b

Hermanus Josias van der Merwe

27069095

Quality Assurance 344

ECSA Data Analysis

Table of Contents

Introduction

This report is to enhance the business understanding, uncover issues that might occur or are common occurrences given the data, understand the operations and measure productivity and discover trends in the data that can be applied or bring attention to the business, for the benefit of the business.

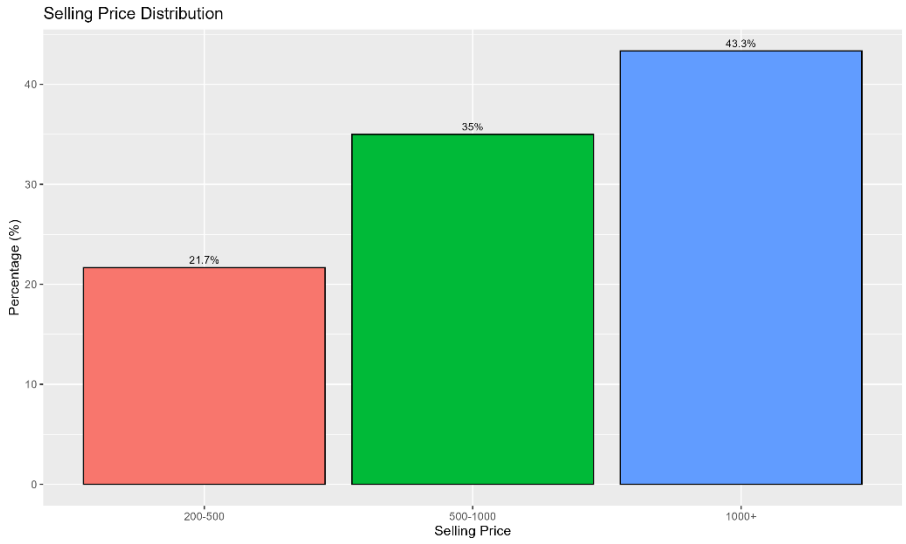
Several datasets will be analysed with the help of R. This data will then be used to create meaningful graphs and statistics. This representations of the data will then be analysed to uncover patterns in the datasets.

The outcome of this analysis is to deepen the operational understanding of the business operation.

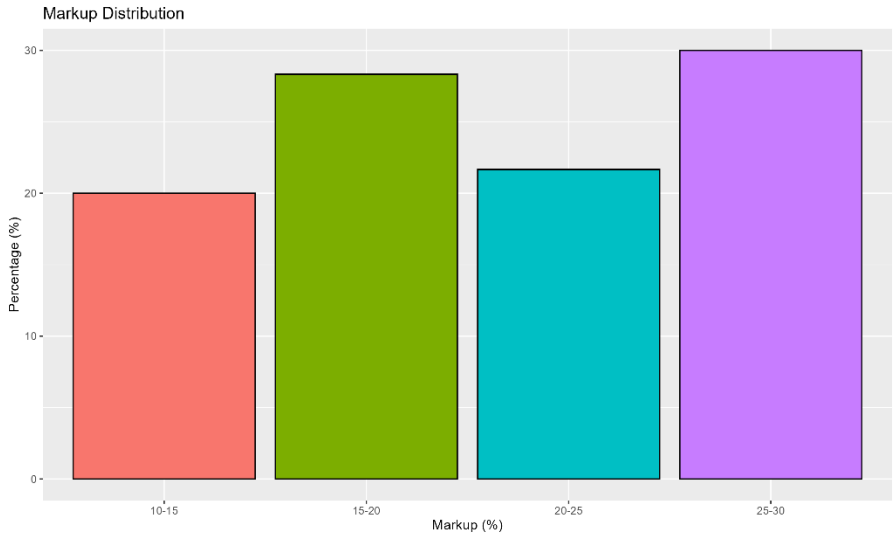
Data set understanding: (Part 1.2)

Graphs:

Pricing Data:



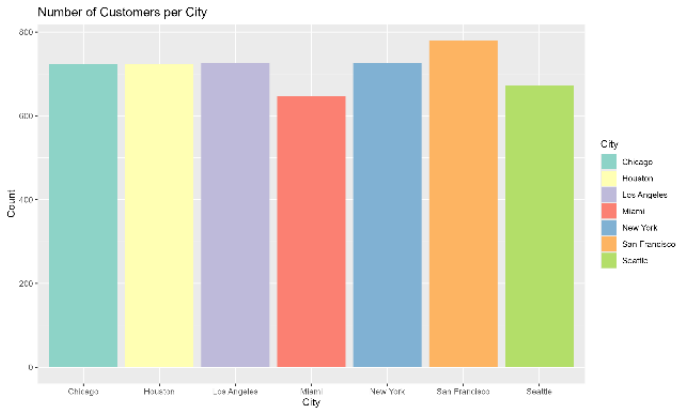
The selling price has a lot of variability. As displayed on the graph, the prices range from 200 – 1000+. However, it can be noted that the majority (43.3%) of the products have a selling price of a 1000+, meaning that although the company covers on a large majority of product price categories, it does tend to have a relatively big influence on the more expensive product ranges (ranging from a 1000 upwards).

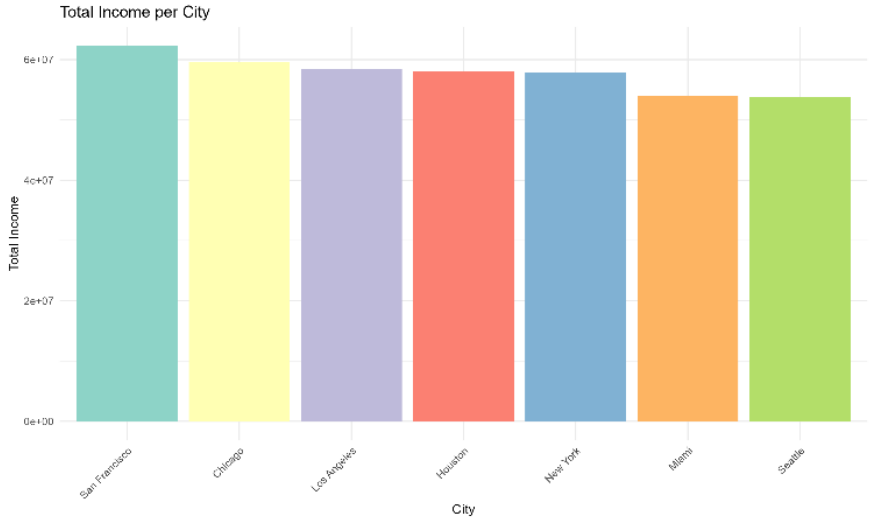


As shown on the graph, the markup distribution ranges from 10 – 30 percent. The distribution tends to be unimodal, meaning that the company does not have a very specific product markup that they tend to sell their products for. The outliers tend to be the 25-30 percent and the 15-20 percent range.

These two markup categories are more than the rest; however, the difference is not significant enough to say that the company mainly focuses on these markup ranges.

