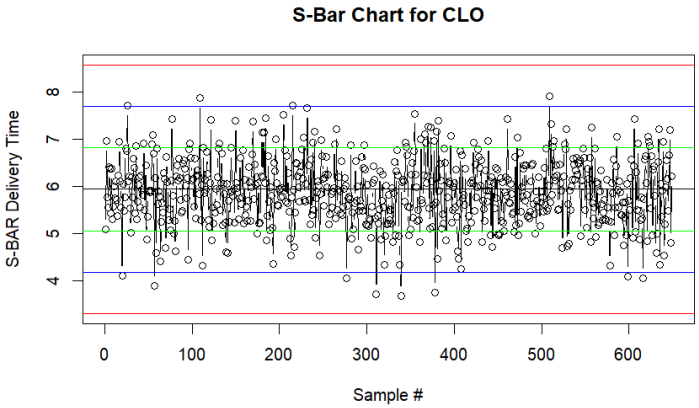
Software





Cloud





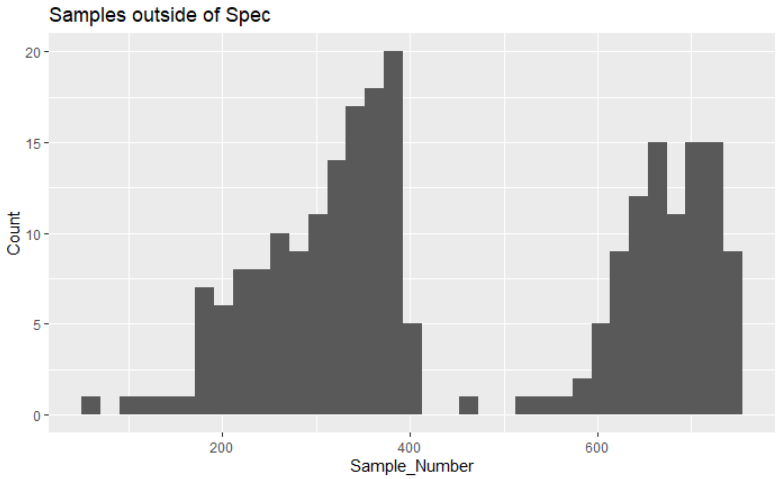
Samples out of Specification

From the below visualisations it can be seen when the process manager is required to adjust or check on the process control. The histograms show bins of sample numbers and then together with that the number of times that those bins of samples fall outside of the specification.

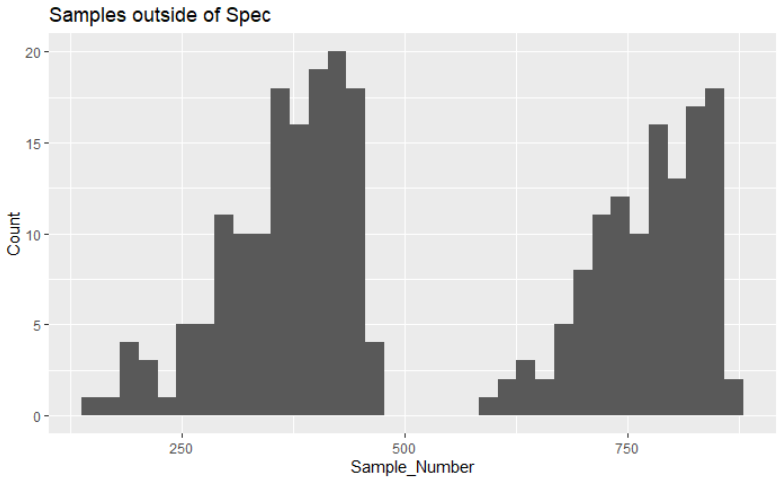
A large count value indicates that the samples associated with that bin fall outside of specification many times after one another. Therefore, adjusting the process at that time is crucial to prevent the process from going out of specification so many times in a row.

All the deliveries first start to deviate from the control chart, then get corrected and then start to deviate again. This could be due to a delivery vehicle starting to deteriorate, a new one then being purchased, and then the new vehicle also starting to deteriorate. From such trends, management should learn when potentially good times will be for predictive maintenance of the vehicles if that were to be the cause of the slower delivery.

