QUALITY ASSURANCE ECSA PROJECT

BC Ramkumar 27411176

Contents

Table of Tables

Table of Figures

Introduction

The purpose of the following project is to use advanced quality assurance, statistical and data analytical techniques to analyse and predict outcomes of datasets. The project finds many different KPI indicators, detects different risks and errors and gives recommendations based of different statistical methods.

The analysis follows a ECSA document and is numbered accordingly. The document starts off with basic descriptive analysis, statistical process control, risk management, performance optimization and performance reliability. Each section is done with detail to explain the process taken.

All analysis was done using R and excel databases. Where all values were calculated, visuals created, and tables displayed. This is where all final decisions will be made based off their statistical calculations. The goal of this project is to highlight area of improvement and areas that are successful through various methods.

1. Descriptive Analysis

1.1. Structure of Data

The data consists of 4 excel files each containing different data. The first file is named "product_data". This file consists of 60 rows of data which has information about the category, description, selling price and markup of different products. The second file is named products "products_headoffice". This file consists of 360 rows and data and 5 columns. This file consists of the same headings as the previous file. The third file is named "sales2022and2023". This file consists of 100000 rows and 9 columns of information. Namely CustomerID, ProductID, Quantity, orderTime, orderDay, orderMonth, orderYear, pickingHours and deliveryHours. The last file is named "customer_data". This file contains 5000 rows and 5 columns of data. Namely CustomerID, Gender, Age, Income and City

1.2. Summary Statistics

The important data to analyse in "products_data" is the selling price. The following table consists of information about the selling price. After updating the file as per request. The values for the following table did not change as only the category column has been manipulated.

Min	1st Quartile	Median	Mean	3rd Quartile	Max
350.4	512.2	794.2	4493.6	6416.7	19725.2

Table 1

In addition, for "products_headoffice" the selling price will also only be analysed and information about it can be found in the table below. As seen below these values are the same as "products_data" file as expected after the change.

Min	1st Quartile	Median	Mean	3rd Quartile	Max
350.4	512.2	794.2	4493.6	6416.7	19725.2

Table 2

The file "sales2022and2023" contains data that is better analysed with graphs due to the data being relatively in interval format. However, the last file can have one of their features analysed. This being income. The information can be found in the table below.

Min	1st Quartile	Median	Mean	3rd Quartile	Max
5000	55000	85000	80797	105000	140000

Table 3

1.3. Missing Values

Using code in R, it was determined that there is no missing values in any of the excel files. Therefore, there is no need for any corrections to be made.

1.4. Data Visualization

We can explore various relations with the following data. However, the most important relationships will be explored here and explained.

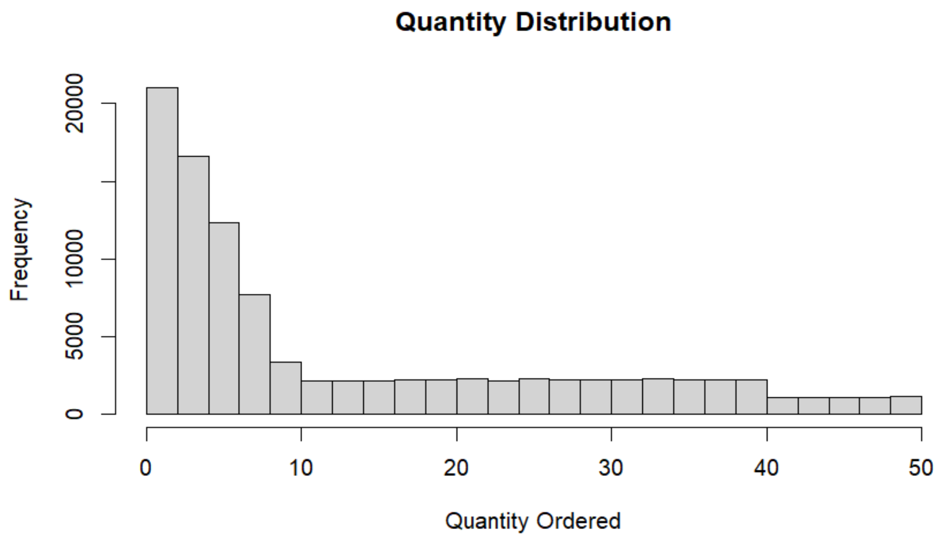


Figure 1

From the histogram above, it is clear that majority of the order have a low frequency. However, there is a few outliers with high quantity. This histogram shows that customers prefer to order our products in a lower quantity of products as expected. Measures can be taken if the quantity ordered wants to be increased such as bulk deals. This information was unaffected by the change in “products_headoffice”.

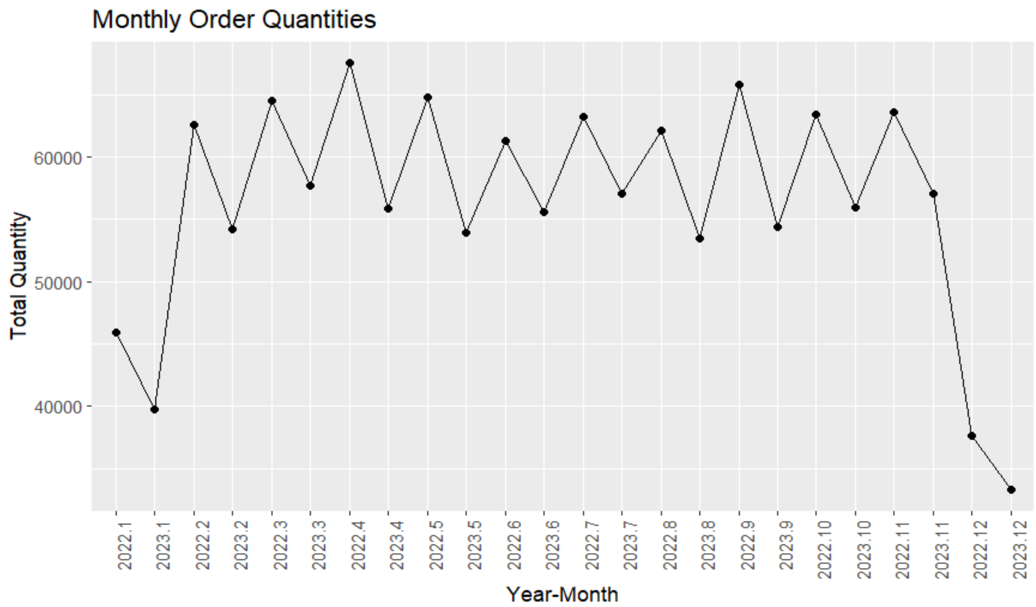


Figure 2

The following graph shows the total quantity ordered in every month of 2022 and 2023. The following graph shows the trend that there in little quality in the summer months. November, December, January and February. The rest of the year the quantity is generally stable with 2 peaks during April and September. Furthermore, there is a general trend that the year 2022 had much more quantity than 2023.