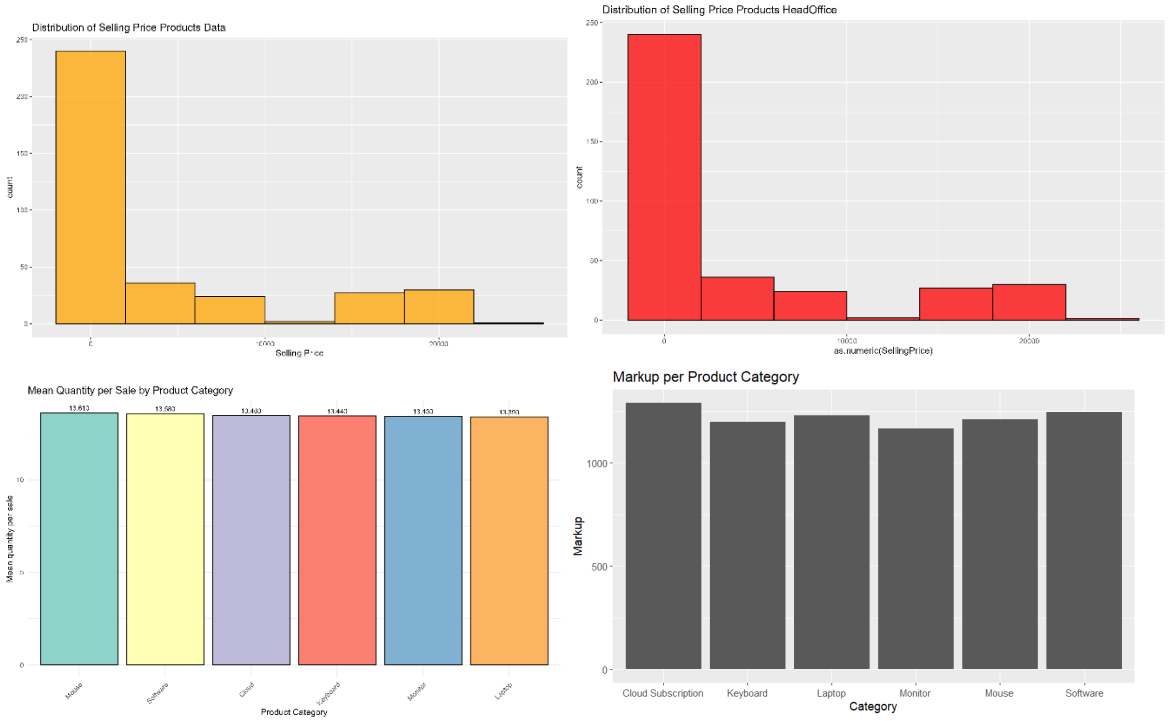
City Specific Customer Data:





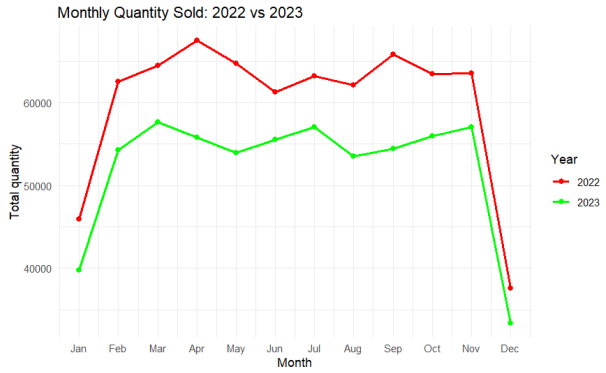
By analysing the two graphs above, although the total income across the different cities does not vary with a large amount, there is a correlation between the total income and the number of customers per city. They can mean that the relatively more well of cities are more likely to be customer of the company or it can be a very simple answer that the more the total customers that are in a city, the more the total income generated from that customers are. Also, it gives us a trend to see which cities are greater supporters of the company than others. Although the distribution of the total income generated per city is unimodal, San Francesco brings in more value to the company than Seattle etc.

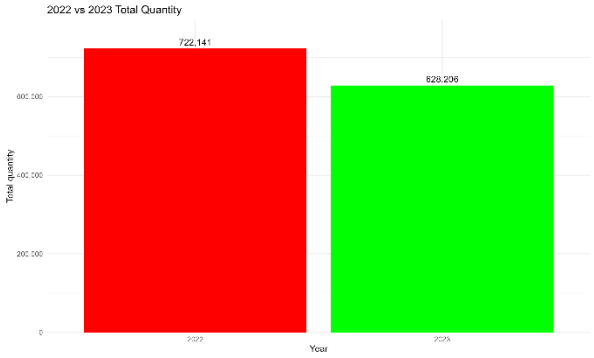
Product Specific Data:



The four graphs above give more insight into the different products that the company is selling. By first glance, looking at the distribution of the selling price, it is very clear that there is a certain category of price ranges that dominates the company's total revenue. This range is between 0 and 250. The company therefore has a clear view on their target market. However, this does not mean that other market categories are not covered by the company, as the selling price goes up to 20000. By further investigation of looking into the "Mean quantity per sale by Product Category" it is clear that the company has a unimodal distribution of all its quantities that it sells per category of products. Similar derivations can be made at looking at the markup per product category.

Sales Data:





By looking at the trend of the sales volumes above, there is a specific trend followed each year. The sales volumes are at an all-time low during December and January, then there are three peaks of similar size during the year in March, Jully and November. Another important factor to take note of is that the sales volumes have dropped from 2022 to 2023. The reasons for this are unknown and might be due to multiple factors but needs to be addressed to ensure growth in coming years.

SPC (Part 3):

(3.1-3.3) Graphs:

x-bar and s-charts were initialized for each product type using R. For each product type the observations were ordered by Year, Month, Day and order time, then grouped into samples of size 24. The initial phase used the first 30 samples (i.e., the first 30 × 24 observations) encountered in that chronological order. Each of the graphs that are formed by looking at the first 30 samples are named as s30-bar or x30-bar to not get confused with the graphs that represents the entire set of data and the graphs that represents only the first 30 observations. Those 30 samples were used to calculate the x-bar and s centre lines and to set the outer control limits as well as the 2-sigma and 1-sigma control limits for the charts.

The formulas used to calculate the statistical control boundaries are as follow:

x-bar:

UCL1 = x_bar_bar + 0.619*sd_bar*(1/3)

UCL2 = x_bar_bar + 0.619*sd_bar*(2/3)

UCL3 = x_bar_bar + 0.619*sd_bar

LCL1 = x_bar_bar - 0.619*sd_bar*(1/3)

LCL2 = x_bar_bar - 0.619*sd_bar*(2/3)

LCL3 = x_bar_bar - 0.619*sd_bar

x_bar_bar = the value obtained from the mean of the first 30 samples

s-bar:

UCL1 = 1.445*sd_bar*(1/3)

UCL2 = 1.445*sd_bar*(2/3)

UCL3 = 1.445*sd_bar

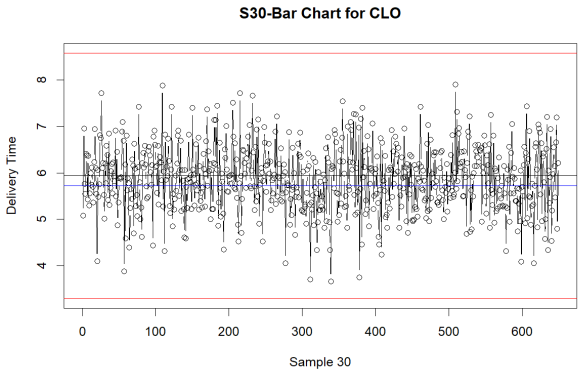
LCL1 = 0.555*sd_bar*(1/3)

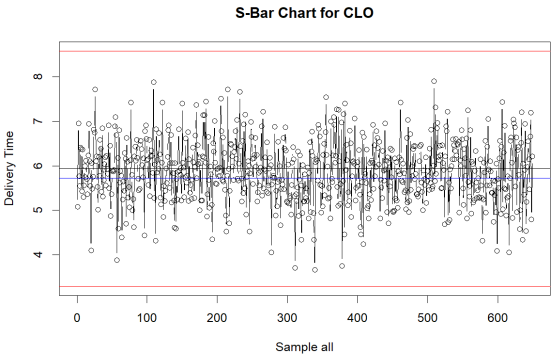
LCL2 = 0.555*sd_bar*(2/3)