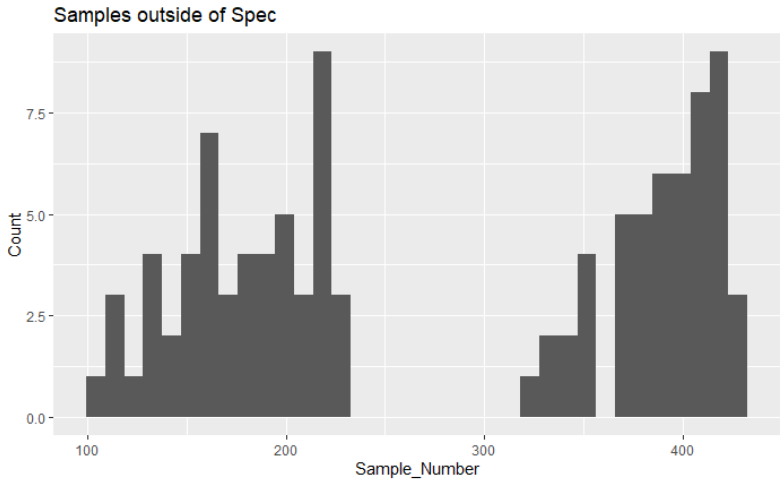
Keyboard



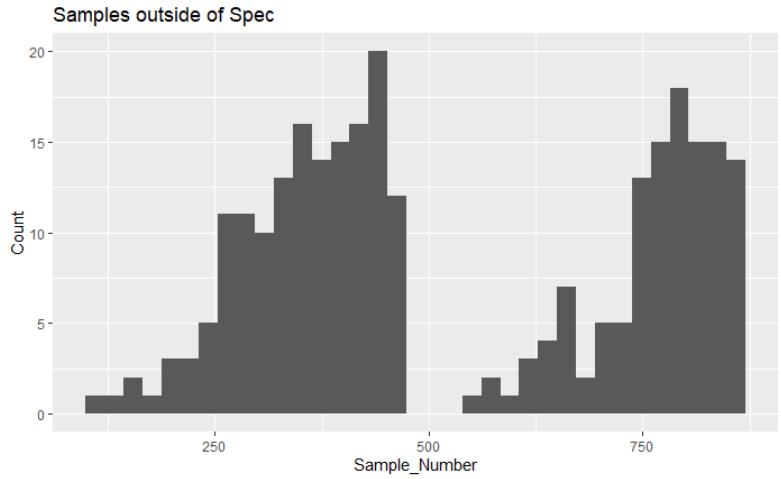
Mouse



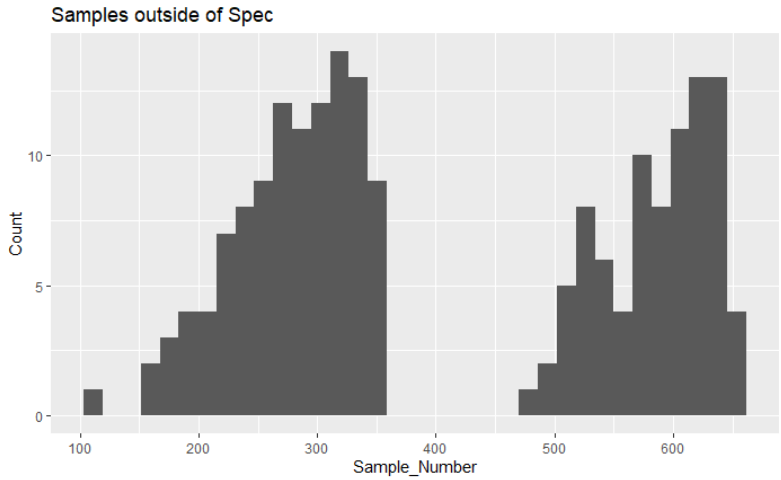
Laptop



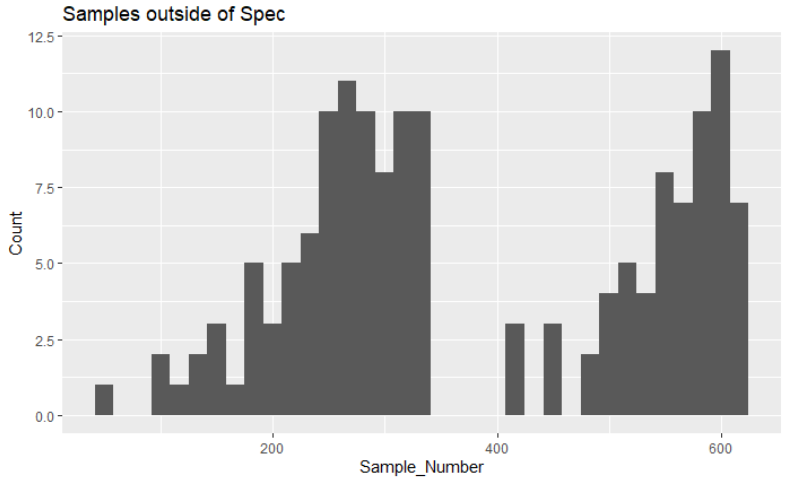
Monitor



Software



Cloud



3.3 Process Capability Indices

Keyboard

Cp	Cpl	Cpu	Cpk
0.9171	1.1049	0.7294	0.7294

Mouse

Cp	Cpl	Cpu	Cpk
0.9152	1.1038	0.7266	0.7266

Laptop

Cp	Cpl	Cpu	Cpk
0.8988	1.1013	0.6962	0.6962

Software

Cp	Cpl	Cpu	Cpk
18.1657	1.0842	35.2473	1.0842

Monitor

Cp	Cpl	Cpu	Cpk
0.8890	1.0785	0.6996	0.6996

Cloud

Cp	Cpl	Cpu	Cpk
0.8977	1.0788	0.7167	0.7167

The above figures comment on the delivery process' ability to stay within the assigned limits for each product type. A LCL and UCL of 0 hours and 32 hours respectively were chosen.

All physical products' Cp values are less than one (< 1) with shows that the delivery times of these products (Keyboard, Mouse, Laptop, Monitor and Cloud) are unable to stay within assigned LCL and UCL. The Cp value of the Software product, however, is many 18.17 which indicates a very definite capability to stay within the assigned limits. Software can be classified as a good product.

Cpk values of less than one (< 1) is seen as uncapable and once again all the physical products fall within this category. This then indicates that the delivery of these products is uncapable of staying within the given bounds. The software delivery Cpk is 1.08. However, due to the delivery process' one sided nature (only late delivery being bad), we must put the focus on the Cpu value of 35.25. This value indicates that the delivery of software is very much the only product that is able to remain within the specified limits and will be able to adhere to the VOC (QA344 Statistics).