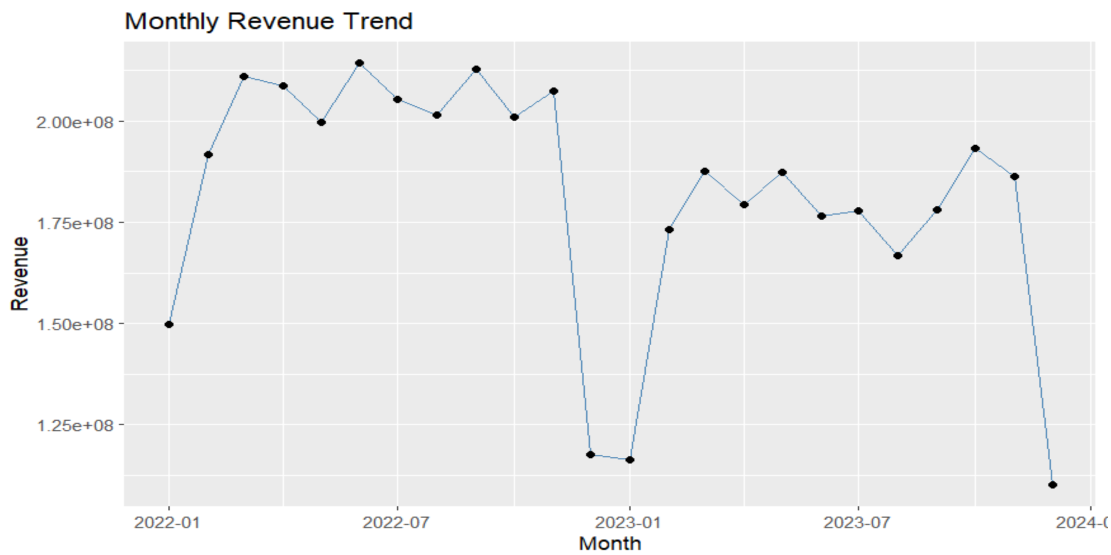
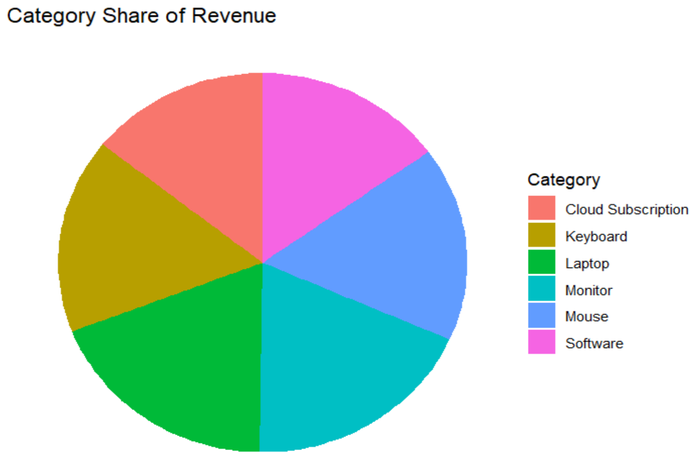


Figure 3

The graph above shows the revenue earned during each month in 2022 and 2023. The year 2022 had a much higher overall revenue than 2023. The Reason why needs to be looked at. Furthermore, you can see the summer months have much less revenue. The company should look for ways to increase this. Furthermore, this graph directly correlates to the quantity graph.

Before Corrections



After Corrections

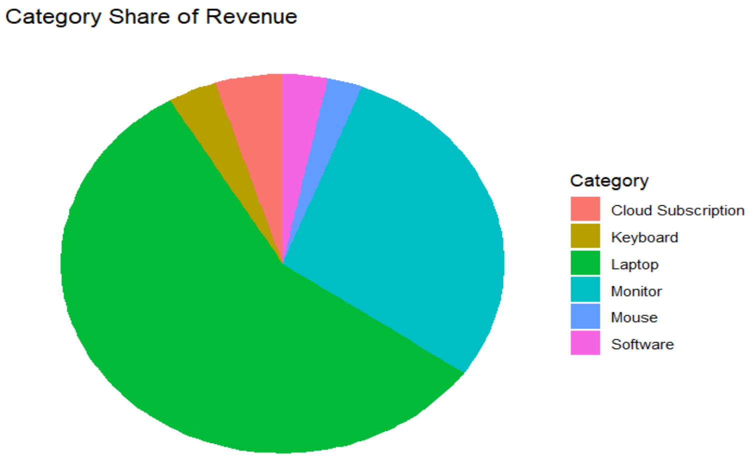
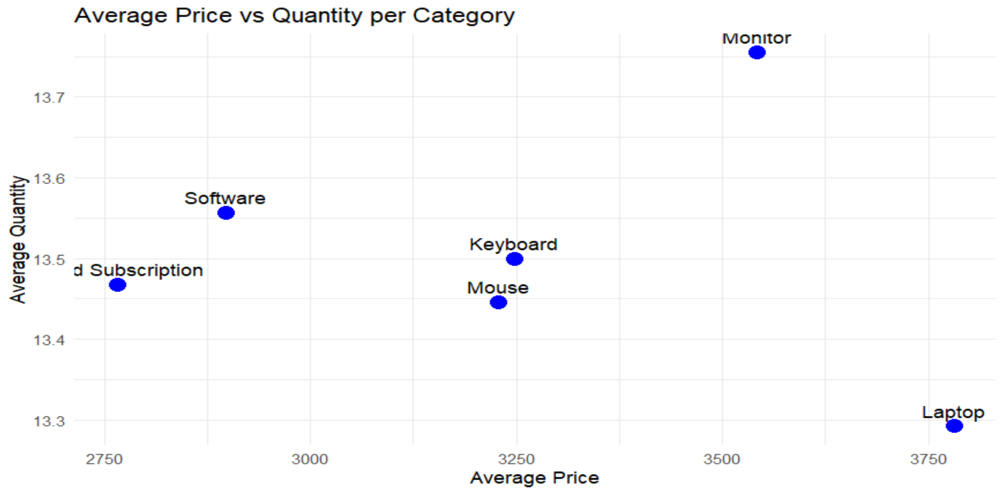


Figure 4

The following pie charts shows the distribution of revenue for each category. Before fixing the error in the old" product_headoffice" file, the pie chart showed each category having an even distribution of revenue. However, after fixing the file, the correct part chart can be seen. It shows that the company sells majority of laptops. Making this product their most important. In addition, they sell a lot of monitors relatively when compared to the other products.

Before Corrections



After Corrections

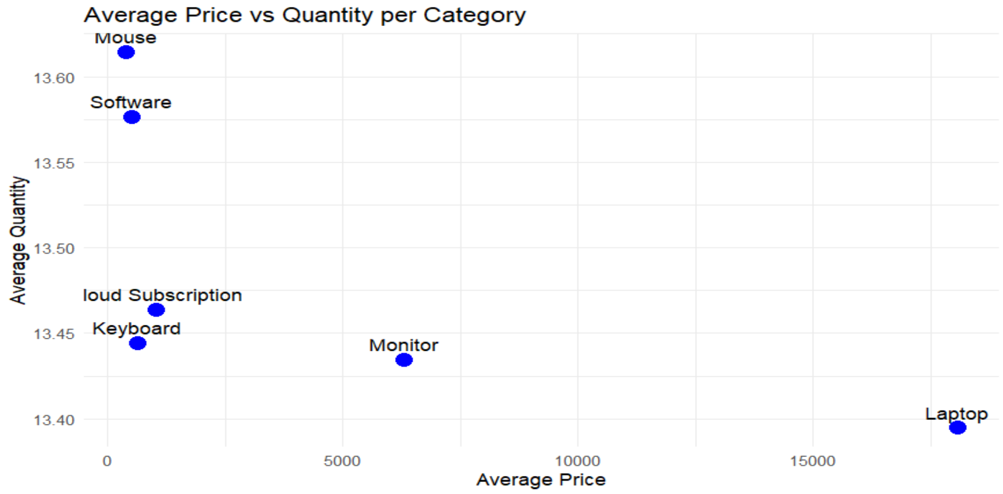


Figure 5

The following scatter plot shows the average quantity vs average price for each category. Before an changes were made to correct the file, the following was examined. Laptops have the highest average price and the lowest average quantity but had the highest revenue. Whereas cloud subscription had the lowest revenue but has the lowest price. After changes were made, all points on the graph shifted upwards in average price. However, it was seen that the quantity of monitors dropped and the quantity of mouses increased. Even though values changed, the two highest revenue products produce around 90% revenue for the company. Which indicates that changes can be made in order to involve other products more frequently.

