ECSA Graduate Attributes Project: Data Analysis Report

Malan Geldenhuys

Student nr: 270 468 77

Date: 24 October 2025

Table of Contents

Introduction:

This report refers to the QA344 2025 ECSA project and focuses on applying statistical methods and techniques for data analysis to practically solve the various problems given. The aim of the project is to use the given data for decision making processes to analyse, monitor, and improve different processes through the use of R programming and solving techniques.

In the first part of the project, descriptive statistics and different visualisations are used to investigate the different customers, products and sales datasets. This then helps to better understand the data and identify any patterns or trends that might be of value.

Part 3 involves using the given delivery times and then applying Statistical Process Control(SPC) using X-bar and S-charts to monitor all of the delivery times for all the different product types. The control limits are set using the first few samples of data and the rest of the samples are tested to identify if and when processes go out of control or given boundaries. The process capability indices (Cp, Cpk, Cpu, and Cpl) are all calculated to check if any of the product types meet the customers' expectations.

In Part 4 the focus of the question is on the risk and error analysis of the data, where Type I and Type II errors are calculated theoretically in order to understand the risks of making any incorrect process control decisions and the impact it has. This section also includes the cleaning of previous datasets and the re-analyses of the data to check for any major differences of the outcomes, that may influence the decision that the company will make.

Part 5 refers to an optimisation problem where data from two different coffee shops are used to determine the optimal number of baristas to ensure the biggest profit, based on service time, cost, and revenue.

In Part 6 of the project an ANOVA or MANOVA is used to test whether or not there are any significant differences in sales between the different datasets such as year 1 vs year 2 or months 1-12. Graphs are included and used to support the results and conclusions made in this question.

The final part 7 looks at the reliability of service at a car rental agency by using probability and cost analysis to estimate how often reliable service can be expected. The optimal amount of staff that should be employed in order to maximise the profit of the company is then determined.

Overall, this project combines statistical analysis with optimisation and process control to solve real-world problems by using the necessary data given. It shows how data can be used to improve the performance of a process, quality of data and efficiency of a company in an industrial environment.

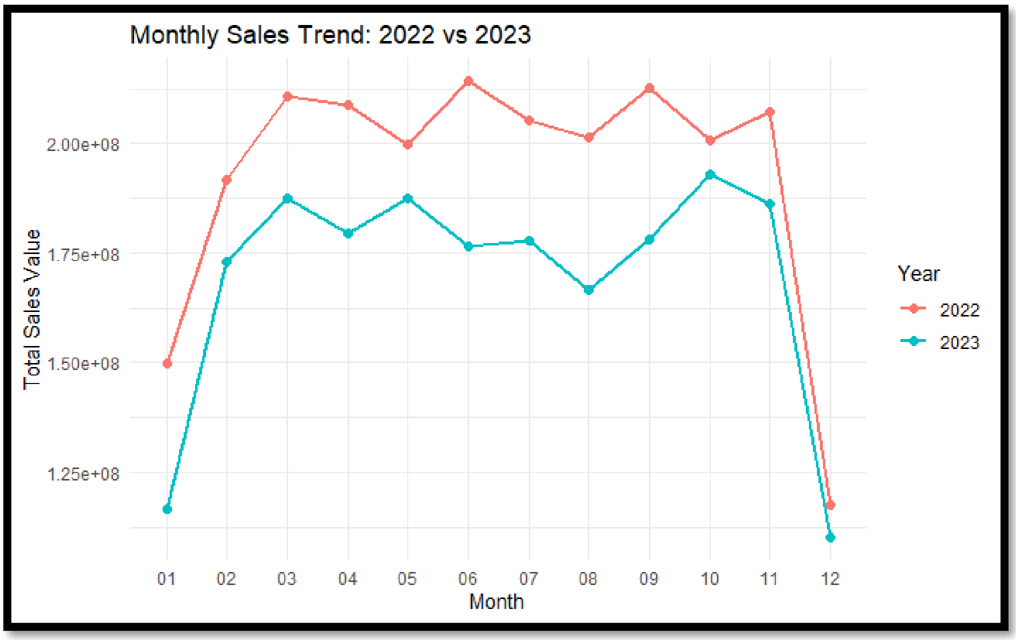
Deliverable 1

Basic Data Analysis 1:

The first data analysis was done on the given sales data through years 2022 and 2023. This graph shows the monthly sales trend over the years by comparing 2022 and 2023 on the same layout. After careful evaluation of the graph, it can be seen that that the overall sales of 2022 were much higher than the number of sales in 2023, with the average difference in sales being 25 million. This indicates a downward trend which is negative for the ROI of the company, I would suggest increasing advertisement of products to boost the sale or run promotions on products that are not as popular anymore.

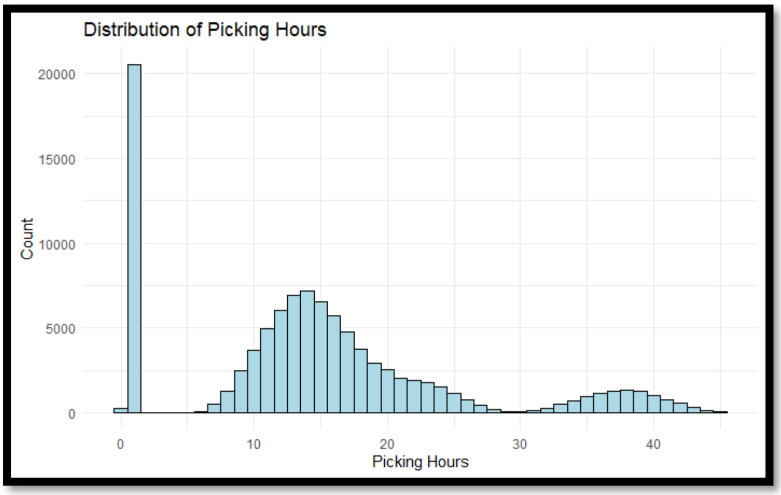
The graph of 2022 shows an incline in sales in the beginning of the year with peaks in sales in June and September. The graph of 2023 indicates peaks in March and May followed by a steep decline and another peak in October. Both of these line plots indicate very low sales in January and February of each year, this can be due to closure of branches or decline in staff over the summer vacation. The company is losing a lot of sales during this period and different options to fix this problem must be assessed.

Option 1 will be to pay staff extra to work over the vacation period or option 2 will be to have better planning of when staff take vacation days to ensure that there will always be sufficient staff to make possible sales. A financial comparison must be made for option 1 to see if it is viable and rules must be set in place to ensure that no one is negatively affected with option 2.

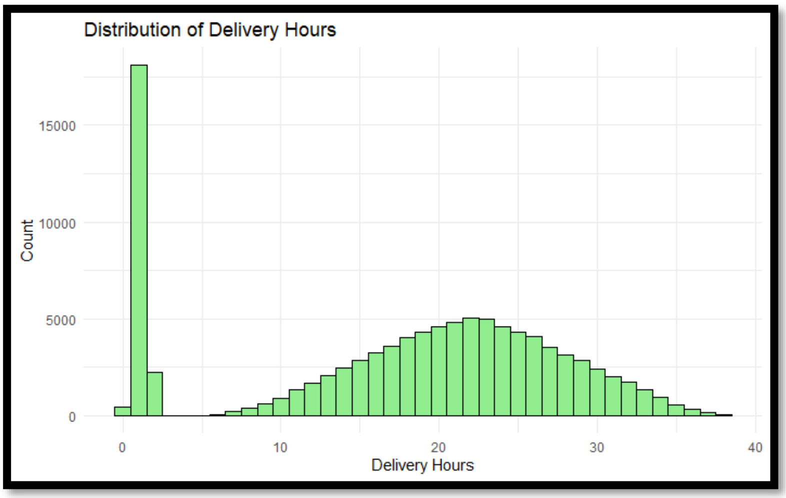


Basic Data Analysis 2:

The second data analysis that was completed is refering to the distribution of the picking hours as well as the delivery hours. The distibution of the picking hours has a very large count of more than 20 000 products being picked within the 1-2 hour mark. This refers to products that are very easily accesible which can include small or automated products. The rest of the graph has a bimodal distribution with the largest peak being between 13-14 hours and a much smaller peak between 35-40 hours. The peak at 13 hours refers to the average time it takes to pick the orders whereas the second peak can refer to some unusually large or complicated orders. This indicates that the warehouse is fairly consistent with a few outliers that can be flagged as exceptional orders. Better systems can be implemented to decrease the picking hours for more complicated orders to move it closer to the mean of the faster picking products.