3.4 Sample Findings

S – Samples Outside of the Upper Control Limit

There are no instances where the standard deviation of a sample goes beyond the specified limit (UCL). All the samples' standard deviations thus remain consistent with that of the first 30 recorded samples. This shows there are very little variance in the delivery process and shows that deliveries that occur within a short period of time from one another should have similar outcomes. This does however not exclude the possibility of a trend occurring over time that will not be captured by this measure. This means that management can be happy with the process but still take the other control levels in mind and re-evaluate if the UCL3 and LCL3 control limits are correct (strict enough).

Consecutive Samples Between LCL1 and UCL1

Keyboard – For the 'keyboard' product type, the maximum number of samples that fall between the LCL1 and UCL1 control limit, is 20 samples. This range starts at sample #725 and ends at sample #744.

Mouse – For the 'mouse' product type, the maximum number of samples that fall between the LCL1 and UCL1 control limit, is 24 samples. This range starts at sample #238 and ends at sample #261.

Laptop – For the 'laptop' product type, the maximum number of samples that fall between the LCL1 and UCL1 control limit, is 19 samples. This range starts at sample #116 and ends at sample #134.

Monitor – For the 'monitor' product type, the maximum number of samples that fall between the LCL1 and UCL1 control limit, is 34 samples. This range starts at sample #238 and ends at sample #271. This is the longest that a product' delivery time fell within the LCL1 and UCL1 control limits and is therefore indicative of very good control. Management should learn from what they did right within this timeframe and apply that knowledge to the other aspects of the delivery process as well. This will be done by monitoring the process to determine the driver behind this success.

Software – For the 'software' product type, the maximum number of samples that fall between the LCL1 and UCL1 control limit, is 19 samples. This range starts at sample #538 and ends at sample #556.

Cloud – For the 'cloud' product type, the maximum number of samples that fall between the LCL1 and UCL1 control limit, is 19 samples. This range starts at sample #151 and ends at sample #169.

Consecutive X-Bar Samples Outside Upper, Second Control Limit

There were many times that more than 4 samples were above the UCL2. These samples reflect on poor delivery since those samples fall above the required time limit that is considered as fine. For simplification purposes, only the first 3 and last 3 instances of 4 or more samples exceeding the UCL2 will be mentioned below for each product type.

Keyboard

Instance	First Sample Number	Streak Length
1	172	4
2	178	4
3	183	4
23	698	21
24	721	4
25	726	21

Mouse

Instance	First Sample Number	Streak Length
1	236	4
2	280	7
3	288	12
22	803	5
23	811	32
24	844	17

Laptop

Instance	First Sample Number	Streak Length
1	119	4
2	130	11
3	158	10
11	364	6
12	374	18
13	393	33

Monitor

Instance	First Sample Number	Streak Length
1	173	5
2	179	4
3	190	5
21	580	29
22	610	4
23	615	4

Software

Instance	First Sample Number	Streak Length
1	237	4
2	260	11
3	278	4
23	803	38
24	843	16
25	860	5

Cloud

Instance	First Sample Number	Streak Length
1	122	4
2	179	5
3	193	4
17	567	36
18	604	23
19	628	22

For explanation purposes, for example, look at the Keyboard product class table. The table says that there were 25 times that 4 of more samples went above UCL2. The first 3 and last 3 of these instances are elaborated in the 2 columns to its right. The 'First Sample Number' column gives the index of the sample where the exceeding streak starts and the 'Streak Length' column gives the length for which the streak continues for (in samples).

The fact that so many of these streaks exist is testimony to the poor quality of the delivery system, but on the positive side of things allow for lots of optimisation in the process. Laptop delivery has the least number of these instances (13) and keyboard, and software delivery is tied for having the largest number of these instances (25).

These consecutive samples that fall outside of the control limits should be addressed by management. This will include monitoring those activities to find the exact reason behind this poor service.

4.1 Estimating Type 1 Errors

A

This is a one-sided probability since it only looks at the probability of the standard deviation going above the UCL3 line.

Probability of Type 1 Error for A is **0.00135**

B

This is a two-sided probability since it looks at the probability of the standard deviation falling between the LCL1 and UCL1 lines.

Probability of Type 1 Error for B is **0.682689**

C

This is a one-sided probability since it only looks at the probability of the sample means going above the UCL2 line in a consecutive manner.

Probability of Type 1 Error for C is **2.67877e-07**

4.2 Estimating Type 2 Errors

Methodology: The probability of making a type 2 error can be defined to be the probability of thinking that the process runs within limits (is controlled), but it is not running correctly. This is possible to determine since we are given that the process mean and standard deviation has changed without management being aware of this. Thus, a type 2 error will be made when the filling value falls within the old LCL and UCL, and outside of the UCL or LCL of the new process values. The probability area where those two cases overlap, will be equal to the probability of making a type 2 error (Bhandari, 2023).