1.5. Total Revenue

Category <chr>	TotalRevenue <dbl>
Laptop	2470814376
Monitor	1258942847
Cloud Subscription	214110418
Keyboard	155002210
Software	142527355
Mouse	111190471

Table 4

City <chr>	TotalRevenue <dbl>
Los Angeles	722173478
San Francisco	674061345
New York	623940704
Houston	598904598
Seattle	587080997
Chicago	574464154
Miami	571962403

Table 5

AgeGroup <fctr>	TotalRevenue <dbl>
18-29	509869702
30-39	554993211
40-49	571736810
50-59	667206343
60+	1933018427
NA	115763184

Table 6

The images above show the actual revenue for each category, cities and age. The images above show what is our best product being laptops. Where our best market is, which is in Los Angeles as well as the age our company sells to the most which is 60+. This yields total revenue of R4 352 587 678.

1.6. Exploring Relationships

The following table is a corelation table of the sales data.

```
              Quantity     orderTime pickingHours deliveryHours
Quantity       1.000000000  0.0058497274 -0.004688722 -0.0027091738
orderTime      0.005849727  1.0000000000 -0.002015252  0.0004686641
pickingHours  -0.004688722 -0.0020152522  1.000000000  0.5831669256
deliveryHours -0.002709174  0.0004686641  0.583166926  1.0000000000
```

As you can see, the only two features with a relatively high correlation is picking hours and delivery hours. This means that they are similar and can affect each other. This is most likely because they follow a similar procedure and therefore have the highest correlation between these features. Furthermore, we can see order time having basically no correlation to any of these features.

2. Statistical Process Control (SPC)

2.1. Objective

The objective of this part of the project is to monitor the consistency and stability of the company's delivery times. This is done by using SPC methods. More specifically, X-bar and S control charts were made to track the mean variation and process variation with regards to the limitations.

2.2. Method

The dataset "sales2026and2027" has been used to generate all the different X-bar and S control charts. They are order by year, month, day and time. ProductID's are arranged in groups of 30 of 24 sales each.

The qcc package in R is used to generate the S-bar and S control charts. The UCL's and LCL's were thereafter automatically calculated. Using the +- 3 sigma deviations. The centre lines were drawn from the means for each graph respectively. This method allows for graphs to be easily analysed.

2.3. Results

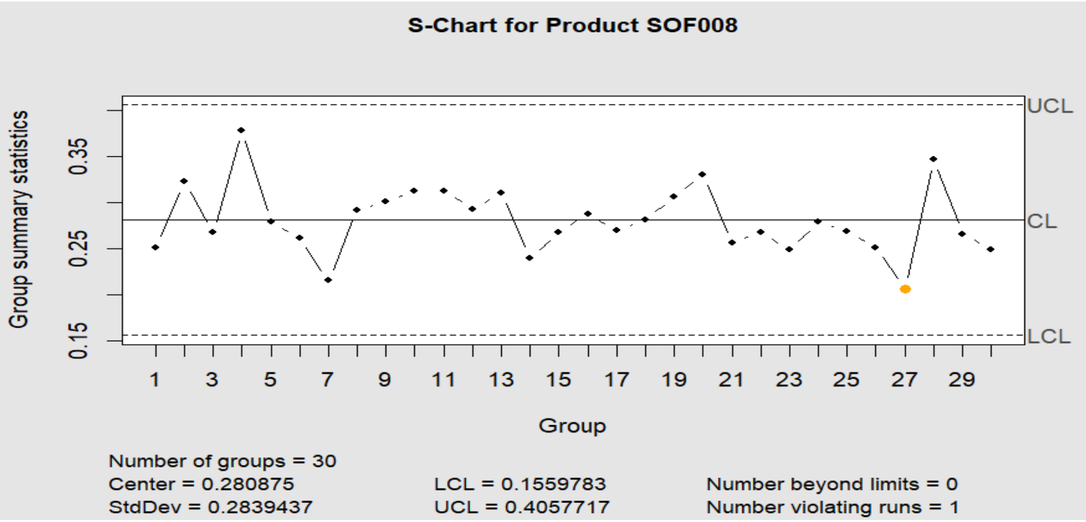


Figure 6

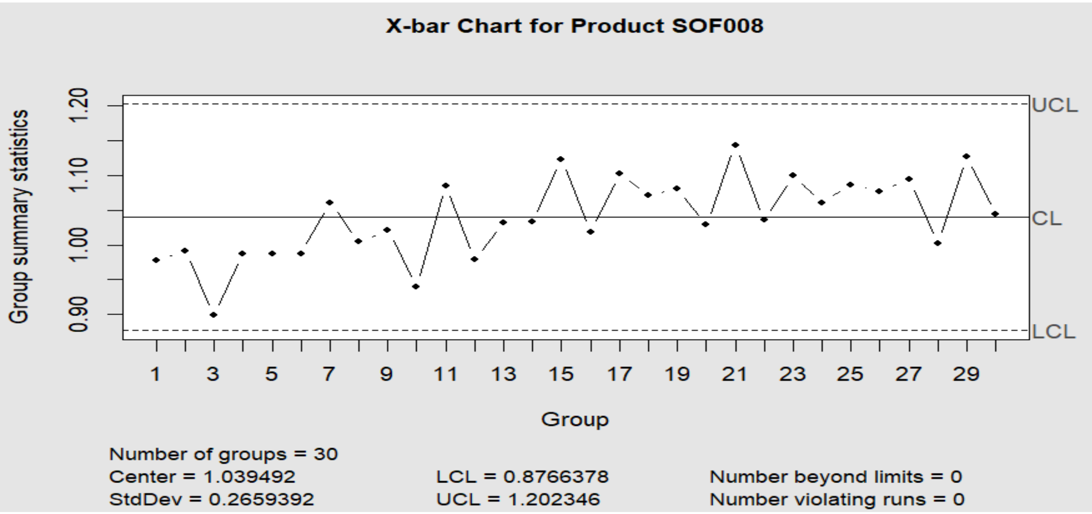


Figure 7

As seen above for product SOF008, all data is stable. For the S control chart, all points are within the upper and lower confidence limits. With only 1 violating run. Furthermore, the same applies for the X-bar chart. Therefore, this products delivery times in in statistical control and no action needs to be taken.