The delivery distribution graph also has a very large peak at 1-2 hours with a count of more than 17500 thus indicating that products were delivered fast. This fast delivery can only be explained by local deliveries or if the company has a same-day delivery policy for certain products. The rest of the graph has a normal distribution with the mean delivery time being between 22 and 23 hours and the rest of the deliveries being fairly spread around the mean. The delivery hours has fewe extreme delays compared to the picking hours. One improvement that can be made for the delivery is to shift the entire normal distribution to the left to decrease the overall delivery time.

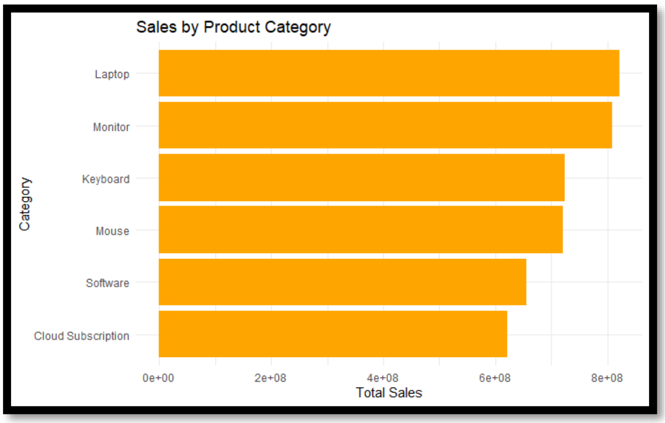


Basic Data Analysis 3:

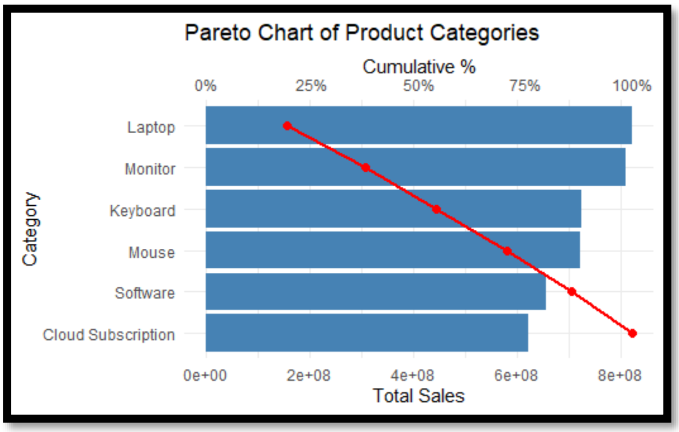
The third data analysis that was completed is a basic horizontal histogram representing the total sales of each product. After closer inspection it can be seen that laptops are the most popular product with monitors being a very close second. There is a bigger gap between the sales of these 2 products and the sales of keyboard and mouses which are both close to each other with about 660 million of each product sold.

If the company is struggling to sell products, like in 2023, certain products can be sold together as a package to make it easier for customers to buy and thus increase overall sales. For example, laptops and monitors or keyboards and mouses can be sold as bundle deals.

Software is the second worst overall seller with cloud subscriptions being sold the least of all the products, indicating that the customers prefer to buy hardware from the company rather than digital products. The sales of digital products can be increased by implementing better marketing strategies and introducing value-adding bundling like buying a laptop with 2 years of cloud storage.



In the Pareto Chart of Product categories is can be seen the cumulative % of products are very close to each other with the cumulative line having almost a perfectly straight line with constant gradient. This indicates that there is not a very large difference between the total amount of sales for each of the products. Laptops and monitors contributes the most with +- 20% each while products like keyboard and mouses contributed +- 17% each and software and cloud subscription only contributing +- 12% each.

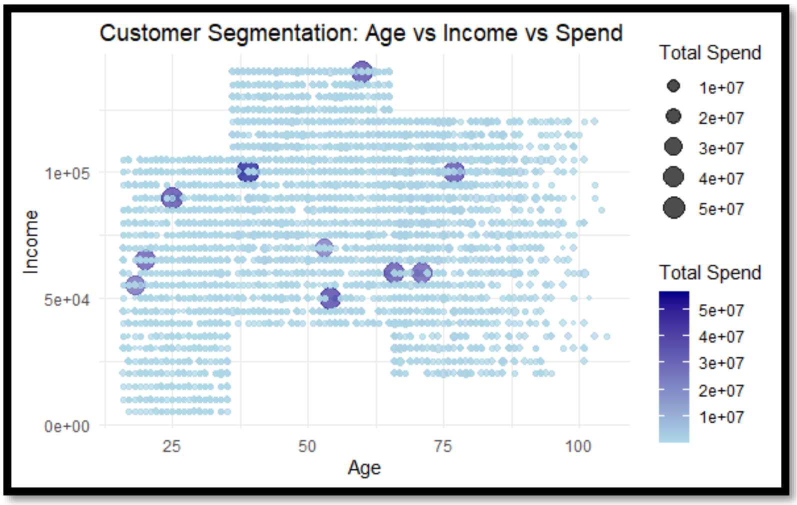


Basic Data Analysis 4:

The fourth data analysis takes a number of factors into account to provide valuable information that can be used to improve advertisement and to determine biggest target market. The graph below shows the segmentation of customers in terms of age, income and the amount that they are willing to spend on products. The following can be interpreted from the graph: customers below 35 has the lowest income while customers between the age of 35 and 67 earns the most and customer above 67 earns the second most.

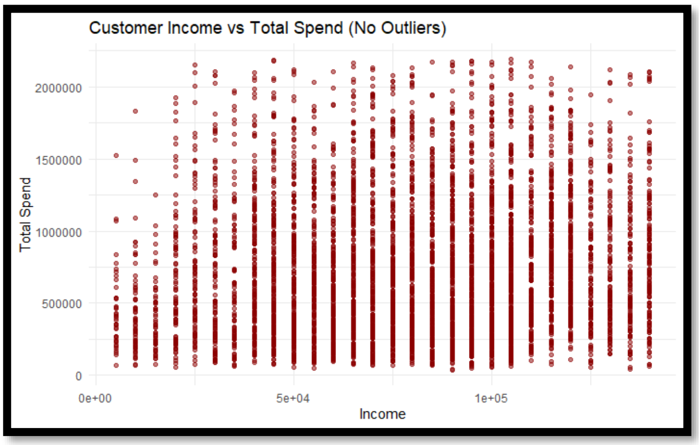
The size and intensity of the blue circles represent the total amount each customer spent. Larger, darker circles indicate higher total spending. The analysis suggests that customers with higher incomes tend to spend more on products, as they have greater purchasing power and may require high-quality equipment for professional use.

A noticeable number of products are also sold to customers below 25 likely driven by the fact that the new generation has a higher interest in gaming/electronic equipment or because they need it for tertiary studies. There is also a spike in sales for customers between 63 and 75 years old which ca be because of people still running consulting businesses or retirees seeking to stay connected with their loved ones through digital technology.