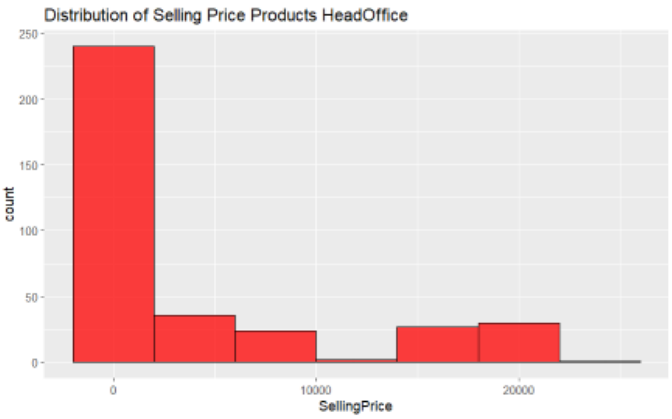
The z values can be determined as follows: $Z_1 = \frac{UCL-mean}{sd}$ and $Z_2 = \frac{LCL-mean}{sd}$

The probability is then determined by: $P(Z \leq Z_1) - P(Z \leq Z_2)$

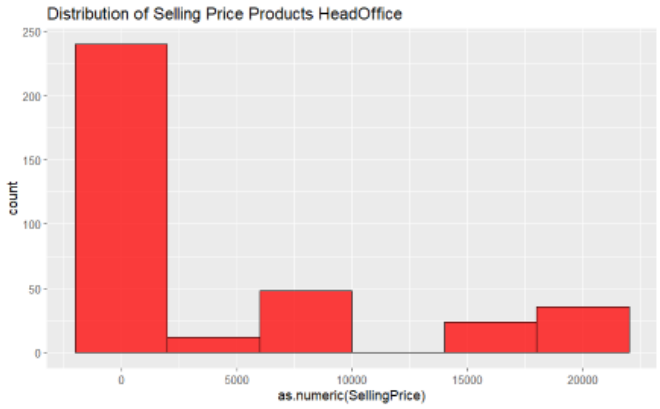
The probability of making a type 2 error is then **0.841178**.

4.3 Data Analysis on Corrected Data

In the previous analysis there was a difference in the distributions of the selling price between the products_data and the products_HeadOffice data. This difference can clearly be seen in the visualisations below.

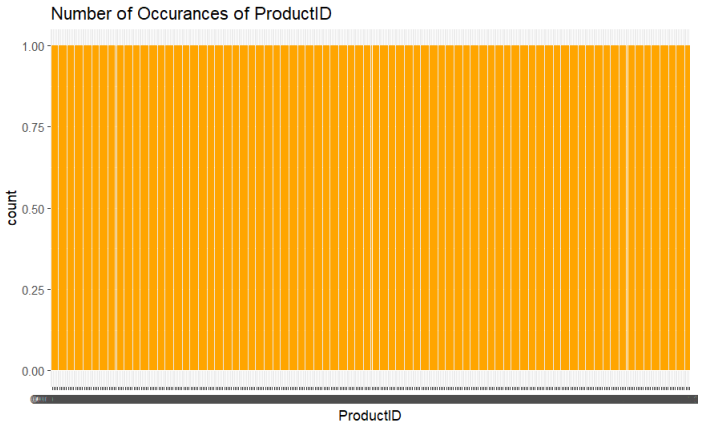


It was found that there were indeed data quality issues in the data. After identifying them and fixing them in the data sheet, it was found fit to redo the analysis on the updated data. The distribution of the selling price of the products_data and the products_Head_Office data sheet is compared below. It is worthwhile to note that the two graphs now look identical. This is due to the error correcting that was conducted in the data sheets.



The visualisation shows that the primary products sold are on the less expensive end of the products available. The visualisation can roughly be described as skewed to the right. This should lead management to ensure that their target audience aligns with the regular customers. This can also allow them insight into which products require more marketing or which products should perhaps be discontinued.

During the previous analysis, data quality issues were suspected due to the irregular appearance of the ProductIDs. Most ProductIDs only occurred once as they should, but some of them appeared six times. After the error correcting process, this plot is now displayed again to show the improvements to the data. It can be seen below that each ProductID now only appears once. This indicates correct data and thus implies more accurate data analysis.



The sales values during the year of 2023 were analysed individually as well and the following numbers were obtained:

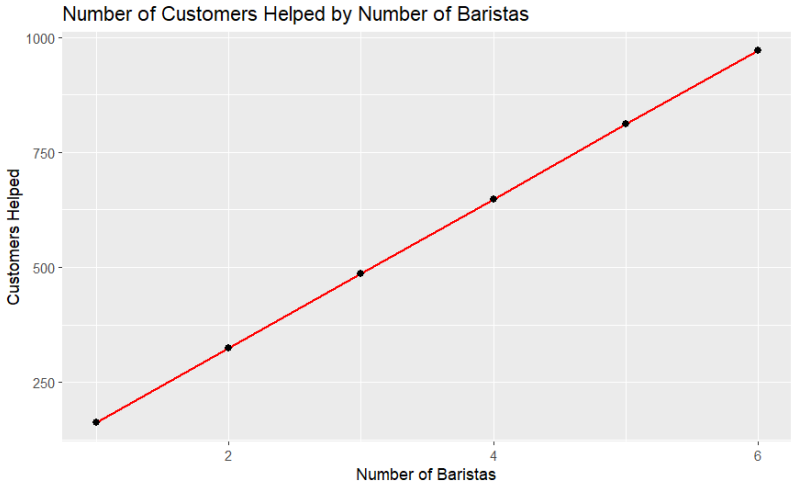
Total Sales During 2023

Product Type	Total Sales Amount [R]	Mean Selling Price [R]
Keyboard	5 378 598.87	644.66
Mouse	3 773 413.87	394.70
Laptop	86 027 413.33	18 086.43
Monitor	43 126 707.90	6 310.53
Software	7 261 887.10	1 019.06
Cloud	4 867 780.65	506.18

The largest amount of income came from Laptop sales, whilst the smallest amount was obtained from Mouse sales. When comparing the Mouse sales amount with its mean selling prices however, there were many sales of this product class since the mean selling price is the lowest out of all the product types. Laptops, on the other hand, has the highest mean selling price and thus explains why its total sales amount might be the lowest. This is an important realisation, since only looking at sales for analysing product success can lead to very deceiving results.