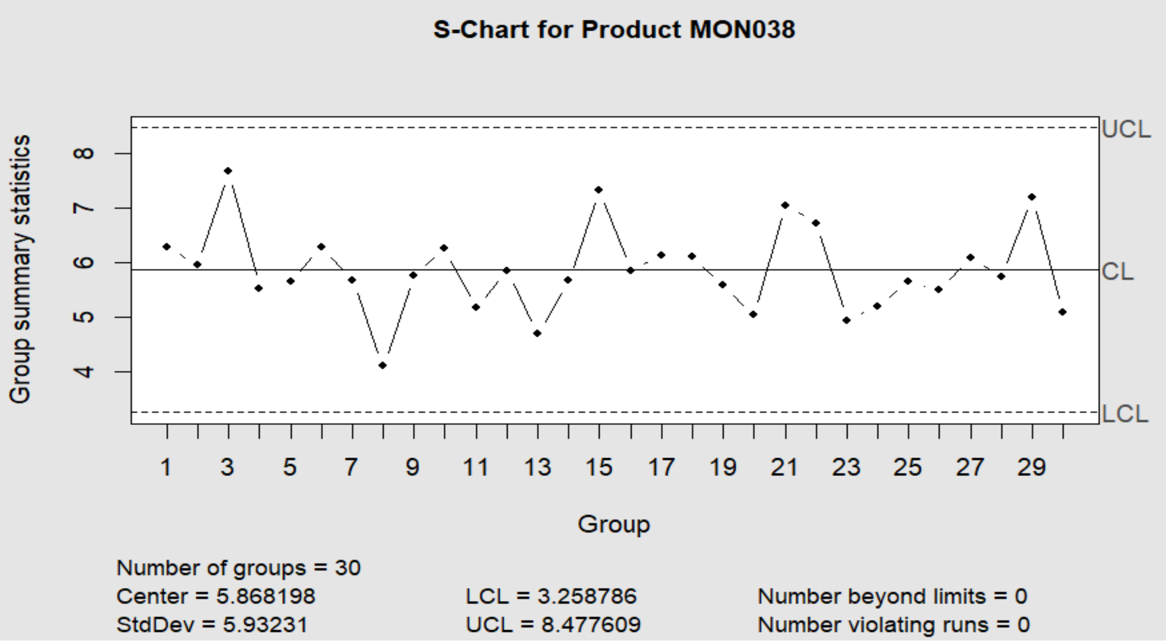


Figure 8

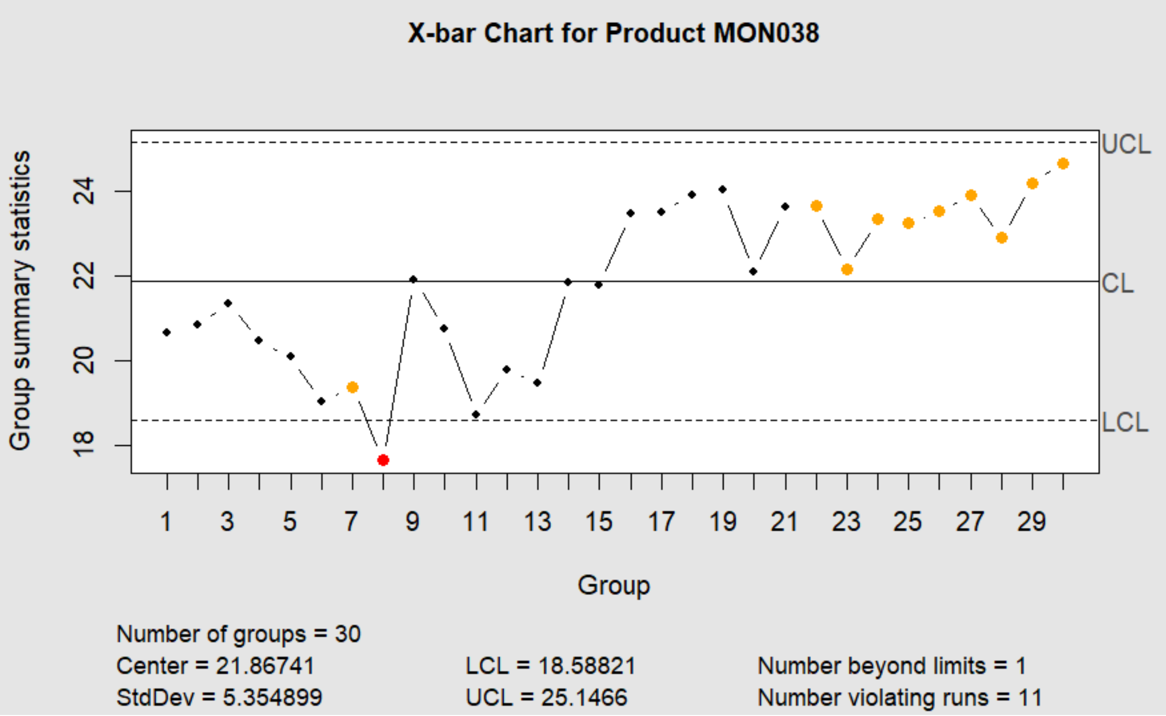


Figure 9

The following Product MON038 has no problem with variation in standard deviation. However, as seen in the X-bar chart- there is 11 violation runs and 1 beyond limit instances. Showing that there is, in this instance an increase in delivery time means. This could be due to various reasons such as wear/tear, human issues and seasonality. However, methods should be taken to investigate this issue.