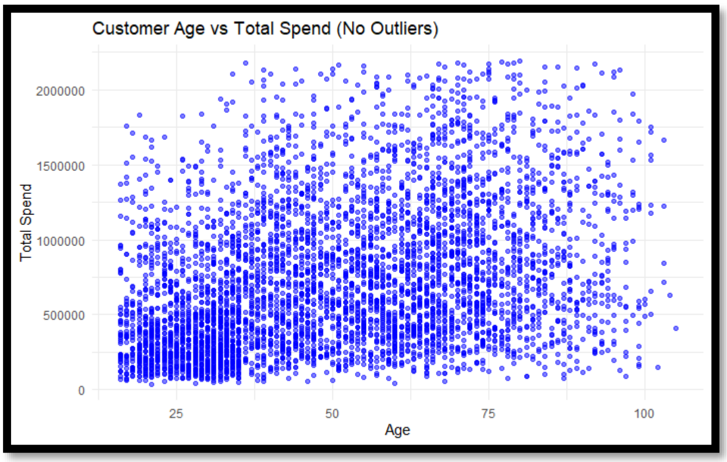


Basic Data Analysis 5:

The analysis shows that customers with higher incomes tend to spend more on products of the company. Whereas lower income levels (less than 50000), the spending is concentrated in smaller amounts, mostly under 600000. However, as income increases, particularly above 100000, there is an upward trend and higher levels of spending, often exceeding 1000000 spend on products. This trend indicates that customers with greater income are not only capable of purchasing but also more actively do so, likely investing in higher-value or good quality electronic products. This suggests that high-income customers form a strong target audience for premium offerings or even extended warranties, and loyalty programs that emphasize quality and exclusivity.



The relationship between customers age and their total spending clearly reveals patterns across customer purchases. The largest concentration of spending occurs among customers aged 25 to 50, making this group the biggest consumer interval even though they do not spend as much on single purchases. Customers below 25 also contribute significantly, which is likely driven by interest in gaming, student needs, or technology for tertiary studies. Interestingly, there is also a noticeable increase in spending among customers aged 63 to 75, which may be explained by individuals still active in consulting businesses, higher disposable income in retirement, or the desire to stay connected with family. This indicates that while the main focus of the company should remain on the 25–50 age group, targeted campaigns towards students and older adults can open up additional growth opportunities, especially when products are positioned for affordability and reliability.



Deliverable 2

3.1 Data Preparation and Sampling

The raw sales data from the 2026 and 2027 excel file was first loaded and inspected to make sure that all the required columns were present. Numeric columns, like Quantity and deliveryHours, were converted to the appropriate data types and rows that contained any missing values in these columns, were removed. An additional column was created with the name orderDate by combining the year, month and day whereafter the data was sorted by this date as well as the orderTime. For the SPC analysis, the delivery hours of each product were divided into fixed size samples to form the base for the X- and s-chart calculations.

3.2 Control Limits Calculation

For each of these products, the initial set of 30 samples were used to calculate the values of the control limits. The CL-central line of the X-bar chart was calculated as the mean of all the sample means and the control limits at 1, 2 and 3 standard deviations were determined using the standard error of the mean. Similarly, the s-chart line was the mean of all the sample standard deviations while the control limits were derived from the statistical formulas. All of these calculations provided the reference points for detecting the variability and special causes in the various samples.

3.3 Monitoring and Special Cause Detection

By using the control limits calculated in question 3.2, all of the samples could be plotted on X-bar and s-charts. By analysing these charts, any unusual patterns or potential special cases can be identified by making use of the following questions.

Question A focusses on the samples where the standard deviation exceeds the 3 standard deviation s-chart limit.Question B highlights the samples that signifies good control by having the longest consecutive run of samples between the -1 and +1 s-chart limits, while question C identifies any sequences of 4 consecutive X-bar samples that are outside the 2 standard deviations limits.

The process capability analysis were performed using the first 1000 deliveries for each product type, with provided specification limits of 0 to 32 hours for delivery time. The indices Cp, Cpu, Cpl, and Cpk was calculated for each ProductID. Product types with Cpk >= 1.33 were considered capable of meeting the customers’ requirements (VOC) while values between 1 and 1.33 are declared as marginally stable.

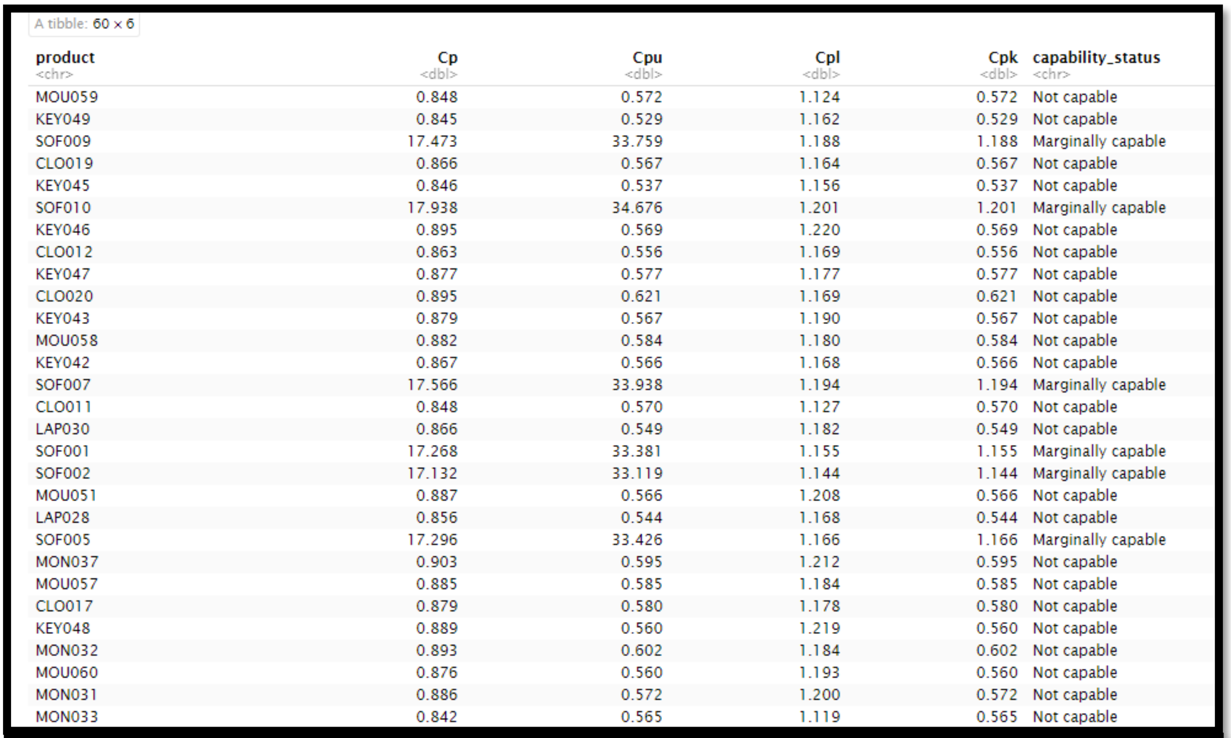
The results show that while some products demonstrated high consistency and capability with Cpk > 1.33, while others exhibited very large variability or mean shifts toward the specification limits thus indicating process improvement opportunities. The following calculations was used to determine all the indices:

$$Cp = \frac{USL - LSL}{6\sigma} = \frac{32-0}{6\times 4} = 1.33$$

$$Cpl = \frac{\bar{x} - LSL}{3\sigma} = \frac{20-0}{3\times 4} = 1.67$$

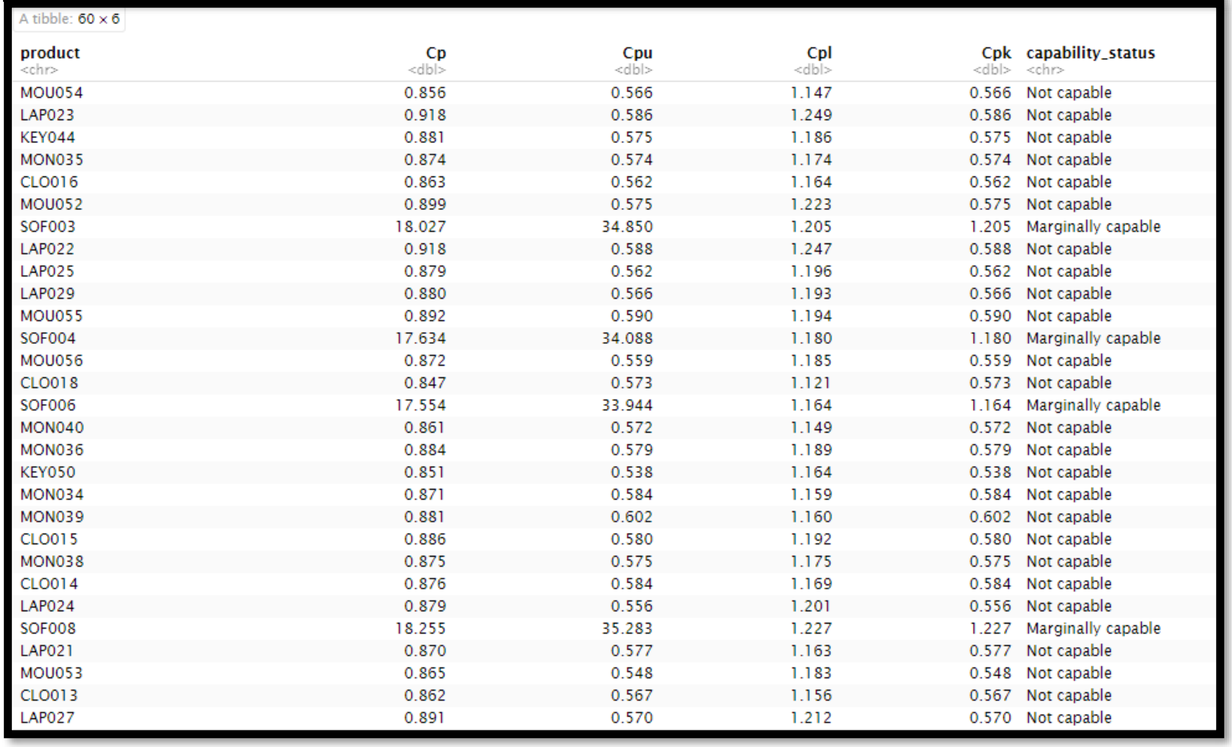
$$Cpu = \frac{USL - \bar{x}}{3\sigma} = \frac{32-20}{3\times 4} = 1$$

$$Cpk = \min(Cpu, Cpl) = \min\,(1.67, 1) = 1$$



A tibble: 60 × 6

product	Cp	Cpu	Cpl	Cpk	capability_status
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
MOU059	0.848	0.572	1.124	0.572	Not capable
KEY049	0.845	0.529	1.162	0.529	Not capable
SOF009	17.473	33.759	1.188	1.188	Marginally capable
CLO019	0.866	0.567	1.164	0.567	Not capable
KEY045	0.846	0.537	1.156	0.537	Not capable
SOF010	17.938	34.676	1.201	1.201	Marginally capable
KEY046	0.895	0.569	1.220	0.569	Not capable
CLO012	0.863	0.556	1.169	0.556	Not capable
KEY047	0.877	0.577	1.177	0.577	Not capable
CLO020	0.895	0.621	1.169	0.621	Not capable
KEY043	0.879	0.567	1.190	0.567	Not capable
MOU058	0.882	0.584	1.180	0.584	Not capable
KEY042	0.867	0.566	1.168	0.566	Not capable
SOF007	17.566	33.938	1.194	1.194	Marginally capable
CLO011	0.848	0.570	1.127	0.570	Not capable
LAP030	0.866	0.549	1.182	0.549	Not capable
SOF001	17.268	33.381	1.155	1.155	Marginally capable
SOF002	17.132	33.119	1.144	1.144	Marginally capable
MOU051	0.887	0.566	1.208	0.566	Not capable
LAP028	0.856	0.544	1.168	0.544	Not capable
SOF005	17.296	33.426	1.166	1.166	Marginally capable
MON037	0.903	0.595	1.212	0.595	Not capable
MOU057	0.885	0.585	1.184	0.585	Not capable
CLO017	0.879	0.580	1.178	0.580	Not capable
KEY048	0.889	0.560	1.219	0.560	Not capable
MON032	0.893	0.602	1.184	0.602	Not capable
MOU060	0.876	0.560	1.193	0.560	Not capable
MON031	0.886	0.572	1.200	0.572	Not capable
MON033	0.842	0.565	1.119	0.565	Not capable

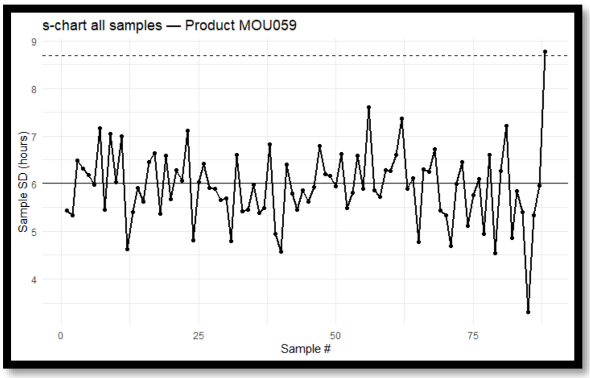


A tibble: 60 × 6

product	Cp	Cpu	Cpl	Cpk	capability_status
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
MOU054	0.856	0.566	1.147	0.566	Not capable
LAP023	0.918	0.586	1.249	0.586	Not capable
KEY044	0.881	0.575	1.186	0.575	Not capable
MON035	0.874	0.574	1.174	0.574	Not capable
CLO016	0.863	0.562	1.164	0.562	Not capable
MOU052	0.899	0.575	1.223	0.575	Not capable
SOF003	18.027	34.850	1.205	1.205	Marginally capable
LAP022	0.918	0.588	1.247	0.588	Not capable
LAP025	0.879	0.562	1.196	0.562	Not capable
LAP029	0.880	0.566	1.193	0.566	Not capable
MOU055	0.892	0.590	1.194	0.590	Not capable
SOF004	17.634	34.088	1.180	1.180	Marginally capable
MOU056	0.872	0.559	1.185	0.559	Not capable
CLO018	0.847	0.573	1.121	0.573	Not capable
SOF006	17.554	33.944	1.164	1.164	Marginally capable
MON040	0.861	0.572	1.149	0.572	Not capable
MON036	0.884	0.579	1.189	0.579	Not capable
KEY050	0.851	0.538	1.164	0.538	Not capable
MON034	0.871	0.584	1.159	0.584	Not capable
MON039	0.881	0.602	1.160	0.602	Not capable
CLO015	0.886	0.580	1.192	0.580	Not capable
MON038	0.875	0.575	1.175	0.575	Not capable
CLO014	0.876	0.584	1.169	0.584	Not capable
LAP024	0.879	0.556	1.201	0.556	Not capable
SOF008	18.255	35.283	1.227	1.227	Marginally capable
LAP021	0.870	0.577	1.163	0.577	Not capable
MOU053	0.865	0.548	1.183	0.548	Not capable
CLO013	0.862	0.567	1.156	0.567	Not capable
LAP027	0.891	0.570	1.212	0.570	Not capable

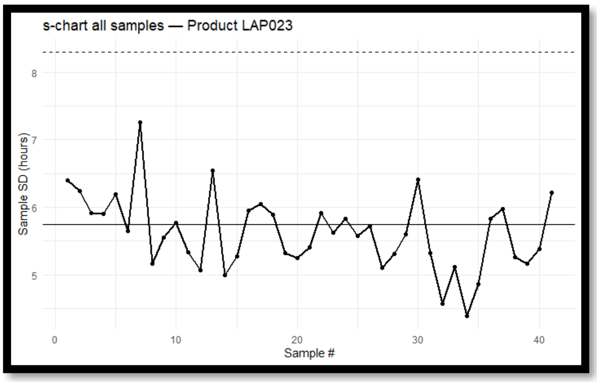
Question A:

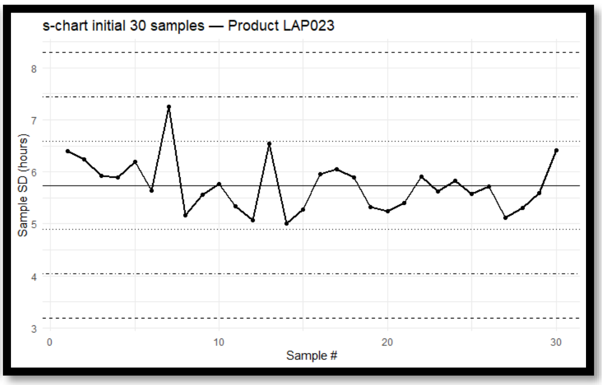
This question focuses on identifying samples with extreme variability, indicated by samples where the standard deviation exceeds the +3 upper sigma limit on the s-chart for all of the samples. This rule indicates a special cause of variation, thus indicating that the process may be experiencing temporarily unusual events. There was only one s-chart identified with a sample above the +3 upper sigma limit and the product identified is MOU059.