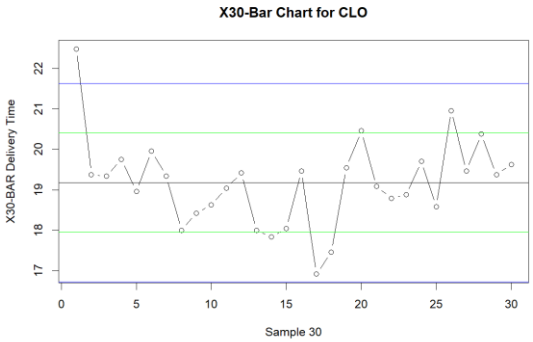
LCL3 = 0.555*sd_bar

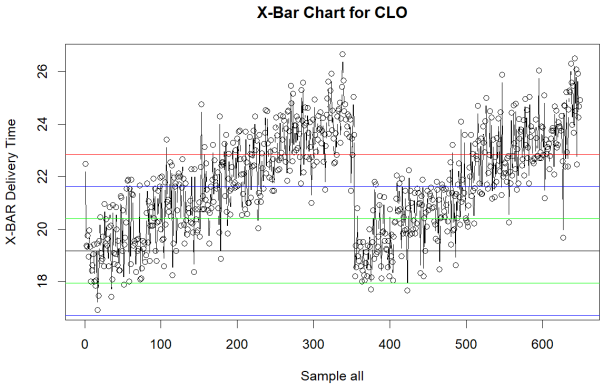
sd_bar = mean calculated from the standard deviation of the first 30 samples

Cloud:

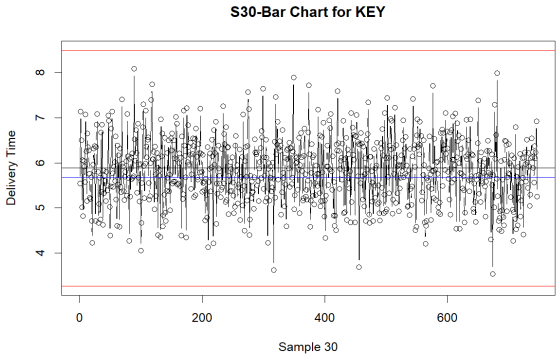

S30-Bar Chart for CLO
Delivery Time
Sample 30

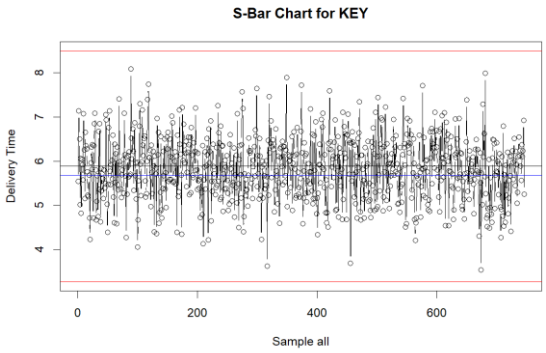

S-Bar Chart for CLO
Delivery Time
Sample all

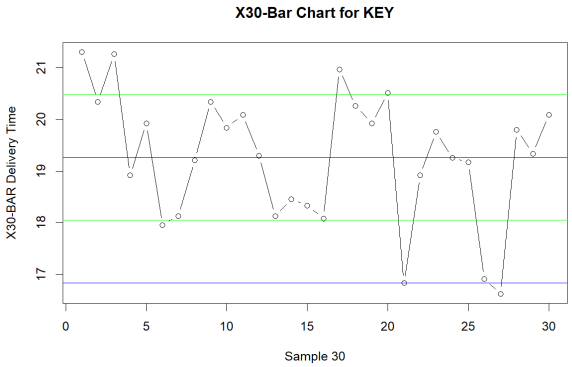

X30-Bar Chart for CLO
X30-BAR Delivery Time
Sample 30

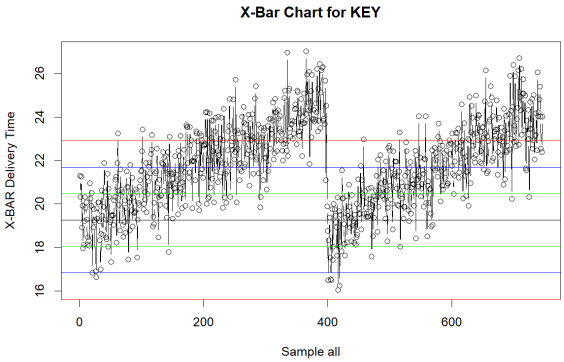

X-Bar Chart for CLO
X-BAR Delivery Time
Sample all

Keyboard:

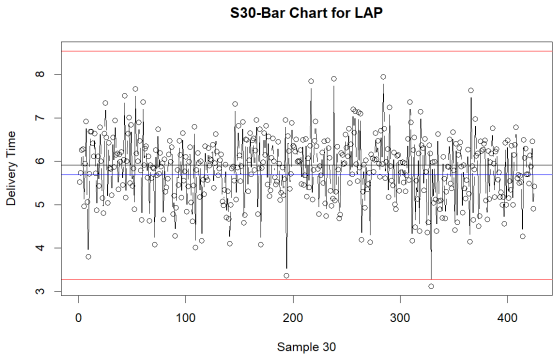


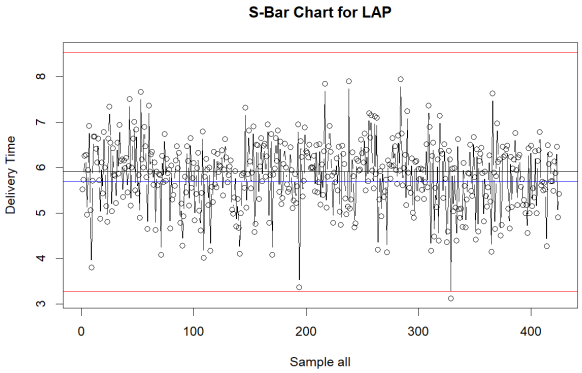


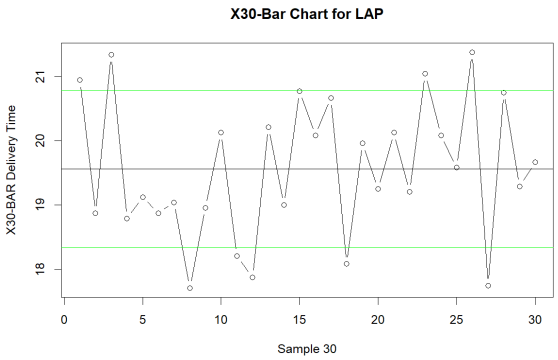


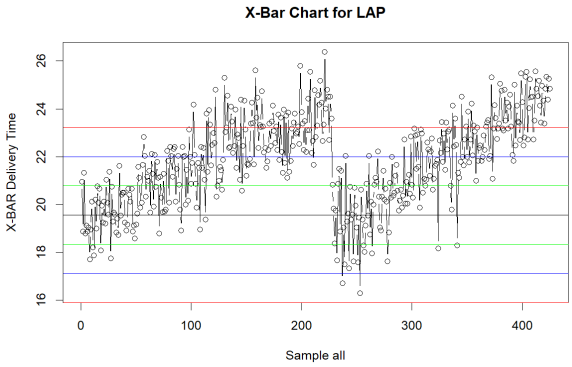


Laptop:

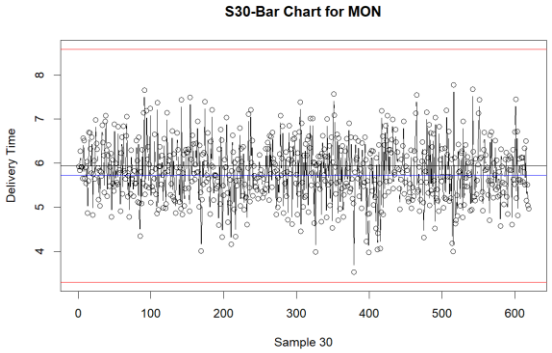


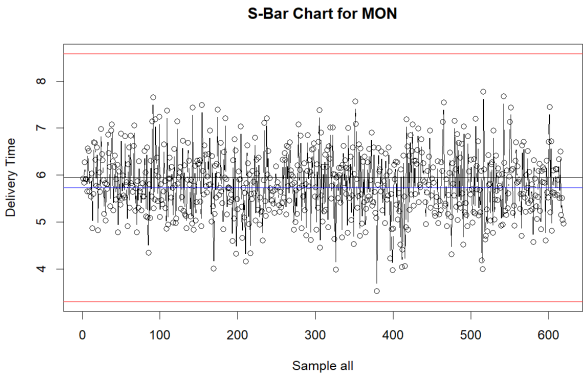


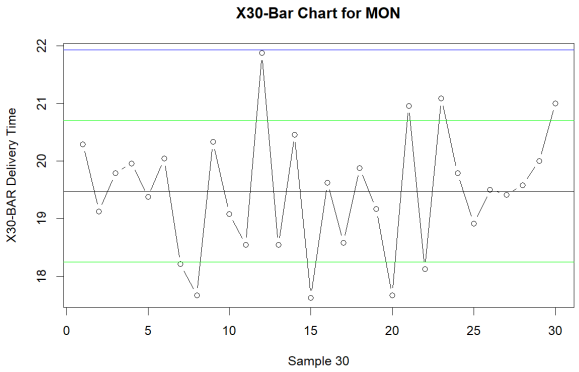


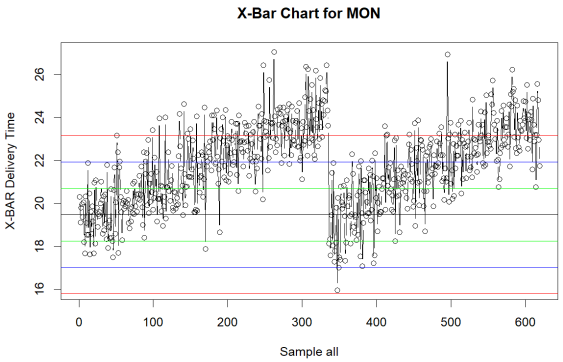


Monitor:

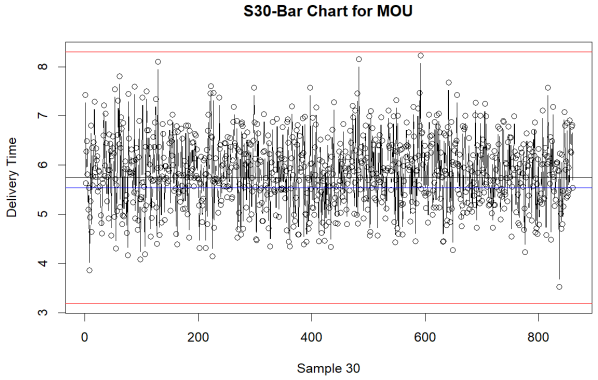

S30-Bar Chart for MON
Delivery Time
Sample 30

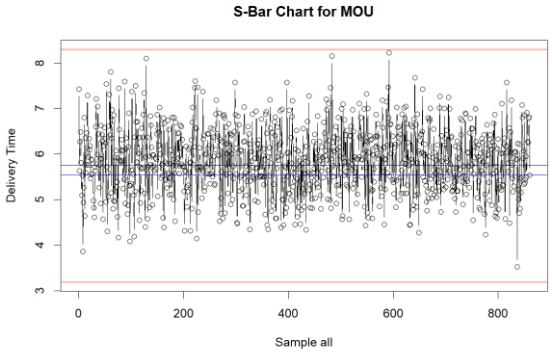

S-Bar Chart for MON
Delivery Time
Sample all

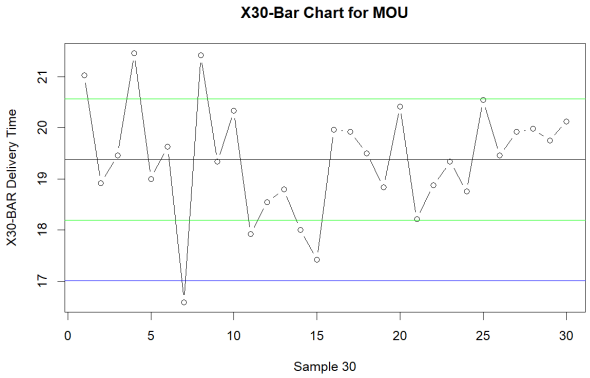

X30-Bar Chart for MON
X30-BAR Delivery Time
Sample 30

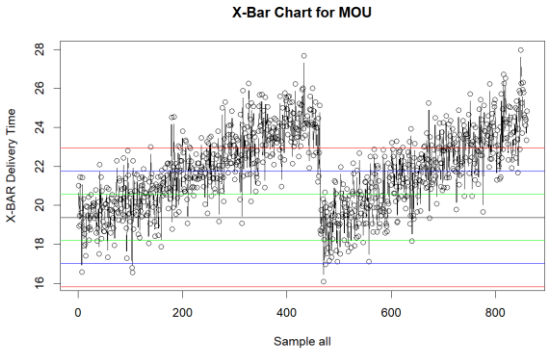

X-Bar Chart for MON
X-BAR Delivery Time
Sample all

Mouse:

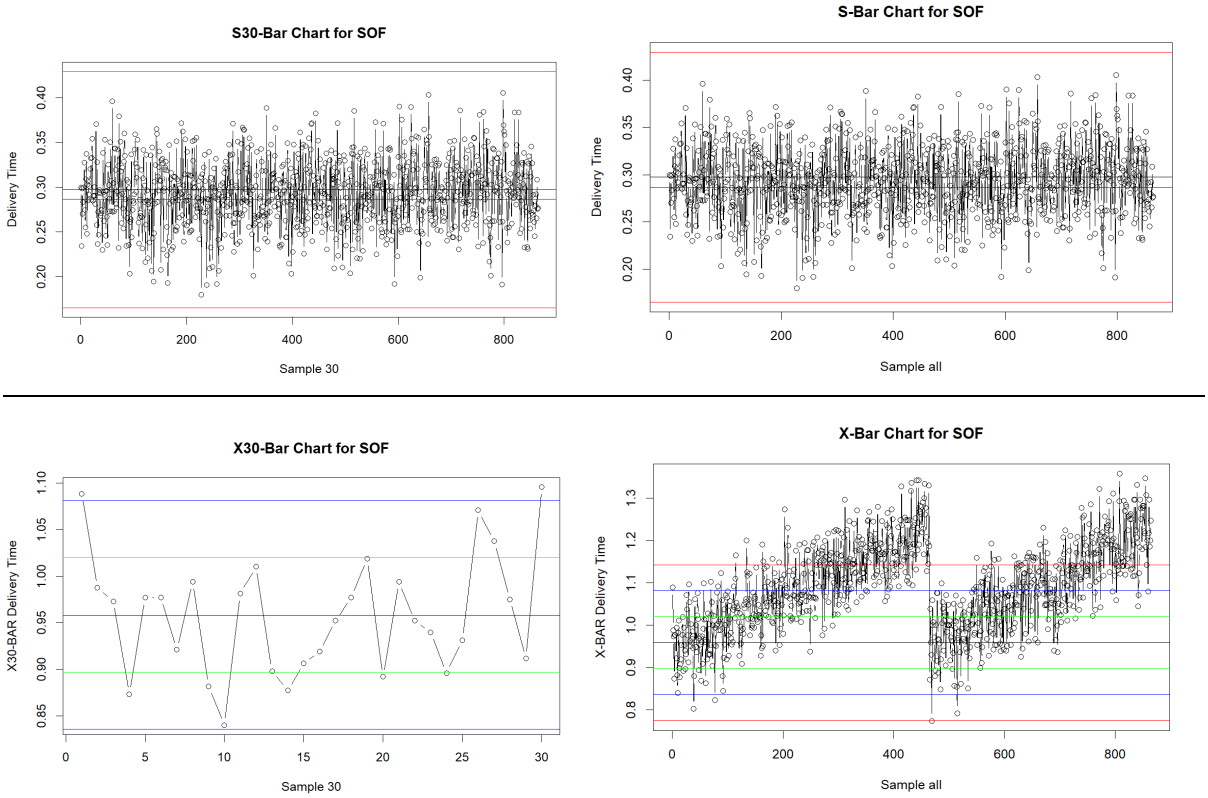

S30-Bar Chart for MOU
Delivery Time
Sample 30


S-Bar Chart for MOU
Delivery Time
Sample all


X30-Bar Chart for MOU
X30-BAR Delivery Time
Sample 30


X-Bar Chart for MOU
X-BAR Delivery Time
Sample all

Software:



Process control analysis:

The process control charts all show a similar trend. The x-bar chart, showing the mean, for the first 30 samples are aligned with the process control limits. However, when looking at the x-bar charts that contains the whole data set, then a common occurrence arises. For the first half of the delivery samples, the samples are within the control limits, with an upward trend. The samples then start get out of hand and consistently rises above the control chart limits. This trend continues until the 400th sample, which is in the middle. The samples then reset back in the bounds of the control limits, it then continues with the same trend as before. This pattern is present within all the product categories.

The reason for this x-bar pattern might be due to the workforce who is motivated and doing their work accurately in the beginning but then slacking off and starting to work slow and ineffective as time goes on. Therefore, after management gives them accountability and checks on them, they will tend to work according to what is required of them again. A possible solution is that management must check on their staff more regularly (preferably every 150-200 deliveries) or have certain procedures in place that checks or motivates the staff members daily to work more effectively and consistently in the long run and not get unmotivated or last as time goes on.

The s-bar charts, which looks at the standard deviation of the product delivery times are mostly within the process control limits and shows no clear trend or pattern.

Key Statistics:

(3.3) Process Capability:

Cp(KEY) = 0.9171	Cpu(KEY) = 0.7294	Cpl(KEY) = 1.1049	Cpk(KEY) = 0.7294
Cp(MOU) = 0.9152	Cpu(MOU) = 0.7266	Cpl(MOU) = 1.1038	Cpk(MOU) = 0.7266
Cp(LAP) = 0.8988	Cpu(LAP) = 0.6962	Cpl(LAP) = 1.1013	Cpk(LAP) = 0.6962
Cp(SOF) = 18.1657	Cpu(SOF) = 35.2473	Cpl(SOF) = 1.0842	Cpk(SOF) = 1.084
Cp(MON) = 0.8890	Cpu(MON) = 0.6996	Cpl(MON) = 1.0785	Cpk(MON) = 0.6996
Cp(CLO) = 0.8977	Cpu(CLO) = 0.7167	Cpl(CLO) = 1.0788	Cpk(CLO) = 0.7167

The charts measure whether delivery times stay within the chosen limits (LCL = 0 hours, UCL = 32 hours). For all the physical products: Keyboard, Mouse, Laptop, Monitor and Cloud, the process capability index (Cp) is below 1, which shows the delivery times are too spread out to reliably meet the 0-32-hour window. Their Cpk values are also under 1, confirming those deliveries are not capable of staying within the limits. Software delivery looks very different, its Cp is extremely high (18.17), indicating very low variation relative to the spec limits. The software's Cpk is 1.08, which suggests it is just capable overall. Because lateness is the only real problem here (only late deliveries matter), the upper-side capability (Cpu = 35.25) is most relevant, and that number strongly indicates the software delivery process reliably avoids late deliveries. Therefore, physical product deliveries need improvement to reduce variation and meet the target window, while software delivery is performing very well, especially at preventing late shipments.

(3.4) Samples Findings:

The three tables below show key findings found in the sample data of the product delivery times.

A: Sample outside of the +3 sigma-control limits

Product	A_total	A_first3_str	A_last3_str
Cloud	0		
Keyboard	0		
Mouse	1	592	592
Software	0		
Laptop	0		
Monitor	0		

Table A shows the number of instances of the standard deviation that was found to fall outside of the upper 3 sigma limit. This is the concerning data points and should be brought to minimum. Only the Mouse product category had an instance that fell outside the limits. Fortunately, it is only 1 occurrence, but the reason for this needs to be identified.

B: Most consecutive samples of standard deviation between -1 and +1 sigma-control limits (measure of good control)

Product	B_max_streak
Cloud	19
Keyboard	20
Mouse	14
Software	19
Laptop	19
Monitor	34

Table B is a measure of how good the control of the process is. The monitor product category shows the most consistent good quality of control, having a consecutive number of samples of standard deviations that falls within the +1 and -1 sigma-control limits of 34, this s good compared to the other products that have a maximum number of consecutive samples of around 20.

C: 4 consecutive x-bar samples outside the upper, second control limits

Product	C_total_above_UCL2	C_total_in_4plus_streaks	C_first3_str	C_last3_str	C_max_streak
Cloud	340	267	56,57,61	647,648,649	48
Keyboard	388	287	40,61,62	744,745,746	92
Mouse	457	328	41,74,87	858,859,860	68
Software	451	326	30,57,70	862,863,864	76
Laptop	207	145	57,58,64	423,424,425	33
Monitor	309	231	12,51,52	616,617,618	35

Table C shows how many samples of the means falls above the second upper sigma-control limit. The results tend to show that this is a common occurrence, with the total number of samples being in a range of 207-457. The number of consecutive samples are also un common, because the range is between 33 and 92.

(4-5) Risk, Optimising for maximum profit; Design of Experiments

4.1 Likelihood estimation of making a Type I (Manufacturer's) Error for A, B and C:

Type 1 error:

A (> +3σ): 0.0013498980316301

B (between ±1σ): 0.682689492137086

C (4 in a row >+2σ): 2.678771559804e-07

The values above give us an indication of the likelihood that we identify something as not being outside of the limits when it is.

4.2 Estimate the likelihood of making type II (Consumer's) Errors for a bottle filling process, that should be centred on 25.05 litres as the process average and CL of an x-bar chart and has an UCL of 25.089 and LCL of 25.011 litres. Unknown to you, the process has moved to an average fill volume of 25.028 litres and now has a x-bar standard deviation of 0.017 instead of 0.013 litres. In the type II error, the Ha is true, but we fail to identify this, due to the sample X-bar and s values being between LCL and UCL of each chart.

<u>Type 2 error:</u>

The beta calculated has a value of 0.9782163. calculating the power of this we use the formula:

Power = 1 – beta

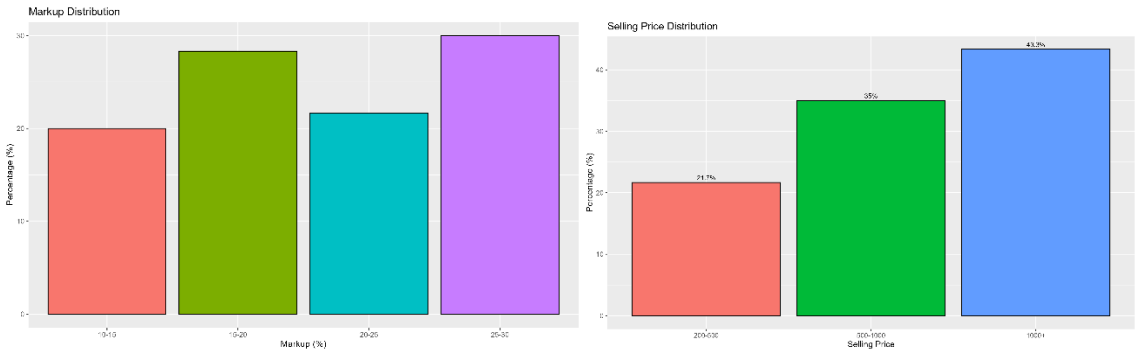
Power = 1 - 0.9782163

Power = 0.02178374

This means that the Type 2 error is 0.02178374. Meaning that the likelihood of us not identifying a sample that is out of specification (out of the control limits), but when it actually is, is 0.02178374.