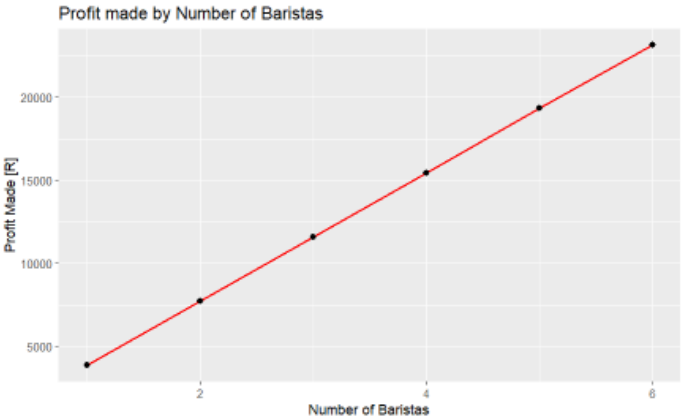
5. Profit Optimisation (Dataset 1)

It is required to maximising the profit that a coffee shop makes by determining the optimal number of baristas that should be hired. This involved calculations for each possible number of baristas that may be hired and then analysing the results and interpreting them accurately. The visualisations below aim to guide in the logic and provide a better understanding of the result that follow thereafter.



The number of customers that the coffee shop can help is directly proportionate to the number of baristas on duty. This can be seen by the linear correlation followed by the graph. The most customers can thus be helped when the number of baristas is at its maximum, that is with six baristas on duty.

The amount of profit made is also directly proportionate to the number of baristas on duty. Thus, a similar result is obtained for the optimal number of baristas that should be hired. A profit of R23 140.49 is made when six baristas are on duty.



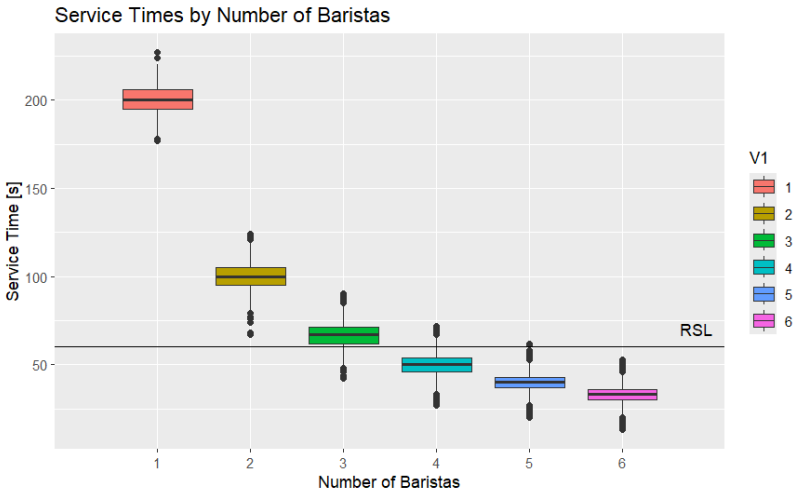
Profit per Number of Baristas on Duty [R]

1	**3856.22**
2	**7703.41**
3	**11 592.02**
4	**15 447.63**
5	**19 323.21**
6	**23 140.49**

It is thus clear to see that **6 baristas** must be hired for the optimal profit of **R23 140.49**. Since the optimal number of baristas is the maximum number allowed the possibility of increasing on this number should be investigated. This problem does not correspond to that of a Taguchi loss function due to its lack of symmetry. A Taguchi loss function looks at the effect of a variable

deviating from its target value. Deviations in a Taguchi loss function on the lower and upper side of the ideal variable value leads to similar results (gain or loss), but in the case of this problem, hiring more and less than the optimal number of baristas lead to different results in profit, respectively. Due to the result's linear characteristic, hiring less than the ideal number, will result in a loss in profit and hiring more than the ideal number of baristas, if it was possible, would lead to an increase in profit according to the linear model.

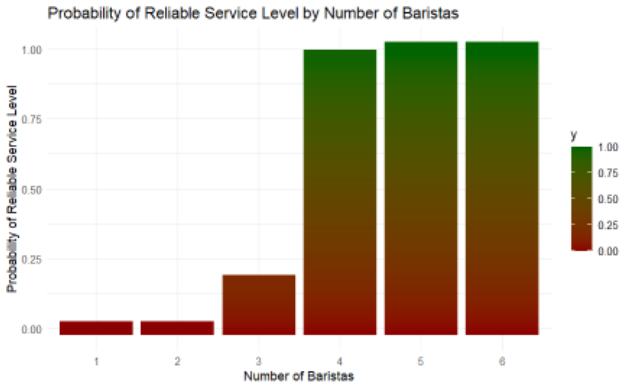
It is useful to estimate the probability that all the customers will be helped within a reasonable time. This will help to estimate the customer satisfaction, which is a key factor for successful companies. The visualisation below shows the distribution of the service times that were achieved with the various number of baristas that may be appointed. Box Plots were chosen to visualise this, as it allows one to see the spread of the service times and not only the average achieved. The reliable service level (RSL) was estimated at **60 seconds**.