Question B:

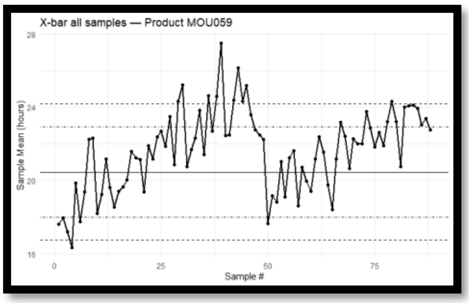
For this question the different charts were examined to find the longest consecutive run of s-chart points within +-1 sigma of the mean. While all points may still technically be in control, a chart with a lot of samples within these narrow limits may indicate very low variation, which can be due to the over-adjustment or constraining of processes. The s-chart with the longest consecutive point was identified as product LAP023 with a total of 24 point within the given area. The +- 1 sigma lines can be observed in the graph on the right-hand side while the chart on the left indicates all the samples of the product. This highlights areas where the process lacks any natural variability and may require further investigation.

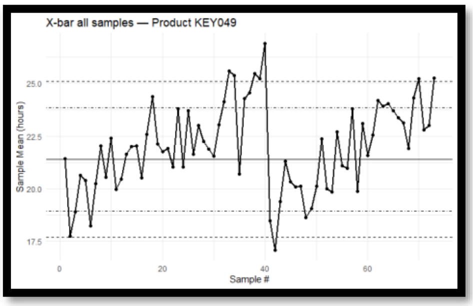


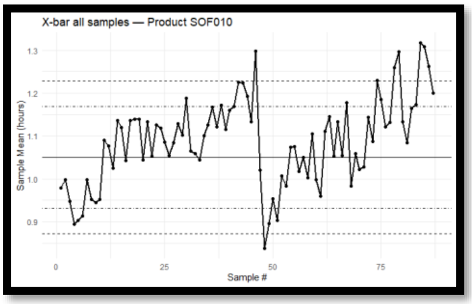


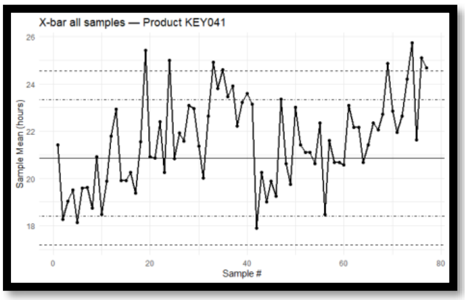
Question C:

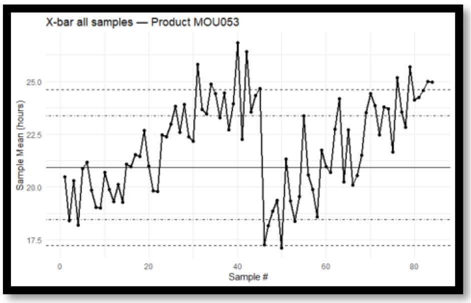
In this question we were asked to look for 4 consecutive X-bar points outside of the +-2 sigma control limits. Although points that are outside of the +-3 sigma are strong signals, a sequence beyond the +-2 sigma can identify any systematic shifts in the process mean, which can thus affect the overall delivery reliability. This helps to identify any trends before they result in larger quality issues that may negatively impact a company. In total 33 products were identified where the first 3 products include MOU059, KEY049, SOF010 and the last 3 products are KEY04, SOF008, MOU053.

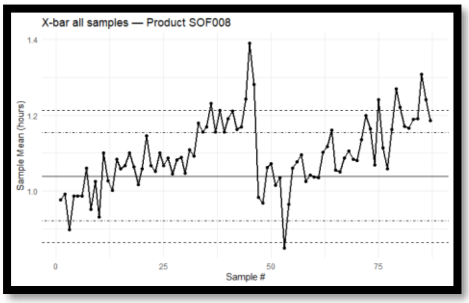












Deliverable 3:

4.1 Type I error

The probability of one sample being above the centreline is 0.5. This is because the standard normal distribution is symmetric, thus meaning half of the values are above the mean or centreline. It then follows that the probability of 7 consecutive samples above the centreline is simply 0.5^7=0.0078125 which is a Type I error because it indicates how often it would wrongly detect an out-of-control sigma.

A: The probability that a single sample exceeds +3 sigma.

$$P(Z > 3) = 1 - \theta(3) \; = \; 0.0013499 \; = \; 0.135\%$$

This indicates that under Ho, about 0.135% of the samples will exceed the upper 3 sigma limit by chance. The per sample false-alarm probability will be found if every single +3 sigma point is flagged.

B: The probability that 4 consecutive samples will lie within +- 1 sigma.

1 sample: $P(-1 < Z < 1) = \theta(1) - \theta(-1) = 0.6826894921$

k samples: $P(k\ consecutive\ in\ [-1,1]) = (0.6826894921)^\wedge k$

$$k \; = \; 3 \; \rightarrow \; 0.682689^3 = 0.31818 = 31.82\%$$

$$k \; = \; 5 \; \rightarrow \; 0.14829 = 14.83\%$$

$$k \; = \; 7 \; \rightarrow \; 0.06911 = \; 6.91\%$$

$$k \; = \; 10 \; \rightarrow \; 0.02199 = 2.20\%$$

These probabilities indicates that about 68% of all single samples will fall inside +- 1 sigma. The longer the number of consecutive samples, within +- 1 sigma, the less likely the probability will be that it will happen. For example, 10 consecutive samples will have a probability of 2.2% whereas the probability for 3 consecutive samples is much higher at 31.82%. Thus, meaning that short runs are more common than longer runs.

C: Probability of 4 consecutive X-bar samples outside of +2 sigma.

1 sample: $P(Z > 2) = 1 - \theta(2) = 0.0227501319 = 2.275\%$

4 samples: $P = \big(1 - \theta(2)\big)^\wedge 4 = (0.0227501319)^4 = 2.6788 \times 10^{-7} = 0.00002679\%$

These values indicate that it is extremely unlikely that there will be 4 consecutive sample means above the upper 2 sigma line. Due to this being very unlikely, one can argue that it is a very strong indicator of an out-of-control shift when observed.

4.2 Type II Error

Centreline = 25.05 litres

UCL = 25.089 litres

LCL = 25.011 litres

True mean fill volume μ1= 25.028

New standard deviation σ = 0.017

$$\beta = P(LCL < X < UXL)\ where\ \mu = \mu 1$$

$$z = \frac{x - \mu}{\sigma}$$

$z_{UCL} = \frac{25.089 - 25.028}{0.017} = \frac{0.061}{0.017} = 3.5882$

$z_{LCL} = (25.011 - 25.028)/0.017 = -0.017/0.017 = -1.0$

$\theta(z_{LCL}) = \theta(-1) = 0.158655$

$\theta(z_{UCL}) = \theta(3.5882) = 0.999834$

$\beta = \theta(3.588235) - \theta(-1.000) = 0.999834 - 0.158655 = 0.8412 = 84.12\%$

$1 - \beta = 0.15582 = 15.88\%$

These probabilities indicates that there is about an 84% chance that the chart will fail to detect this shift of the true mean thus leading to poor sensitivity. The power of the chart is very low coming in at only 15.88%, so it is unlikely to flag the small mean shift. To increase accuracy the sample size must be increased in order to reduce the standard deviation.