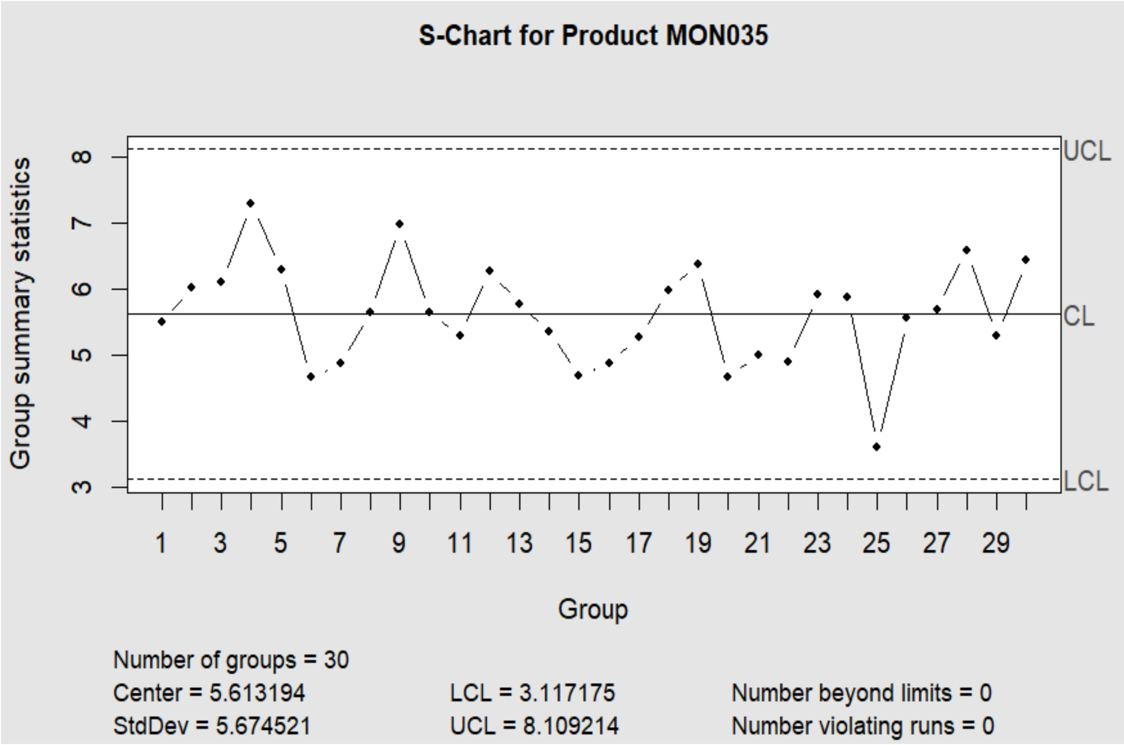


Figure 10

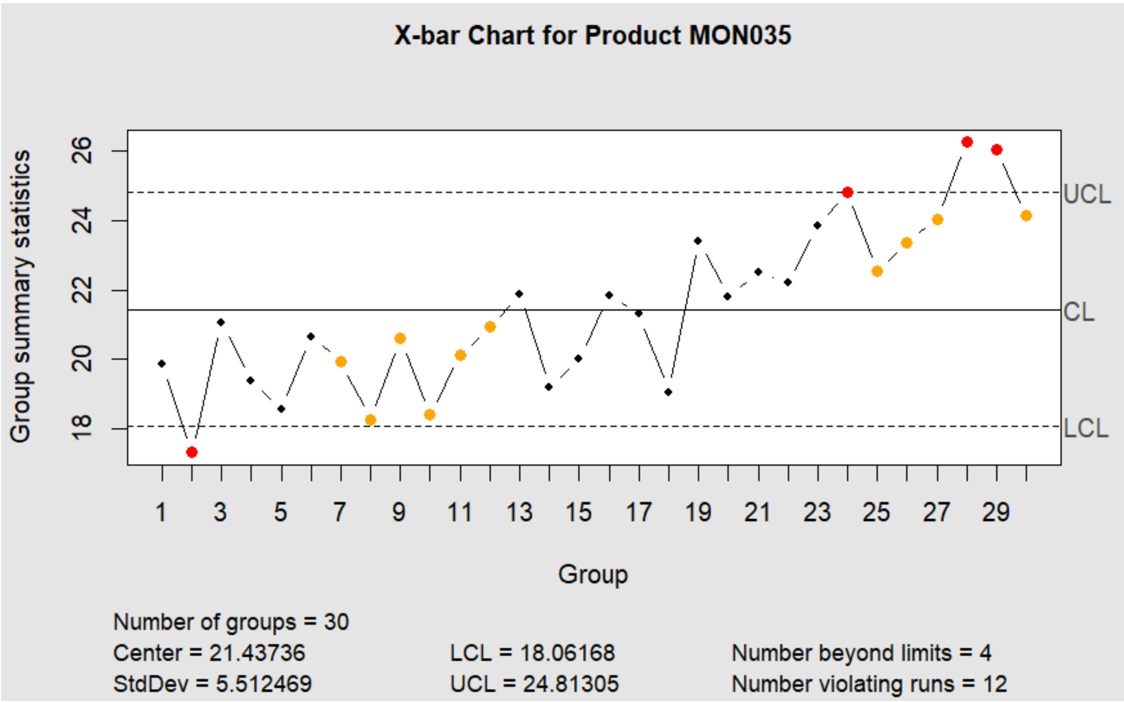


Figure 11

The following product MON035 has a unstable mean. It shows lots of variation as well as 4 beyond limit points as well as 12 violation runs. Therefor 16 out of the 24 sales have had problems. Showing that there is a big problem that needs to be investigated.

2.4. Process Capability Indices

The process capability indices such as Cp, Cpu, Cpl, Cpk were calculated using the first 1000 instances in the database. As seen in the image below, these are the 10 best Cpks of the ProductID's. Furthermore, these are the only Cpk's above 1.

ProductID <chr>	mean_deliv <dbl>	sd_deliv <dbl>	Cp <dbl>	Cpu <dbl>	Cpl <dbl>	Cpk <dbl>	count <int>
SOF008	1.075675	0.2924531	18.2365410	35.2470450	1.226037	1.2260370	1000
SOF003	1.069425	0.2954763	18.0499544	34.8934667	1.206442	1.2064420	1000
SOF010	1.069425	0.2964676	17.9895963	34.7767848	1.202408	1.2024078	1000
SOF007	1.085750	0.3044887	17.5157027	33.8428008	1.188605	1.1886046	1000
SOF009	1.085725	0.3050057	17.4860119	33.7854613	1.186563	1.1865625	1000
SOF004	1.069825	0.3042944	17.5268865	33.8818541	1.171919	1.1719188	1000
SOF006	1.059400	0.3019032	17.6657052	34.1617200	1.169691	1.1696905	1000
SOF005	1.078225	0.3083722	17.2951168	33.4247319	1.165502	1.1655017	1000
SOF002	1.064625	0.3082288	17.3031630	33.4549897	1.151336	1.1513362	1000
SOF001	1.069425	0.3100505	17.2014995	33.2532669	1.149732	1.1497321	1000

Table 7

Therefore, showing that these products are marginally reliable to meet delivery times. Whereas the other products should be looked at to improve their reliability of their delivery times. However, there are no Cpk greater than 1.33. Meaning that there is no fully reliable products in terms of product reliability.

2.5. Process Control Rules

2.5.1. Rule A

The following instance falls outside the limits of rule A. Where the sample is outside of the upper limit is +-3 sigma.

ProductID <chr>	sample <dbl>	mean_sample <dbl>	center <dbl>	sigma <dbl>	UCL <dbl>	LCL <dbl>	OutOfControl <lgl>
CLO018	67	30.046	21.70885	2.358349	28.78390	14.63380	TRUE
LAP029	43	37.092	22.07540	3.139652	31.49436	12.65645	TRUE
MON039	64	34.046	21.78646	2.316038	28.73457	14.83834	TRUE

Table 8

2.5.2. Rule B

The following are the 10 best instances that have the most consecutive samples between -1 and 1 sigma control limits.

ProductID <chr>	LongestStableRun <int>
MON039	23
SOF007	22
SOF010	22
MOU055	21
SOF008	21
SOF004	19
SOF005	19
MOU051	18
KEY044	17
MOU053	17

Table 9

2.5.3. Rule C

The following instances have 4 consecutive upper second control limits.

ProductID <chr>	MaxConsecutiveOut <dbl>	TotalViolations <int>
KEY048	2	3
MON035	2	3
MOU051	2	3
MOU056	2	2
MOU058	2	5
SOF003	2	2
SOF008	2	4
SOF010	2	5
CLO011	1	1
CLO012	1	2

Table 10

3. Risk Management

3.1. Type I Error

A type I error is when you reject Ho but Ho is in fact true. Essentially a false positive or In other words a "false alarm".

3.1.1. Rule A

The probability that Rule A has a type I error was calculated at: Rule A (1 above $+3\sigma$): 0.001350 (0.135%). Therefore, the probability that 1 sample will be above the $+3\sigma$ is very low.

3.1.2. Rule B

The probability that Rule B has a type I error is calculated at: Rule B (7 above centreline): 0.007812 (0.781%). This indicates that there is still a low probability that there will be 7 consecutive samples above the centre line. Indicting a low false alarm rate.

3.1.3. Rule C

The probability that Rule C has a type I error is calculated at: ule C (4 above $+2\sigma$): 0.0000002679 (0.000027%). Which is an extremely low probability and likely to never happen. Meaning no false alarm risk.

3.2. Type II Error

The following type II error was calculated on the basis that Ha was true but we failed to identify this. This was the information given to make the predictions. Target mean = 25.05, LCL = 25.011, UCL = 25.089, shifted mean = 25.028 and $\sigma = 0.017$. The following results were calculated using R: z_L = -1.000, z_U = 3.588, β (Type II error) = 0.841 (84.1%), Power ($1 - \beta$) = 0.159 (15.9%).

The results show that there will be an 84.1% chance that the shift will not be detected. Which implies that the control charts are not sensitive enough for small mean changes. Only 15.9% of the time will there be a shift detected which is to low. To increase this percentage and reduce type II errors the company can reduce sample size and narrow control limits.

4. Optimised Performance vs Efficiency (Dataset 1)

4.1. Objective

The objective of this section of the report is to optimise the number of baristas to balance customer reliability and overall profit. This objective is done using the dataset "timeToServe".

4.2. Method

The dataset concludes of two columns with 200 000 rows of data. The columns represent Barista and Service Time respectively. The barista's column represents the number of baristas working and the service time column represents the number of seconds it takes to serve a customer.

The assumption was that the baristas worked an 8-hour day. Which means that a workday consists of 28 800 seconds. The mean service time was calculated for each amount of baristas. After that is calculated, the amount of services per day is calculated using 28800/Mean Service Time. There after we know that each customer roughly gives us a revenue of R30. As well as each barista costs R1000 per day. Therefore, a revenue can be calculated as was as net profit.

4.3. Results

Baristas <dbl>	MeanService <dbl>	ServicesPerDay <dbl>	Revenue <dbl>	StaffCost <dbl>	NetProfit <dbl>
1	200.15588	143.8879	4316.636	1000	3316.636
2	100.17098	287.5084	8625.253	2000	6625.253
3	66.61174	432.3562	12970.686	3000	9970.686
4	49.98038	576.2261	17286.784	4000	13286.784
5	39.96183	720.6876	21620.629	5000	16620.629
6	33.35565	863.4220	25902.661	6000	19902.661

Table 11

As seen above, the number of records and their respective statistics for each number of baristas are shown. It can be seen that 1 and 2 baristas lead to a long mean service time. Therefore, decreasing the amount of services per day. Which means that their net profit is way to low. This means that 1-2 number of baristas cannot be a solution for this problem. However, the higher number of baristas you have the higher amount of services per day, resulting in a higher net profit regardless of the increase in barista cost.