From the above visualisation, the probabilities of service quality can now be determined. The assumption was made that a reliable service time is service under 60 seconds. Thus, the methodology behind the probability calculations is using the past data that is available and calculating the percentage of times that customers were helped within 60 seconds per number of baristas present. This percentage can then be used as a probability for future customer service estimates. This method is based on the concept that the past can be a representation of the future. This might not be possible for this specific scenario, so the result should only be seen as an estimate and service level should still be monitored constantly in practise. This should only provide advise on the amount of baristas to hire and not as a guarantee of good service level.

Probability of Reliable Service by Number of Baristas

1	**0**
2	**0**
3	**0.16461**
4	**0.97229**
5	**0.99996**
6	**1**

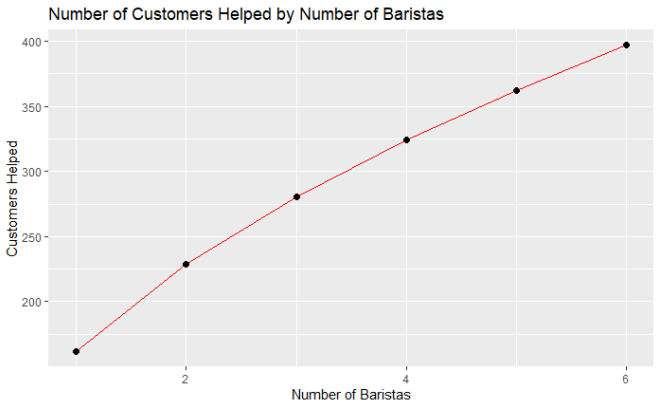


From the above table it is clear to conclude that at least **4 baristas** should be hired in order to achieve a high probability of reliable customer service. This number of baristas hired may be lowered if a lower level of reliability becomes acceptable, but under current circumstances at least four baristas will have to be hired.

In the case of the optimal number of baristas being hired, 100% reliable service level can be promised. This is a wonderful figure to be able to show customers and will bring in even more sales.

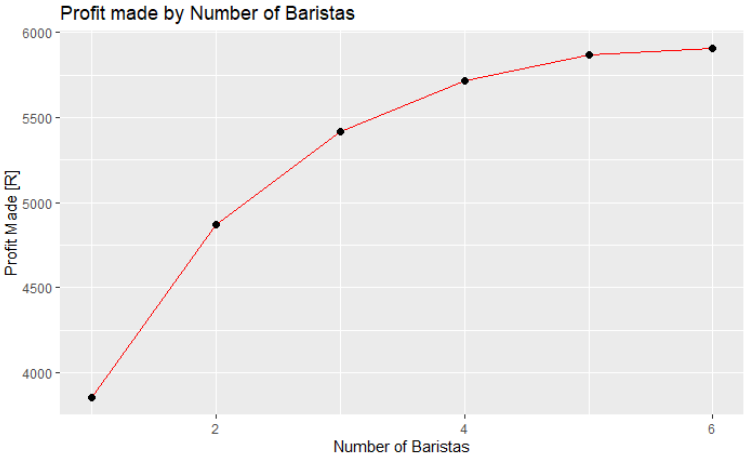
Profit Optimisation (Dataset 2)

It is required to maximising the profit that a coffee shop makes by determining the optimal number of baristas that should be hired. This involved calculations for each possible number of baristas that may be hired and then analysing the results and interpreting them accurately. The visualisations below aim to guide in the logic and provide a better understanding of the result that follow thereafter.



The number of customers that the coffee shop can help is directly proportionate to the number of baristas on duty. This can be seen by the linear correlation followed by the graph. The most customers can thus be helped when the number of baristas is at its maximum, that is with six baristas on duty.

The amount of profit made is also directly proportionate to the number of baristas on duty. Thus, a similar result is obtained for the optimal number of baristas that should be hired. A profit of R23 140.49 is made when six baristas are on duty.

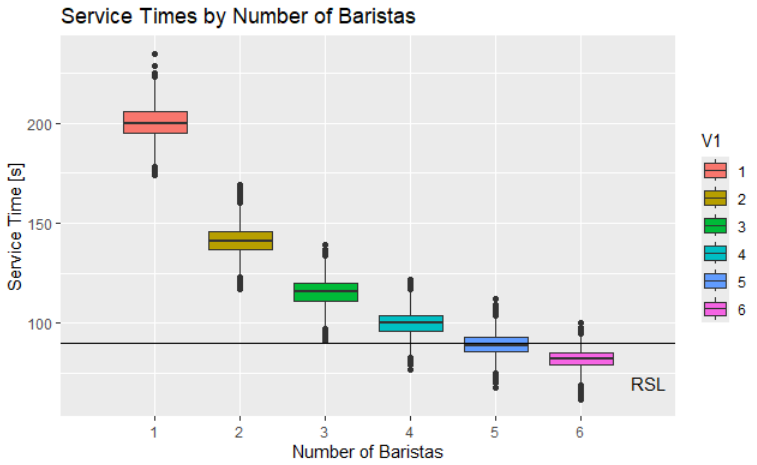


Profit per Number of Baristas on Duty [R]

1	**3855.90**
2	**4868.55**
3	**5419.89**
4	**5718.52**
5	**5868.11**
6	**5905.53**

Profit-wise, it is the best to employ **6 baristas** as the profit (**R5905.53**) is the most for that number. The profit does however start to plateau at six baristas, so it can be assumed that having more than six will result in a reduced profit. This is however above the maximum allowed number of baristas, but nonetheless important to note and may warn that having six baristas is already a stretch on space capacity within the coffee shop.