4.3 Updated data Analysis

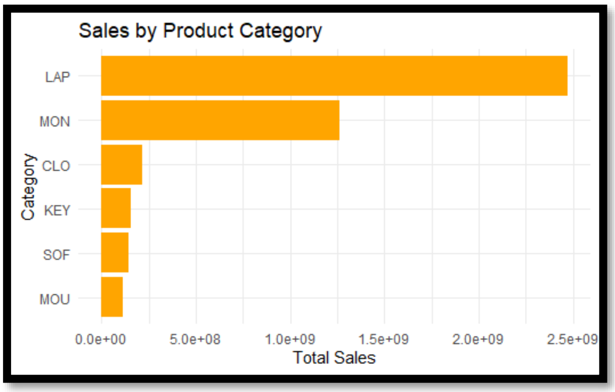
For this question the original product and head-office datasets were found to contain inconsistencies in ProductIDs, selling prices, and markup values. To address this, the productsHeadoffice.csv file was corrected by updating the ProductIDs and ensuring that the selling prices and markup values repeated consistently every ten rows per product type based on the authoritative local productsdata.csv.

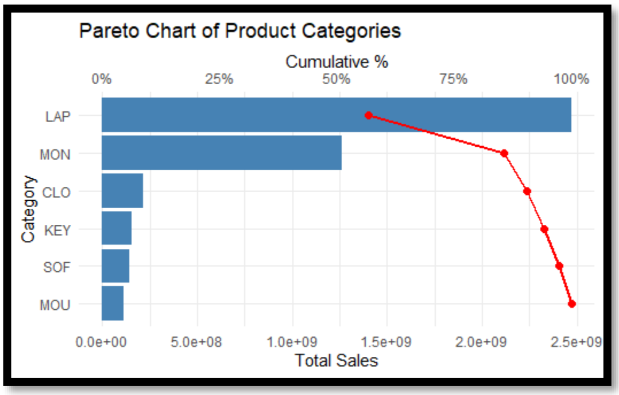
Additionally, the category column in the productsdata.csv file was updated to correspond correctly with each ProductID. The corrected datasets was then saved as productsHeadoffice2025.csv and productsdata2025.csv, and the basic data sales analysis was calculated again using these updated files to assess any differences in the outcomes.

Most of the analysis graphs remained largely unchanged, with only a few notable exceptions showing significant differences. One such graph is the Sales by Product Category, which changed considerably after the necessary corrections were applied to the dataset. In the updated graph below, it is clear that the overall sales of laptops and monitors now dominate, far exceeding those of other product categories. In comparison with the previous dataset that showed a more evenly distributed sales across the various products, with an average sales value of approximately 7e+08. This indicates that the original data provided a misleading picture of product popularity, which could have led to incorrect business decisions and poor planning for future inventory and marketing strategies.

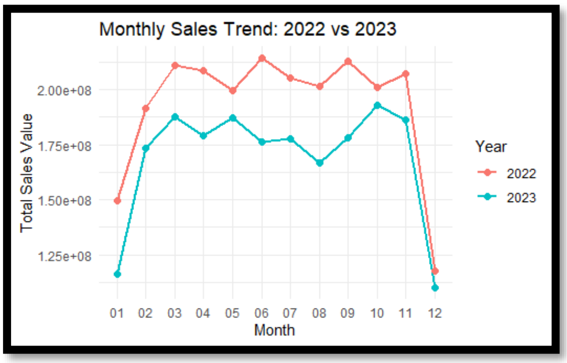
In the Pareto Chart of Product Categories, it is clear that the cumulative percentage distribution has changed significantly after the dataset corrections. Laptops now account for more than 50% of total sales, while monitors contribute roughly 25%, making them the two dominant product categories. This can be seen from the cumulative percentage line, meaning a smaller gradient indicates a gradual contribution between consecutive products, whereas a steep gradient reflects a large gap in sales contribution.

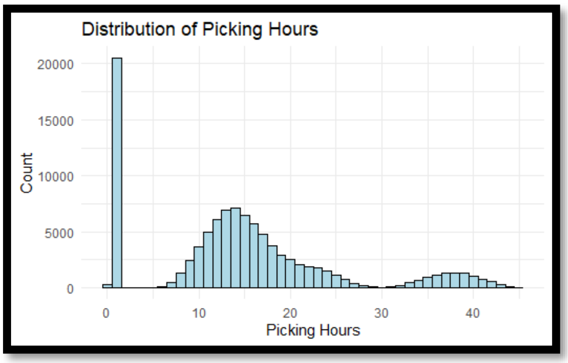
The sharp change in gradient after monitors highlights the large difference between these two categories and the remaining products. Compared to the previous dataset, which suggested that sales were mostly evenly distributed across products, the updated chart now clearly identifies the top-selling items. This correction provides a more accurate view of popular products within the company and helps to prevent misleading conclusions that could lead to overstocking or poor inventory planning.

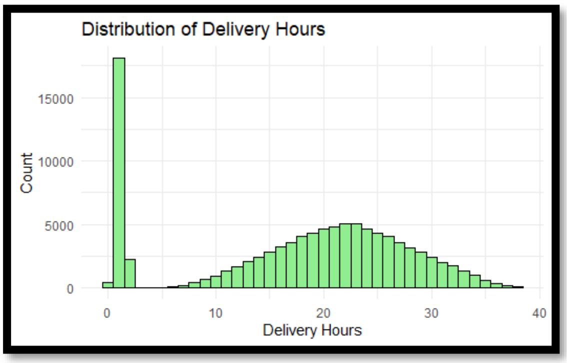


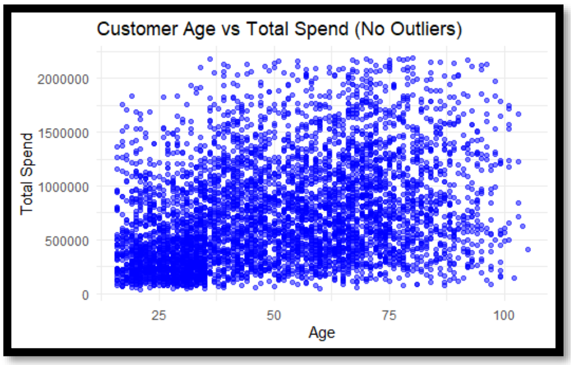


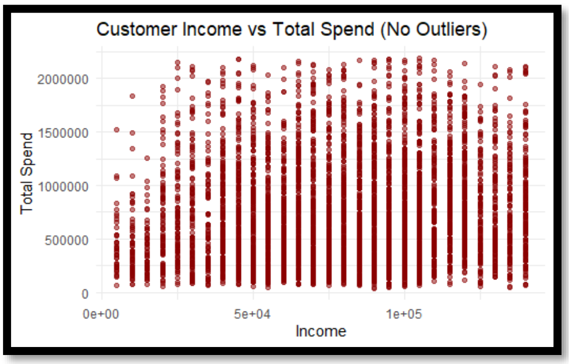
Most of the graphs produced from the corrected dataset remained consistent with the original analysis created from the dataset with errors. These include visualisations such as Sales by Month, Distribution of Picking Hours, Distribution of Delivery Hours as well as Age vs Income Spend, all of which showed minimal or no changes in their overall trends and patterns. This indicates that while some product-specific data were affected by the corrections, the general sales performance and seasonal trends of the business remained stable across both datasets. The following graphs stayed consistent throughout the different datasets:

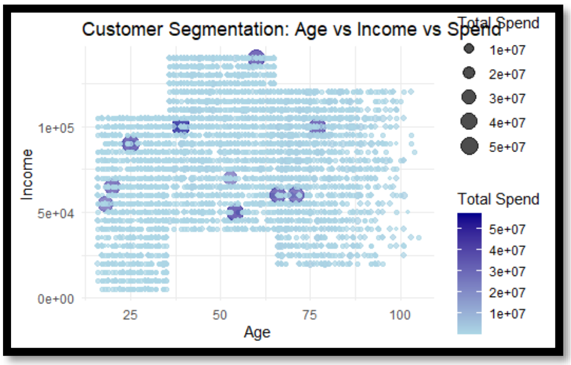












5. Optimisation of Total Baristas

To determine the optimal number of baristas required to maximise the profit of the coffee shop, the following procedure was followed. The provided dataset, representing the one year of sales, was analysed to explore the relationship between the number of baristas currently working and the average time in seconds taken to serve a customer. The mean service time for each total of baristas was calculated while assuming a standard 9-hour workday.