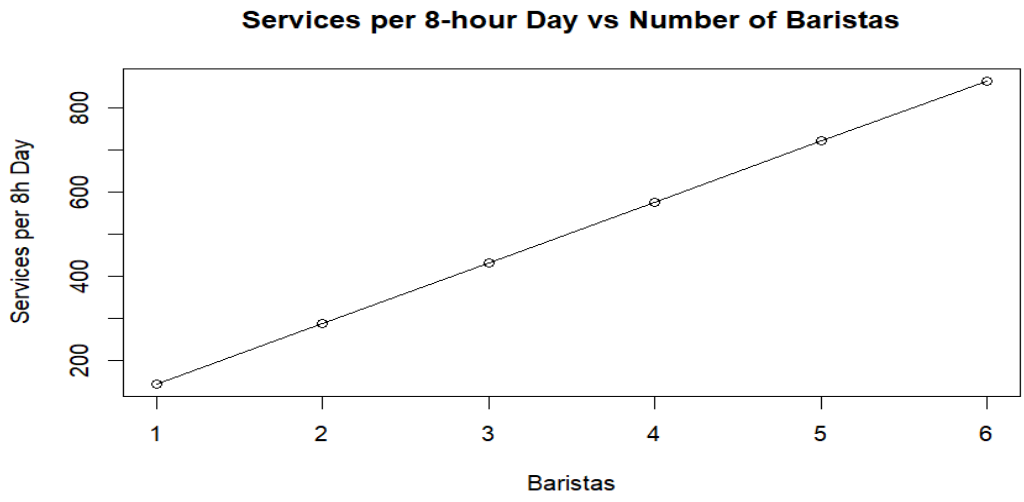


Figure 12

The graph above shows the number of services per day. It shows a general trend of the higher the baristas, the higher the number of services.

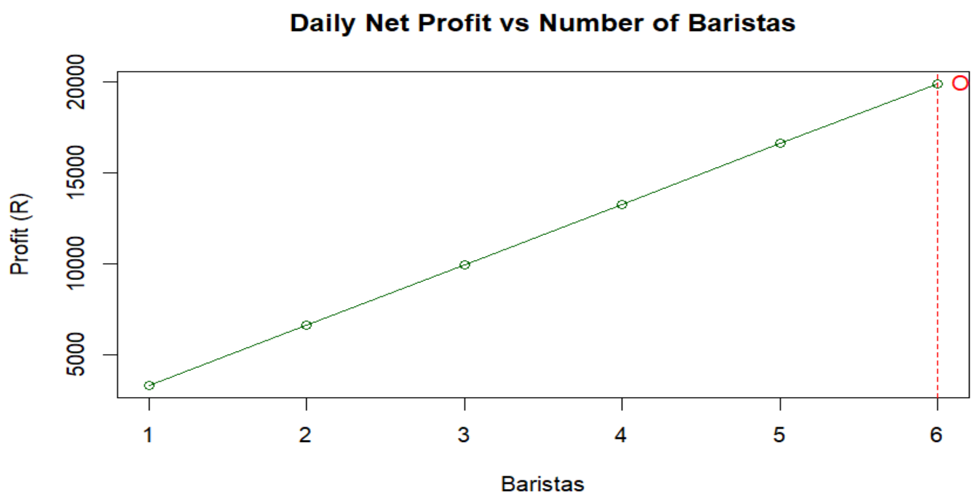


Figure 13

This profit graph shows the more baristas the higher the profit. For this problem 5-6 baristas can be seen as the best options.

Drawing from evidence from the graphs and tables above. The ideal solution for the existing problem is to have the best possible services correlating to the best possible profit. With the assumptions that all baristas are busy all day as there is enough customers which make the baristas fully utilized. The optimal number of baristas would be 6 due to it having the highest net profit and the highest amount of services per day. However, in a real setting especially when it is during a non-peak period. Where demand is finite and not dependant on mean service time. 3-4 baristas would be optimal as staff wouldn't be underutilized. Possibly 5 baristas on a peak day. However, for the dataset provided, this is not the case. Meaning having 6 baristas is optimal for the current dataset conditions.

5. Optimised Performance vs Efficiency (Dataset 2)

5.1. Results

The objective and method is exactly the same as the previous dataset. However, since the dataset is different, the results are very different.

Baristas <dbl>	MeanService <dbl>	ServicesPerDay <dbl>	Revenue <dbl>	StaffCost <dbl>	NetProfit <dbl>
1	200.16894	143.8785	4316.354	1000	3316.354
2	141.51462	203.5125	6105.376	2000	4105.376
3	115.44091	249.4783	7484.348	3000	4484.348
4	100.01527	287.9560	8638.681	4000	4638.681
5	89.43597	322.0181	9660.543	5000	4660.543
6	81.64272	352.7565	10582.695	6000	4582.695

Table 12

As seen above, we can see that services per day between 3-6 baristas are much tighter together. This essentially leads to the revenue being much closer together as well. Therefore, the staff cost has a much bigger impact.

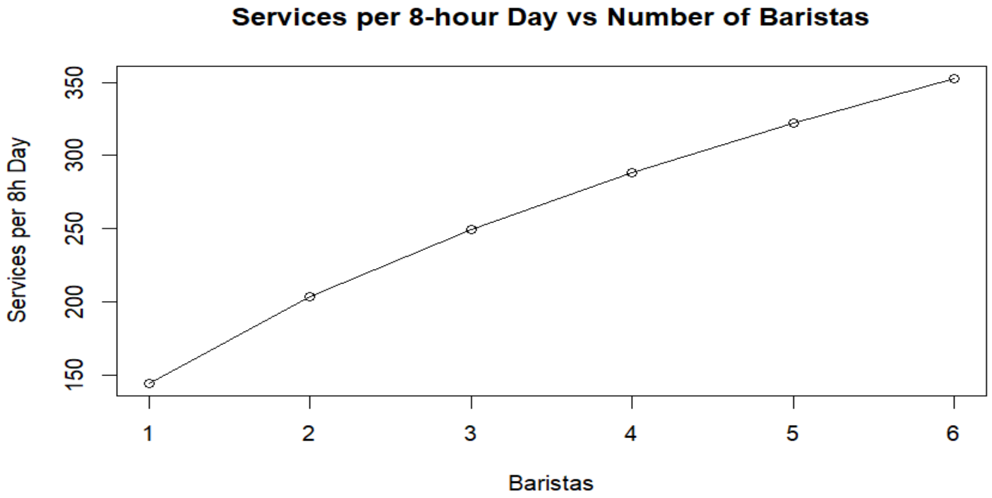


Figure 14

The services per day graph follows a logarithmic distribution compared to the last dataset which follows a straight-line graph. This basically means that this will plateau out eventually when there is to many baristas, however in this case the most number of services per day is still at 6 baristas as expected.