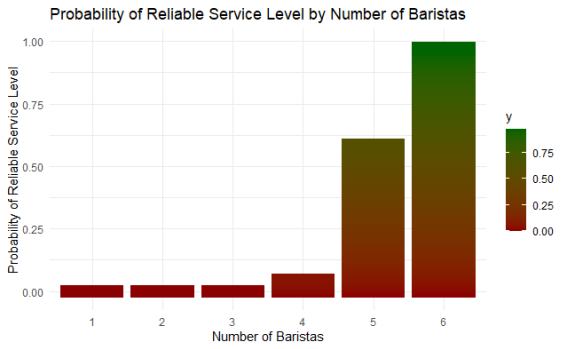
It is useful to estimate the probability that all the customers will be helped within a reasonable time. This will help to estimate the customer satisfaction, which is a key factor for successful companies. The visualisation below shows the distribution of the service times that were achieved with the various number of baristas that may be appointed. Box Plots were chosen to visualise this, as it allows one to see the spread of the service times and not only the average achieved. The reliable service level (RSL) was estimated at 90 seconds.



From the above visualisation, the probabilities of service quality can now be determined. The assumption was made that a reliable service time is service under **90 seconds**. Thus, the methodology behind the probability calculations is using the past data that is available and calculating the percentage of times that customers were helped within 90 seconds per number of baristas present. This percentage can then be used as a probability for future customer service estimates.

Probability of Reliable Service by Number of Baristas

1	**0**
2	**0**
3	**0**
4	**0.0455099**
5	**0.5842098**
6	**0.9744077**



From the above table it is clear to see that at least 5 baristas need to be hired to achieve a 58% probability of providing reliable service. For a reliability percentage of 97%, which is much advised, **6 baristas** need to be hired. Since this number of baristas correspond to the optimal number of baristas, this is a wonderful figure to be able to show customers and will bring in even more sales.

6. ANOVA on Sales Data

After re-analysis of the results from part 3, the following hypotheses were chosen to be tested. The difference between the delivery times between the years 2022 and 2023 will be tested for each of the product types. One-way ANOVAs will be performed with the result of 6 ANOVAs being tested. The results of the ANOVAs are shown and explained below (Kenton, 2025).

Keyboard

```
               Df Sum Sq Mean Sq F value  Pr(>F)
orderYear       1    302  302.42   8.088 0.00446 **
Residuals   17918 669951   37.39
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA outputs a p value of 0.00446 and a F value of 8.088. This shows a statistically significant effect of the ‘orderYear’, but also that it explains almost none of the variation. Since the p value is so small, we reject the null hypothesis and conclude that orderYear has a statistically significant relationship with the response, but also that the variance is not practically meaningful.

Mouse

```
               Df Sum Sq Mean Sq F value Pr(>F)
orderYear       1     20   20.13    0.53  0.467
Residuals   20660 784450   37.97
```

Since the p value is much bigger than 0.05, the null hypothesis fails to get rejected and thus it can be said that there is no statistically significant effect of ‘orderYear’ on the response variable. The low F value also shows that there is not such a big difference in the variances of the groups. The variable ‘orderYear’ likely does not contribute explanatory power to the model and it could be removed without the loss of predictive performance.

Laptop

```
               Df Sum Sq Mean Sq F value Pr(>F)
orderYear       1     19   18.92   0.513  0.474
Residuals   10205 376427   36.89
```

The ANOVA outputs a low p and F value which shows similarly to the Mouse ANOVA that the null hypothesis fails to get rejected and that it can thus be said that there is no significant statistical effect of ‘orderYear’ on the response variable. The low F value also shows that there is not such a big difference in the variances of the groups. The variable ‘orderYear’ likely does not contribute explanatory power to the model and it could be removed without the loss of predictive performance.

Monitor

```
               Df Sum Sq Mean Sq F value Pr(>F)
orderYear       1     17   17.38   0.472  0.492
Residuals   14862 547395   36.83
```

The above ANOVA also says the same as the ANOVA for 'Mouse' and 'Laptop' delivery times. The effect of 'orderYear' once again does not have a statistically significant effect on the delivery times. The null hypothesis fails to get rejected due to the p value exceeding 0.05. The low F value also shows that there is not such a big difference in the variances of the groups. The variable 'orderYear' likely does not contribute explanatory power to the model and it could be removed without the loss of predictive performance.

Software

```
               Df Sum Sq Mean Sq F value Pr(>F)
orderYear       1      0 0.01695   0.179  0.672
Residuals   20747   1966 0.09475
```

The p value is much greater than 0.05, so the null hypothesis fails to get rejected. Thus, there is once again no statistical significance between the 'orderYear' and the delivery times of the product. The F value of 0.179 shows that the between-group variance is smaller than the within-group variance. The predictor has no meaningful explanatory power to the model.

Cloud