By using this information, the total daily income was calculated by dividing the total available work time by the average service time and then multiplying the result by the profit earned per customer which is given as R30. Since the staffing cost was R1 000 per barista per day, the total daily profit was determined by subtracting total staff costs from the total income.

Coffee Shop 1:

$$Total\ Profit = (Customers\ \times Customer\ profit) - (Total\ baristas \times Cost\ per\ barista)$$

$$Profit\ 1\ Barista = \left(\frac{9 \times 60 \times 60}{200.156} \times 30\right) - (1 \times 1000) = 3856.21$$

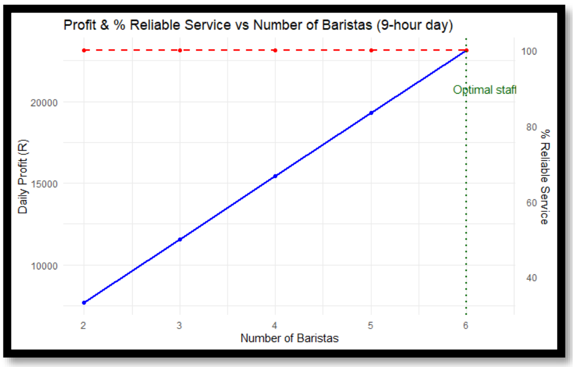
$$Profit\ 2\ Barista = \left(\frac{9 \times 60 \times 60}{100.171} \times 30\right) - (2 \times 1000) = 7703.41$$

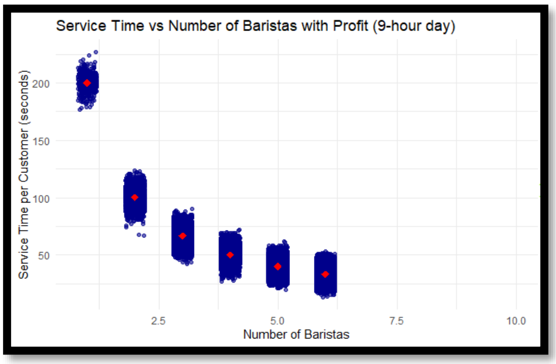
$$Profit\ 3\ Barista = \left(\frac{9 \times 60 \times 60}{66.613} \times 30\right) - (3 \times 1000) = 11591.75$$

$$Profit\ 4\ Barista = \left(\frac{9 \times 60 \times 60}{49.98} \times 30\right) - (4 \times 1000) = 15447.78$$

$$Profit\ 5\ Barista = \left(\frac{9 \times 60 \times 60}{39.96} \times 30\right) - (5 \times 1000) = 19324.32$$

$$Profit\ 6\ Barista = \left(\frac{9 \times 60 \times 60}{33.356} \times 30\right) - (6 \times 1000) = 23140.18$$





<u>Coffee Shop 2:</u>

$$Total\ Profit = (Customers\ \times Customer\ profit) - (Total\ baristas \times Cost\ per\ barista)$$

$$Profit\ 1\ Barista = \left(\frac{9 \times 60 \times 60}{200.16894} \times 30\right) - (1 \times 1000) = 3855.90$$

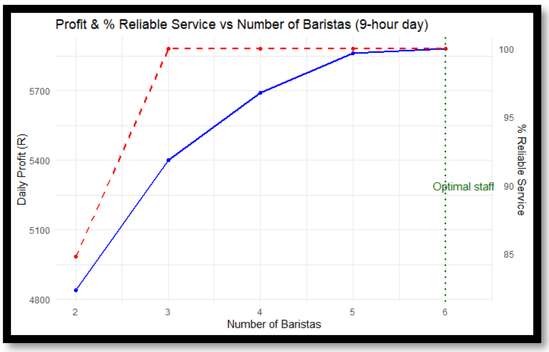
$$Profit\ 2\ Barista = \left(\frac{9 \times 60 \times 60}{141.51462} \times 30\right) - (2 \times 1000) = 4868.55$$

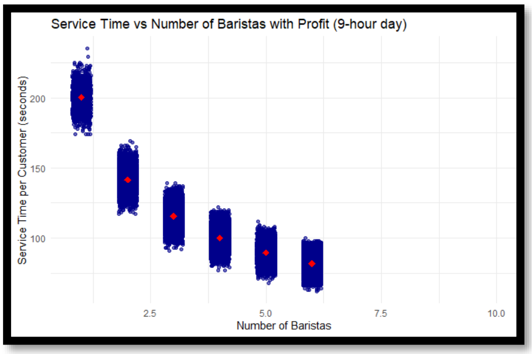
$$Profit\ 3\ Barista = \left(\frac{9 \times 60 \times 60}{115.44091} \times 30\right) - (3 \times 1000) = 5419.89$$

$$Profit\ 4\ Barista = \left(\frac{9 \times 60 \times 60}{100.01527} \times 30\right) - (4 \times 1000) = 5718.52$$

$$Profit\ 5\ Barista = \left(\frac{9 \times 60 \times 60}{89.43597} \times 30\right) - (5 \times 1000) = 5868.11$$

$$Profit\ 6\ Barista = \left(\frac{9 \times 60 \times 60}{81.64272} \times 30\right) - (6 \times 1000) = 5905.53$$





After completing the necessary calculations and analysis for the **first coffee shop** it can be seen that the profit increased notably with the increase of baristas thus peaking at approximately R23 140.18. The similar calculations and analysis were performed on the **second coffee shop** and it can be observed that the profit gap was very large with the initial increase of baristas but then grew smaller with the increase in baristas, thus peaking at approximately R5905.53. Based on both calculations, the optimal staffing level for maximum profitability at both of the coffee shops are six baristas per day. Above all the tables of comparison along with their calculations can be seen as well as the Service Time vs Baristas graphs for both of the coffee shops.

Coffee Shop 1		
Total Baristas	Average time (seconds)	Profit (Rand)
1	200.15588	3856.21
2	100.17098	7703.41
3	66.61174	11591.75
4	49.98038	15447.78
5	39.96183	19324.32
6	33.35565	23140.18

Coffee Shop 2		
Total Baristas	Average time (seconds)	Profit (Rand)
1	200.16894	3855.90
2	141.51462	4868.55
3	115.44091	5419.89
4	100.01527	5718.52
5	89.43597	5868.11
6	81.64272	5905.53

Deliverable 4

6. DOE and ANOVA

In order to evaluate whether there were any significant monthly changes in the delivery time performance of each product type, an ANOVA was conducted for each product using the delivery time data grouped per month. The data was first ordered in chronical order to ensure a realistic time sequence.

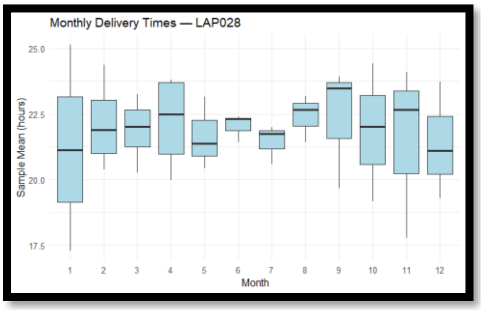
The ANOVA tested the null hypothesis (Ho) that determined that there are no significant differences in the average delivery times between months 1-12 for a given product. The alternative hypothesis (H1) tested that there is at least one month's average that differs from the rest. To visualize these differences, boxplots were created for each product type from 1-12 months thus displaying the spread and central tendency of the overall delivery times.

The p-value from each ANOVA was then used to determine the statistical significance at the commonly used 5% level. If the products exhibit high p-values, typically much larger than 0.05, it indicates that there were no statistically significant differences in the different delivery times between the various months. This then indicates that the delivery process remained stable and consistent throughout the entire year.