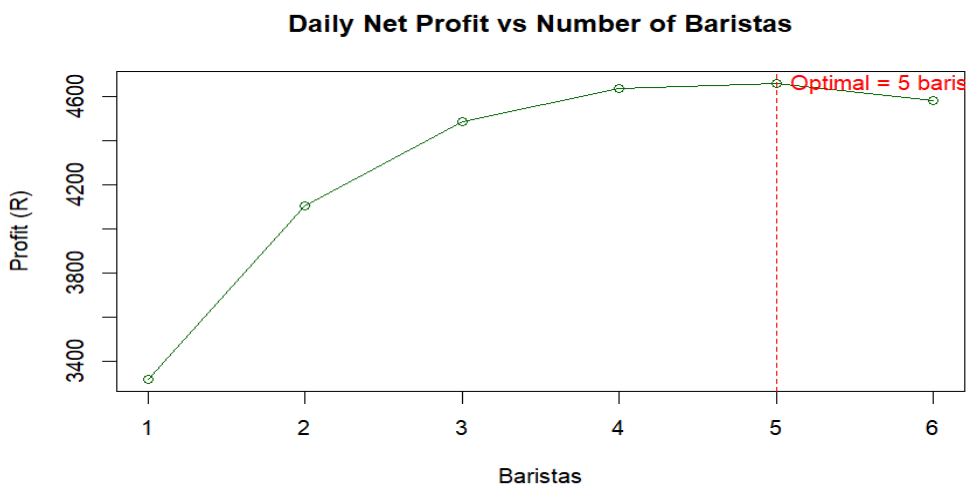


Figure 15

The Daily Net profit graph can be seen above. It as well follows a logarithmic distribution, but the plateau can be seen. This graph is completely different to the previous dataset. The optimal number of baristas in this dataset is 5. As it has the highest daily net profit. As 6 baristas does not create enough revenue compared to 5 to cover the extra barista cost. Therefore, making 5 baristas optimal. In addition, on low demand days 4 baristas could be best as all baristas can still be utilized fully. This dataset shows a more realistic results as sometimes more baristas is not the best. Due to not all baristas being utilized fully. Therefore, paying extra for a barista that is not making enough profit to cover the cost.

6. DOE and Anova

6.1. Objective

The objective is to test if there is a significant difference is delivery time for a specific product type for years 2022 and 2023. As well as for each month. The chosen product was laptops due to it being the companies most important product.

6.2. Method

The "sales2022and2023" database was filtered for laptops. Thereafter, the mean and standard delivery time was computed. Then the ANOVA code was run in R to generate results and graphs were also computed.

6.3. Results

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Order Year	1	19	18.92	0.513	0.474
Residuals	10205	376427	36.89		

Table 13

This was the output of the ANOVA model for the years 2022 and 2023.

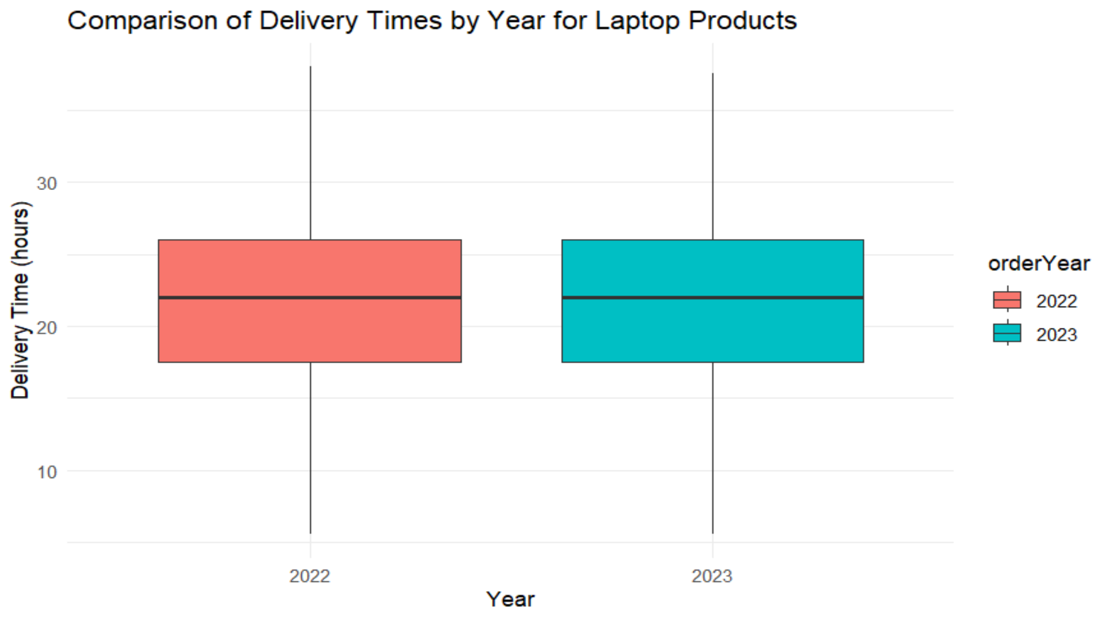


Figure 16

This box plot is a visual representation of the delivery times between 2022 and 2023 of laptops.

The p values calculated for this comparison is 0.474 which is greater than α = 0.05. Which means we reject the null hypothesis. Essentially, this means that there is statistically no difference between 2022 and 2023 as backed up by the visual box plot. Therefore, there is no difference in delivery times for laptops in 2022 and 2023.

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Order Month	11	24380	2216.2	64.18	0.00001525878
Residuals	10195	352066	34.5		

Table 14

This table represents the output of the ANOVA model for the order months.

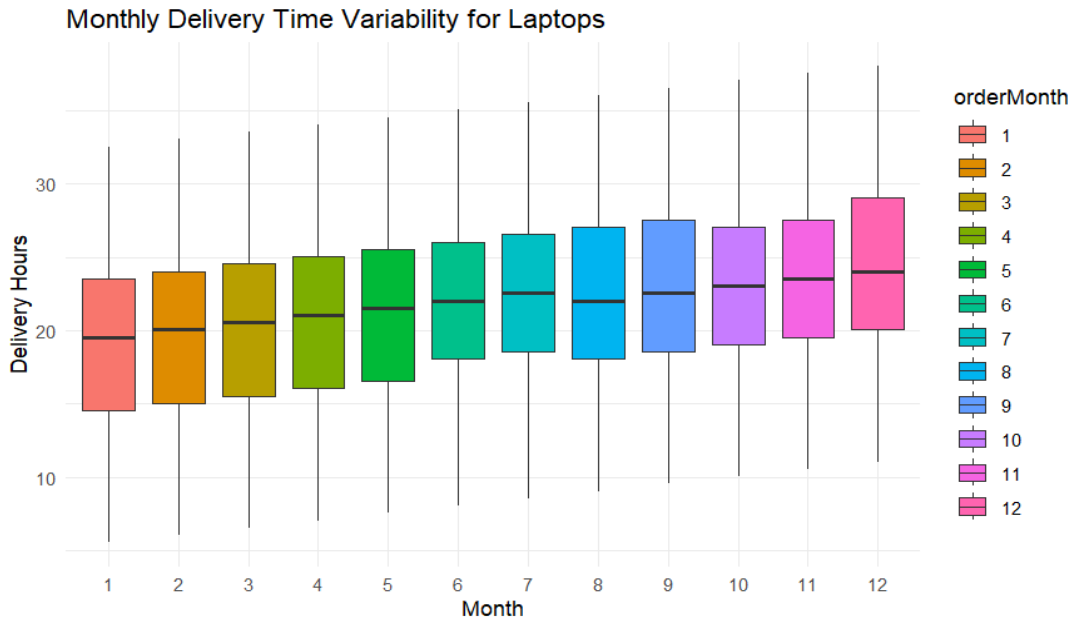


Figure 17

This box plot is a visual representation of the comparison of the mean times for each month.

For comparing the delivery hours for each month, the following was concluded. The F values is extremely large at 64.18 while the p value is very small at 0.00001525878. This means that the results are statistically significant, and we reject the null hypothesis. Meaning that there is at least one month at differs in delivery times. However as seen in the graph above, all the months differ in delivery time. With December being the slowest month. This most likely represents seasonal demand changes or operational changes during slower months. These results shows that capacity planning and logistics optimization need to implement during high demand months.