```
               Df Sum Sq Mean Sq F value Pr(>F)
orderYear       1      2    1.85   0.049  0.825
Residuals   15596 588231   37.72
```

The above ANOVA shows that the p value is much greater than 0.05 and therefore we fail to reject the null hypothesis, and it can be said that there is no statistically significant effect of the delivery time due to the 'orderYear'. The F statistic is also very small (0.049) which means that the average differences between the groups are virtually non-existent.

7.1 Reliable Service at Car Rental Agency

For this part of the analysis, it is assumed that reliable service occurs when **15 workers** are present. The percentage of times that reliable service levels were obtained will first be obtained from the past data available and then that percentage will be used to estimate the number of days per year that reliable service can be expected in the future. This methodology will only work if the assumption of the past being able to predict the future, is correct.

Firstly, the percentage of days with reliable service during the previous 397 days were the following:

This calculation consists of the number of days where there were 15 or more employees on duty, divided by the total number of days (397), and then multiplied by 100 to convert the fraction into a percentage.

$$Percentage\ of\ reliable\ service\ from\ past\ data = \frac{270 + 96}{1 + 5 + 25 + 96 + 270} * 100$$

This yields a percentage of 92.19%

Thus, the corresponding number of days per year that reliable service (15 workers) can be expected in the future, will be the following:

$$Number\ of\ days\ per\ year\ with\ reliable\ service = 0.9219 * 365$$

This gives an average number of days per year with reliable service of 336.5, which rounded down, gives 336 days out of a year consisting of 365 days. This level of reliability can be considered good. In practice, the actual quality of service level should still be constantly monitored, and customer feedback should ideally be recorded for optimal performance. Continuous improvement should always be the focus point.

7.2 Profit Optimisation for Car Rental Agency

The profit of the car rental agency will be optimised by graphing the costs that correspond to the number of employees hired and doing so for 16 employees through to 20 employees. From there, the optimal number of employees will be determined as the number where the total cost is the lowest. The costs included in this optimisation problem is the cost of hiring more than 16 employees. Every employee that is hired after 16 has already been hired, incurs a cost of R25000 per month. Then, the cost of not providing reliable service (not having at least 15 employees) also contributes quite heavily to the total cost. Every day there is less than 15 employees present, it costs the company R20000 in sales for the day. The graph below takes all these costs into account.

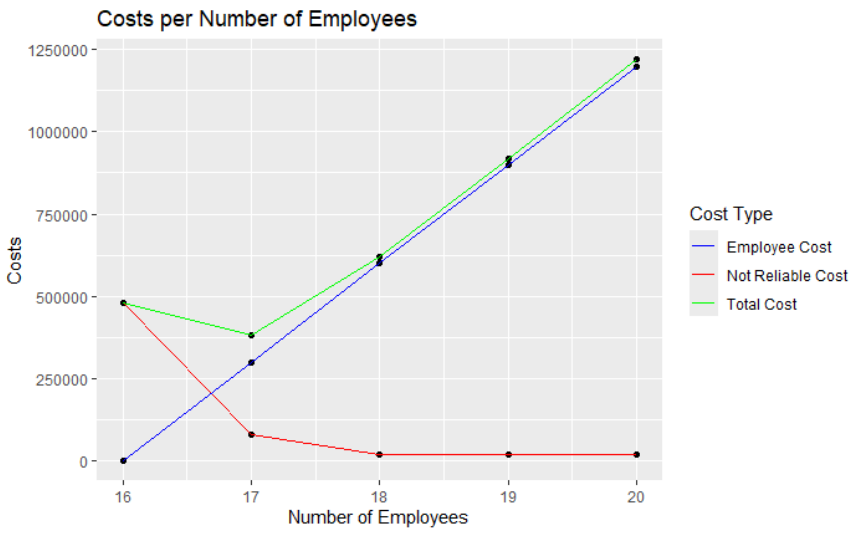
The probability of having less than 15 workers on a day was estimated using a binomial distribution. The probability used for the binomial distribution was calculated by using the data available for the past 397 days.

The methodology for the calculation of the probabilities is the following:

$$P(X < 15) = \binom{n}{x} p^x (1-p)^{n-x}$$

where n = 15, x = the number of baristas and p = the probability of a worker arriving at work (determined from the past data).

This probability is then multiplied by the number of days in a year (assumed at 365) to obtain the number of days during which problems are to be expected. This number of days is then multiplied by the cost of having problems occur, which is a daily cost. That cost is then added to the cost that it will be to hire the specific amount of workers. These calculations were then done for each of the number of employees that are potential options and the total costs, as well as the individual costs, were then graphed below and evaluated accordingly.



The optimal number of employees to hire is **17**. This number leads to the lowest amount of total cost for the company and then also to the most profit. The total cost is indicated by the green line on the graph on the left. Hiring 17 employees will roughly lead to 19 days with problems.

Conclusion

In conclusion, this report highlights many quality assurance principles. Initially, data analysis was done on the main data set to identify potential data quality issues and to potentially extract hidden insights from the data. We then calculated control limits for similar type of data. These control limits are then used to determine the process' ability to stay within these limits and to make it easier to see where the process starts to deviate from the control limits. Data correction was then performed to correct the data quality issues identified during the initial data analysis. Two types of companies' profit were also optimised by making use of various techniques and visualisation methods. ANOVAs were also implemented with the goal of clearing up conclusions that can be made with regards to the changes in delivery times between two different years in the relevant data set for each type of product.

The report concludes with the goal of showcasing many data analysis and quality assurance techniques and contribute to meaningful results and/or conclusions to be seen from the data, companies and models that were looked at.

References

Bhandari, P. (2023). Type I & Type II Errors | Differences, Examples, Visualizations. *Scribbr*.

Kenton, W. (2025). What Is Analysis of Variance (ANOVA)? *Investopedia*.

QA344 Statistics. (n.d.).