On the other hand, if any p-values are below 0.05 it indicates that there will be significant difference between the various months. The boxplots are used to visually support these findings by showing similar information graphically thus confirming the performance of the process as seen in the ANOVA. The following ANOVAs and the correlating boxplots for products LAP028, SOF005 and MOU053 is shown below:

LAP028

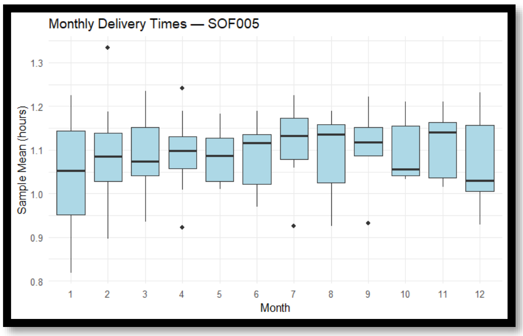
The ANOVA table of the product indicated a P-value equal to 0.9873428 which means that we fail to reject the null hypothesis. The boxplot of LAP028 shows that the median delivery time for each month is around 21-23 hours with similar interquartile ranges. There is no evidence of large seasonal variations although the variation months 6,7 and 8 are smaller than the rest of the months.



Product: LAP028					
Source	Sum of Squares (SS)	Degrees of freedom (DoF)	Mean Square (MS)	F	P-Value
Treatment	7.61	11	0.69	0.18	0.9977812
Error	137.46	36	3.82	---	---
Total	145.07	47		---	---

SOF005

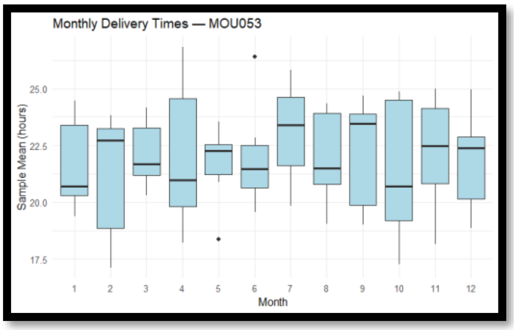
The ANOVA table of the product indicated a P-value equal to 0.9919493 which means that we fail to reject the null hypothesis. The boxplot of SOF005 shows that the median delivery time for each month is between 1.02 and 1.18 hours, which is very faster than the delivery of LAP028 meaning that this product is easier and faster to transport. There is no evidence of any large variations throughout the various months.



Product: SOF005					
Source	Sum of Squares (SS)	Degrees of freedom (DoF)	Mean Square (MS)	F	P-Value
Treatment	0.03	11	0	0.25	0.9919493
Error	0.91	84	0.01	---	---
Total	0.94	95	---	---	---

MOU053

The ANOVA table of the product indicated a P-value equal to 0.0.9873428 which means that we fail to reject the null hypothesis. The boxplot of MOU053 shows that the median delivery time for each month is between 21-23.5 hours. This delivery brackets correlates closer to product LAP028 than product SOF005 meaning that it might have similar delivery challenges than the products of LAP028.



Product: MOU053					
Source	Sum of Squares (SS)	Degrees of freedom (DoF)	Mean Square (MS)	F	P-Value
Treatment	17.63	11	1.6	0.28	0.9873428
Error	408.62	72	5.68	---	---
Total	426.26	83	---	---	---

A one-way ANOVA was performed to compare the mean delivery times across 12 months for products. The results indicated that there was no statistically significant difference in mean delivery times between the various months, as all of the product's P-value is far above the 5% significance level. The boxplots for the identified products all show that the median delivery time remains fairly consistent across all of the months, with little to no outliers and no months showing a large deviation. These graphs are consistent with the results of the given ANOVAs, which supports the statistical conclusion that there are no significant differences between months.

7. Reliability of Service

7.1 Reliable service days per year

In order to determine the number of days with reliable service we used the given staff data to estimate the probability that there will be enough workers present each day. This probability was then multiplied by 365 days to estimate the expected number of days per year with reliable service.

Given:

- Reliable service = days with >= 15 workers
- Total days = 397 days
- 365 days in a year
- R20 000 less per day if less than 15 workers
- R25 000 per extra worker

$Total\ reliable\ service\ days = 96 + 270 = 366\ days$

$Reliable\ proportion = \frac{366}{397} = 0.9219$

$Reliable\ Service\ days = 0.9219 \times 365 = 336.5\ days = 336\ days/year$

7.2 Optimisation of Company

To optimise the amount of staff we modelled the daily staff presence as a binomial process and compared the total expected loss from all the unreliable days with the cost of adding additional staff. The number of staff that minimised the total annual cost was then chosen as the optimal amount.

$$Average\ number\ workers = \frac{1(12) + 5(13) + 25(14) + 96(15) + 270(16)}{397} = 15.58\ workers$$

$$Rate\ of\ attendance = \frac{15.58}{16} = 0.975$$

$$Loss = 20000 \times 365 \times P(X < 15)$$

$$Total\ Cost = Loss + Staffing\ cost$$

For a total of 16 of workers:

Binomial Formula: $P(X < 15) = 1 - P(X \geq 15) = 1 - (P(15) + P(16))$

$$where\ P(15) = 16(0.975)^{15}(0.025) = 0.301$$

$$and\quad P(16) = (0.975)16 = 0.66$$

$$thus\ giving\ P(X < 15) = 1 - (0.301 + 0.66) = 0.039 = 3.9\%\ chance\ of\ unreliable\ service$$

$$Loss = 20000 \times 365 \times 0.039 = R284700$$

$$Total\ Cost = 284700 + 16(12)(25000) = R5084700$$

For a total of 15 of workers:

Binomial Formula: $P(X < 15) = 1 - P(15) = 1 - (0975)^{15} = 1 - 0.688 = 0.312 = 31.2\%$

$$Loss = 20000 \times 365 \times 0.312 = R2277600$$

$$Total\ Cost = 2277600 + 15(12)(25000) = R6777600$$

For a total of 17 of workers:

Binomial Formula: $P(X < 15) = P(X = 14) + lower\ terms = 0.004 = 0.4\%$

$$Loss = 20000 \times 365 \times 0.004 = R29200$$

$$Total\ Cost = 29200 + 17(12)(25000) = R5129200$$

After careful evaluation of all the calculations it can be seen that a total of 16 workers provides the optimal solution with the lowest overall cost of R5084700. A close second option, with a total cost of R5129200, is to add an additional worker to have a total of 17 workers. The worst decision will be to lay of a worker to only have 15 workers, because the total cost then rises to R6777600 and then offers a higher likelihood for unreliable service.

Conclusion

This project demonstrated the practical application of statistical analysis, process control and optimisation techniques to real-world business scenarios. By making use of descriptive analysis, patterns and trends in sales the various delivery times, and customer behaviours could be identified and thus highlighting some key insights such as the dominance of certain products over other and the purchasing behaviour of different customer segments. Statistical Process Control (SPC) was applied to monitor the reliability of deliveries as well as identifying areas of high variability and potential special causes, while the process capability indices quantified how well the processes met the given specifications. The risk analysis done through Type I and Type II errors highlighted the sensitivity of the control charts and the importance of the accuracy of sample sizes.

By completing the necessary data corrections, it ensured that the analyses reflected the true business performance thus revealing the most popular products and improving decision making accuracy. The optimisation exercises, like determining the optimal number of baristas needed and staff at the car rental agency to ensure reliability. The combined probability models were used in order to maximise the profitability of the company while taking the total cost into consideration to ensure that a high service reliability is maintained. ANOVAs were created and then analysed to confirmed that the monthly variations in delivery times were not statistically significant, thus indicating a stable operational performance.

In conclusion, this project made use of a variety of techniques from data analysis, statistics and optimisation to help make better decisions for the company, improve how the different processes run, and boost overall business performance. The approaches we used give a clear framework that can be applied to other business or operational problems, thus making sure that the decisions are based on given data and are cost effective.

References

- Montgomery, D.C., 2020. *Introduction to Statistical Quality Control*. 8th ed. Hoboken, NJ: John Wiley & Sons.
- Ross, S.M., 2017. *Introduction to Probability and Statistics for Engineers and Scientists*. 6th ed. Academic Press.
- Field, A., Miles, J. and Field, Z., 2012. *Discovering Statistics Using R*. London: SAGE Publications.
- Hyndman, R.J. and Athanasopoulos, G., 2021. *Forecasting: Principles and Practice*. 3rd ed. Melbourne: OTexts. Available at: https://otexts.com/fpp3/ [Accessed 20 October 2025].
- R Core Team, 2024. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.
- Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.