4.3 Data Analysis (redone)

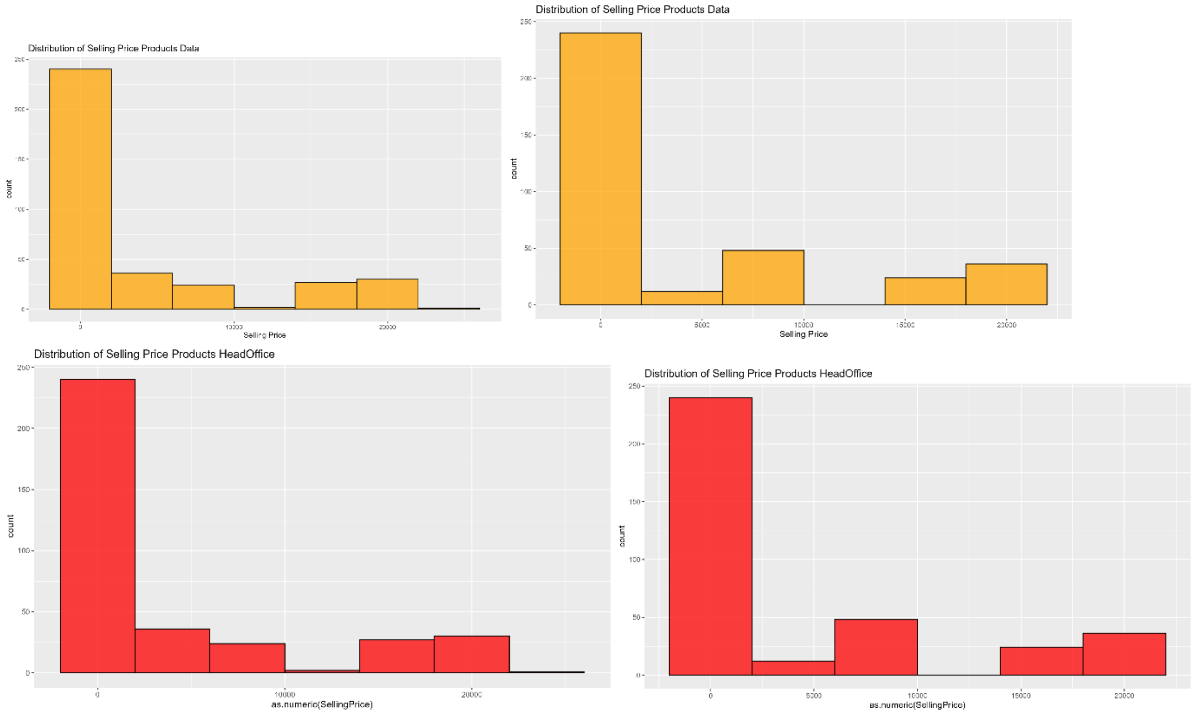
<u>Pricing Data:</u>



<u>City Specific Customer Data:</u>

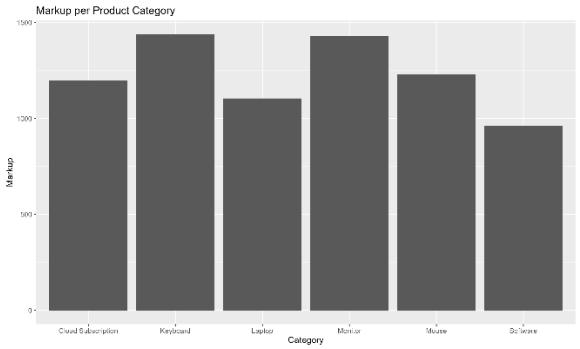
The City Specific Customer Data as well as the Pricing Data graphs does not show any difference after the data have been corrected. Therefore, the analysis for these graphs remains the same.

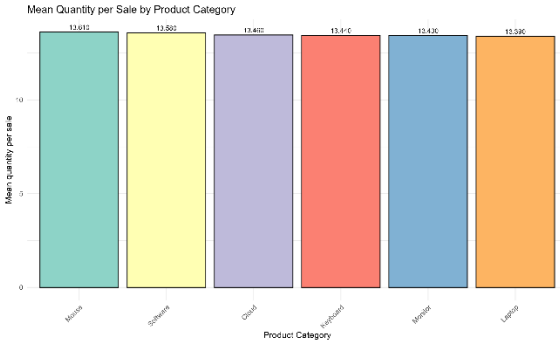
Product Specific Data:



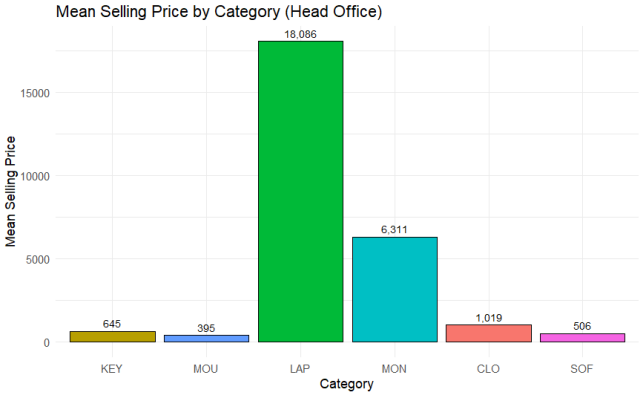
The pricing distribution data seem to have differences that starts to arise.

On the left-hand side are the graphs generated from the data before it has been cleaned and to the right are the graphs generated with the new and corrected clean data. The difference is at the 2nd and 3rd bin; in the pre-cleaned generated graphs the selling price distribution has a gradual descent from the 2nd to 3rd bin. However, with the cleaned data, the opposite is true, the selling price distribution has a gradual increase from the 2nd to 3rd bin. In the big picture, the data has relatively the same distribution, but it is now more clearly and accurately represented that there are more products with a selling price closer to a 700 and a 1000 than around 300 and 500.



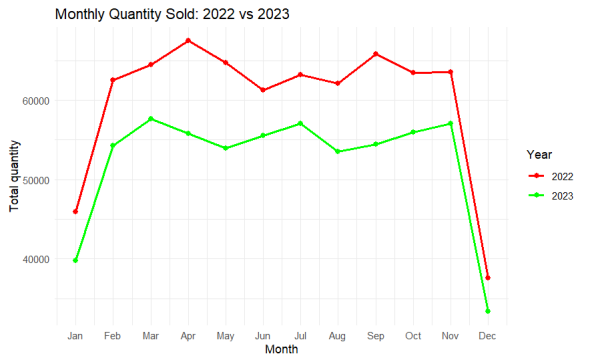


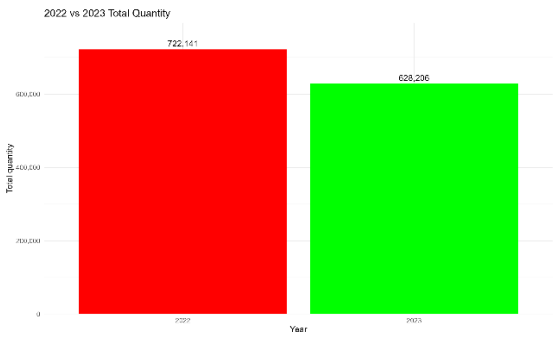
Moreover, the mean quantity sold per product category remains the same, but the markup per product category have changed, the distribution is less uniform than previously showed. It is now clear to see that keyboards and Monitors have a higher markup put on the products and software has the lowest markup put on the price of the products.