7. Reliability of Service

7.1. Car rental service

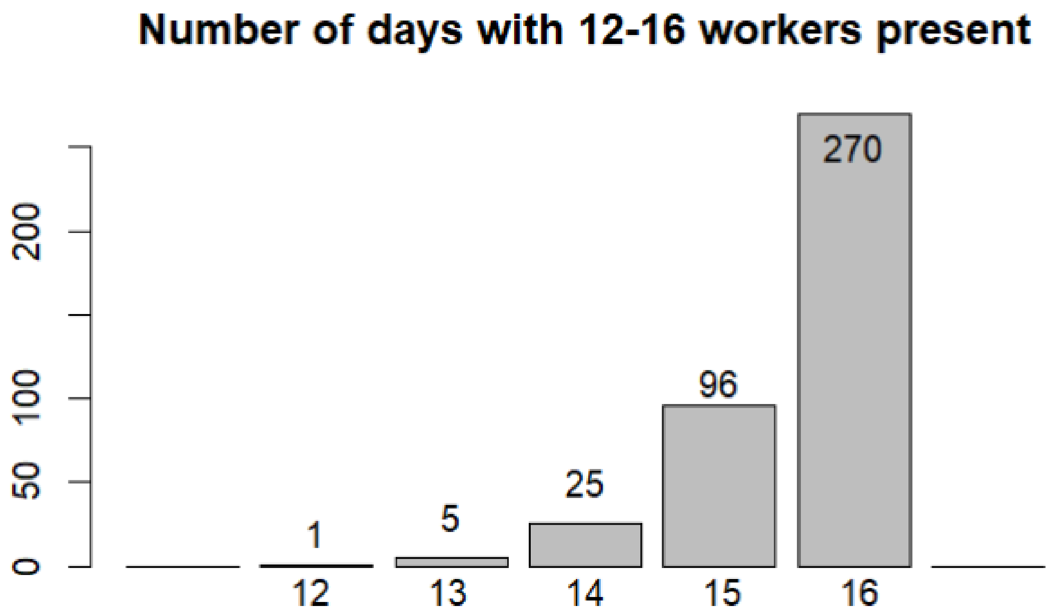


Figure 18

It is assumed that ≥ 15 number of employees ensure an reliable service. Therefore, the total number of days worked by all number of employees is 397. The total number of days worked by ≥ 15 is 366 days. Therefore, offering a reliable service 366/397 which is 92.19 percent of the time.

7.2. Model to Optimise profit

A binomial reliability model was made to optimise the net profit of the data provided in the graph above. The model was used to estimate the amount of "problem days" if there was less than 15 staff working. This would cost the business R20 000 per day. Currently before the model is built the company is experiencing 31 problem days out of the total 397 days, if this is converted into days per year and the revenue according to that is calculated they would lose R570 000 per year. The following results were calculated:

k <int>	problem_days_after <dbl>	annual_loss_after <dbl>	loss_reduction <dbl>	staff_cost <dbl>	net <dbl>
0	31	570025.19	0.0	0e+00	0.00
1	6	110327.46	459697.7	3e+05	159697.73
2	1	18387.91	551637.3	6e+05	-48362.72
3	0	0.00	570025.2	9e+05	-329974.81

Table 15

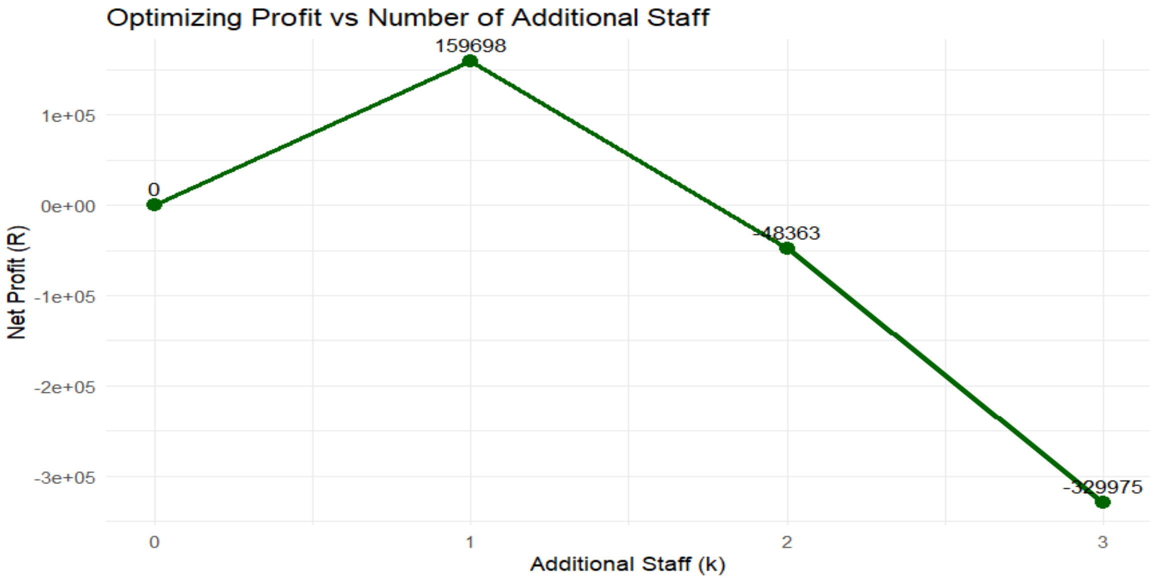


Figure 19

The results show that hiring 1 staff member is the optimal solution to this model. As hiring 1 staff member decreases the problem days to only 6 days. Where is an 80.65 % decrease in problem days. Which in turn increases the business reliability to 98.49%. Furthermore, it decreases the annual loss to R110327.46 whilst increasing the by R159697.73. As hiring more than 1 employee may drop the annual problem days but it significantly decreases the net profits die to salary increases. Therefore, making it the best option for this problem.

Conclusion

This project demonstrates the application of advanced quality assurance, statistical and data analytical techniques to multiple business problems. For the descriptive analysis section, it was very evident that laptops and monitors were the company's bestselling product. Whereas other categories had room for improvement.

SPC analysis showed that no products were fully reliable to meet their delivery times as none of them had a Cpk greater that 1.33. However, some products did have a high reliability. Showing there is much room for improvement in this part of the business.

For risk analysis the probability of Type I errors and Type II errors were calculated. It was seen that there was a very low probability for Type I error, confirming the statistical reliability of those thresholds. However, the probability of the Type II error indicated that the system was sensitive to small process shifts. It shows that their needs to be tighter control limits or increased sampling.

Optimising the barista problem showed two different databases. One which concluded the more baristas the more profit, this database was unrealistic and ignored problems such as shortage of demand. For this database 6 barista was optimal as it was the maximum number of baristas. In comparison, the other database showed that to many baristas comes at an expense as they are not utilized fully and therefore, not generating a profit. For this database, it was concluded that 5 baristas were the optimal solution.

The ANOVA results showed that there is no statistical difference in delivery hours between 2022 and 2023 for laptops. However, the results showed that the delivery hours between months were different. This could be due to various reasons, but it clearly showed that the delivery time in December was significantly higher. Showing that the delivery time is either affected by holidays or some operational problem.

Lastly the reliability of the car services showed that it was optimal for the business to hire an extra employee. This ultimately increase their net profit by R159697.73. Therefore, benefitting the company. Overall, all analysis done had various different conclusions that can benefit the company.

References

1. Bhandari, P. (2021). *Type I & Type II Errors | Differences, Examples, Visualizations*. [online] Scribbr. Available at: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/.
2. ChatGPT (2020). *ChatGPT*. [online] ChatGPT. Available at: https://chatgpt.com/c/68d83fbf-cdf8-8330-87df-f3062fea2d5b [Accessed 20 Oct. 2025].
3. Hayes, A. (2024). *Descriptive statistics: Definition, overview, types, example*. [online] Investopedia. Available at: https://www.investopedia.com/terms/d/descriptive_statistics.asp.
4. Hessing, T. (2019). *X Bar S Control Chart | Six Sigma Study Guide*. [online] Six Sigma Study Guide. Available at: https://sixsigmastudyguide.com/x-bar-s-chart/.
5. Kenton, W. (2025). *What is Analysis of Variance (ANOVA)?* [online] Investopedia. Available at: https://www.investopedia.com/terms/a/anova.asp.
6. R (2016). *R: Getting Help with R*. [online] R-project.org. Available at: https://www.r-project.org/help.html.