Laptops are the most expensive products sold by the company, followed by Monitors, each of these categories falling in their own class of total selling price per product. All the other categories are in a different class, having a selling price much less than that of Laptops and Monitors. There is a clear wide range (395 – 18086) of selling price distribution of the products, because the category of the product will drastically influence the cost of the product. This is common in electronics as the different categories serve different roles and have varying degrees of complexity, laptops for instance having much greater technological advancements packed into it and therefore costing more than a simpler technological product like a Mouse for a laptop.

Selling Data:





The selling data is still the same and still shows the same trends from before the data cleanup. Therefore, the analysis stays the same as previously calculated.

What has changed though is the fact that is now possible to calculate the sum of sales per product category. This was previously not possible as an error occurred.

<u>The following statements are what the output of the R code gave:</u>

[1] "The sum of KEY sales in 2023 is 5378598.87"

[1] "The sum of MOU sales in 2023 is 3773413.87"

[1] "The sum of LAP sales in 2023 is 86027413.33"

[1] "The sum of MON sales in 2023 is 43126707.9"

[1] "The sum of CLO sales in 2023 is 7261887.1"

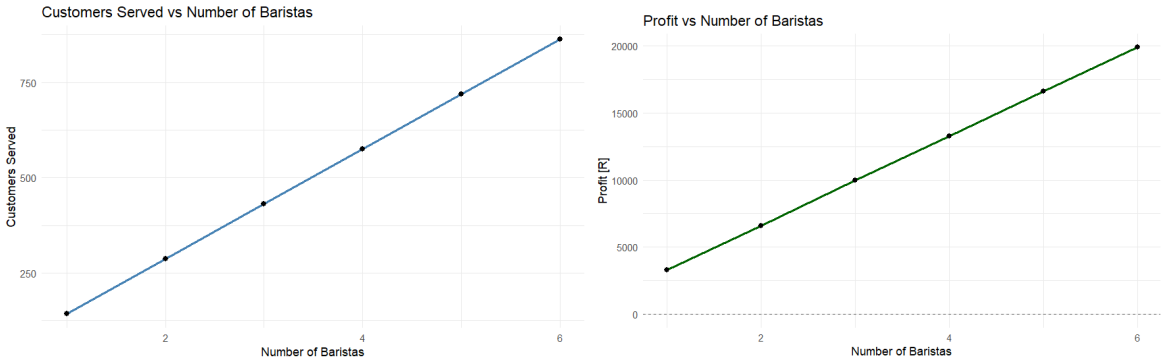
[1] "The sum of SOF sales in 2023 is 4867780.65"

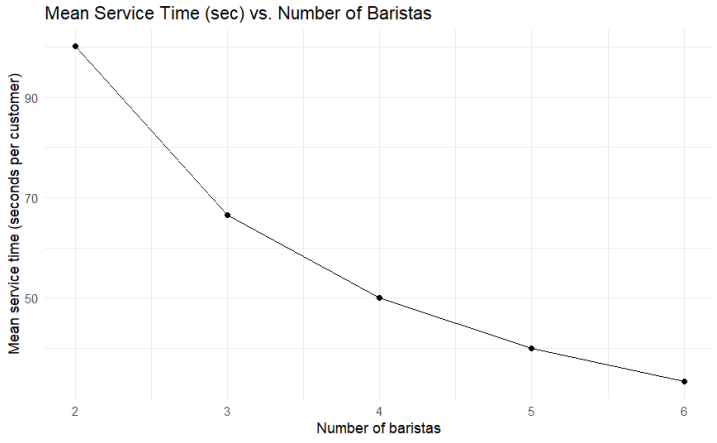
Analysing these six sums generated for each product category shows that the product categories have varying ranges of popularity as the it ranges from 3773413.87 to 86027413.33. Laptops are therefore the most sold product and Mouse the least.

5. Optimization:

First dataset (timeToServe):

For this profit optimization problem, the brute force method was used on R.





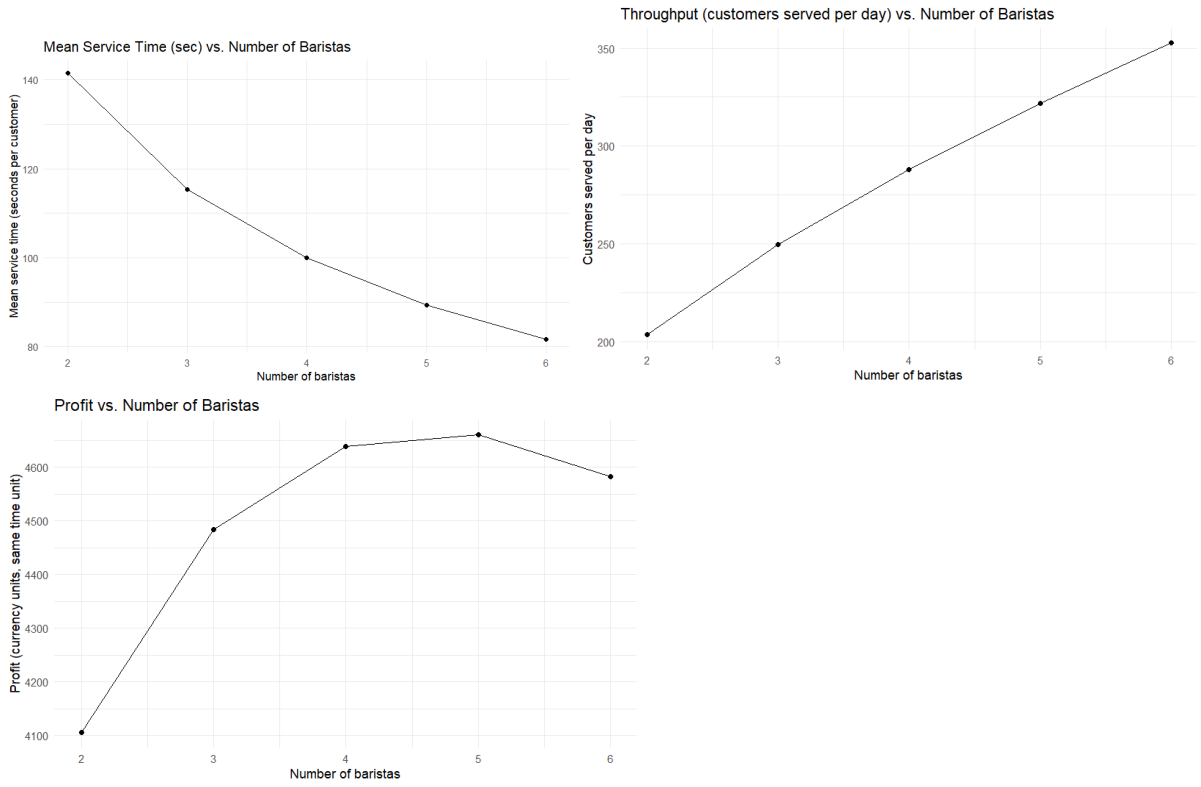
The graphs were generated with the following mathematical equation:

profit = revenue_per_customer * customers_served - add_personnel_cost_per_barista * n_baristas

This was done to test for the maximum amount of profit that the company can make given a number of baristas. Based on the data that has been given, a linear trend is generated. This gives the most optimal number of baristas to be at 6. This data does not however seem to consider the space taken up per barista and so more constraints are needed to make a more educated decision.

By looking at the mean service time per barista graph however, there is a logarithmic trend generated. This means that the contribution of mean time reduction per customer per barista decreases the more baristas are added. So, the optimal solution in a practical world view that it is most likely more optimal to use 4 baristas as your optimal amount of profit for the least number of baristas.

Second dataset (timeToServe2):



The new dataset shows a different trend, which is more realistic in a real-world scenario.

The throughput rate of customers is increasing when more baristas are present. However, when looking at the profit made by the company, a parabolic shape starts to appear. This profit vs number of baristas graph indicates that the profit is increasing up until 5 baristas, thereafter the profit starts to decrease. This means that the optimal number of baristas for the maximum amount of profit is 5.

Part 4: (6-7)

6.2

ANOVA table for each product:

Cloud:

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

Explanation:

For the Cloud product category, the ANOVA produced an F-statistic of 0.049 with a p-value of 0.825.
This very small F value indicates that differences in mean delivery hours between order years explain almost none of the overall variation in delivery times.
Because the p-value is much greater than 0.05 (an estimation of a widely used alpha value), we fail to reject the null hypothesis, meaning there is no statistically significant difference in mean delivery hours between 2022 and 2023 for Cloud.
Any observed differences are likely due to random variation rather than a meaningful year-to-year change.

Software:

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

Explanation:

For the Software product category, the ANOVA produced an F-statistic of 0.179 with a p-value of 0.672.
This very small F value indicates that differences in mean delivery hours between order years explain almost none of the overall variation in delivery times.
Because the p-value is much greater than 0.05 (an estimation of a widely used alpha value), we fail to reject the null hypothesis, meaning there is no statistically significant difference in mean delivery hours between 2022 and 2023 for Software.
Any observed differences are likely due to random variation rather than a meaningful year-to-year change.

Laptop:

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

For the Laptop product category, the ANOVA produced an F-statistic of 0.513 with a p-value of 0.474.
This small F value indicates that the variation in mean delivery hours between the two order years is very minor compared to the overall variation within the data.
Since the p-value (0.474) is much greater than the conventional significance level of 0.05, we fail to reject the null hypothesis.
This means there is no statistically significant difference in the mean delivery hours between 2022 and 2023 for Laptop products.
Any observed differences are likely due to random variation rather than a meaningful year-to-year change.

Monitor:

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

For the Monitor product category, the ANOVA produced an F-statistic of 0.472 with a p-value of 0.492.
This small F value indicates that the variation in mean delivery hours between the two order years is very minor compared to the overall variation within the data.
Since the p-value (0.492) is much greater than the conventional significance level of 0.05 (an estimation of a widely used alpha value), we fail to reject the null hypothesis.
This means there is no statistically significant difference in the mean delivery hours between 2022 and 2023 for Monitor products.
Any observed differences are likely due to random variation rather than a meaningful year-to-year change.

Mouse:

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

For the Mouse product category, the ANOVA produced an F-statistic of 0.53 with a p-value of 0.467.
This small F value indicates that the variation in mean delivery hours between the two order years is very minor compared to the overall variation within the data.
Since the p-value (0.467) is much greater than the conventional significance level of 0.05 (an estimation of a widely used alpha value), we fail to reject the null hypothesis.
This means there is no statistically significant difference in the mean delivery hours between 2022 and 2023 for Mouse products.
Any observed differences are likely due to random variation rather than a meaningful year-to-year change.

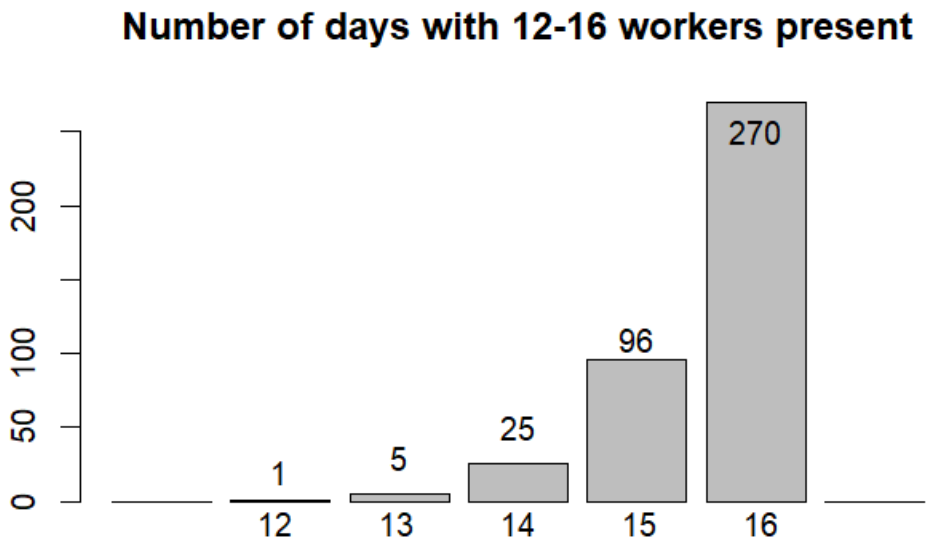
Keyboard:

```
               Df Sum Sq Mean Sq F value  Pr(>F)
orderYear       1    302  302.42   8.088 0.00446
Residuals   17918 669951   37.39
```

For the Laptop product category, the ANOVA produced an F-statistic of 8.088 with a p-value of 0.00446.
This large F value indicates that the variation in mean delivery hours between the two order years is of significant difference compared to the overall variation within the data.
Since the p-value (0.00446) is less than the conventional significance level of 0.05 (an estimation of a widely used alpha value), we reject the null hypothesis.
This means there is statistically significant difference in the mean delivery hours between 2022 and 2023 for Keyboards products.

7.1

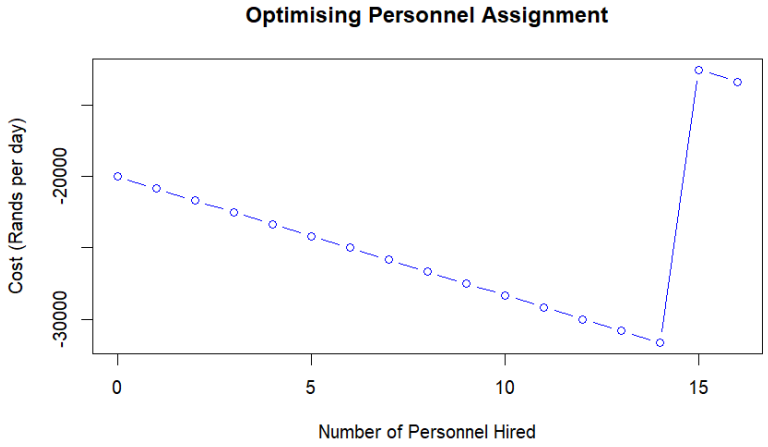
The following graph reflects the number of people on duty over 397 days.



Given the data on the graph above, the estimation for the number of days we should expect reliable service can be the sum of all the days that has between 12 and 16 workers present.

This value is calculated to be: 397. This is the total number of days being analysed and therefore a stricter assumption will be made: that only with 16 workers present, the service is at its most reliable form. This then brings the number of days with reliable service to 270 out of the 397 days.

7.2



The following formula was used for the optimization of the cost modejyl:

cost_model = (-Problem_day_cost * Problem_day) - (Cost_per_person_day * nr_hired)

Optimal number of personnel: 15

minimum daily cost: -12500

The cot model above shows the optimal number of personnel that the company should hire to minimize the cost involved in their operations. This optimal number results to be 15 people, and more or less than that will negatively impact the cost.

Conclusion

Several datasets have been reviewed and analysed in this report. Some graphs and statistics show that certain occurrences or values of data can be uniform in nature, or some have specific binomial distributions. Other trends have also been discovered, and it is made clear which products fit into what price categories and that the products have varying levels of influence in the profit generated by the company.

The key thing to note is that the data showed significant insights, some indicating that actions need to be taken, some being useful for the knowledge of business strategy. Different aspects have been seen, and these aspects are all useful for the business to take note of and get better understanding of their operations and reasons for doing things.

The data should therefore be taken into consideration when that business makes decisions, to make more informed, concise and beneficial decisions.

References