ECSA REPORT

Moss, A, Mr [26923823]

Contents

Table of Figures

List of Tables

Introduction

The following report is based on analysing datasets in order to evaluate process performance, reliability and profit optimisation. The report starts with a basic data analysis to gain insight into sales data and then follows with statistical process control (SPC) to monitor delivery times. Process accuracy is then assessed by calculating type one and two errors, after which data correction needs to be put in place ensuring consistency between datasets. Profit models then need to be developed for 2 coffee shops using service-time data. ANOVA testing and binomial modelling form the last part of the report with the goal of assessing operational efficiency.

1. Basic Data Analysis

1.2 Data Loading and Inspection

There are four data sets required to complete the data analyse, which are customer data, products data, sales data and product head office data. The customer data has 1000 rows and 5 columns (features), the column headings include CustomerID, Gender, Age, Income and City. The product data and product data head office also have 1000 rows and 5 columns, the Column headings include ProductID, Category, Description, Selling Price and Markup. The sales data consists of 1000 rows and 12 columns. The datasets contains a mix of different variable types including numerical variables (quantity, age, income) and categorical variables (city, gender).

	CustomerID	ProductID	Quantity	orderTime	orderDay	orderMonth	orderYear	pickingHours	deliveryHours	Gender	Age	Income	City	Category	Description	SellingPrice	Markup
1	CUST1791	CLO011	16	13	11	11	2022	17.7216667	24.5440	Male	39	100000	Los Angeles	Keyboard	burlywood silk	1070.54	16.41
2	CUST3172	LAP026	17	17	14	7	2023	38.3908333	31.5460	Female	58	90000	Chicago	Cloud Subscription	aliceblue silk	18711.72	13.51
3	CUST1022	KEY046	11	16	23	5	2022	14.7216667	21.5440	Female	20	95000	Seattle	Monitor	blueviolet silk	708.18	17.72
4	CUST3721	LAP024	31	12	18	7	2023	41.3908333	24.5460	Female	66	60000	Miami	Mouse	blueviolet marble	18366.92	29.35
5	CUST4605	CLO012	20	14	7	2	2022	15.7216667	24.0440	Female	70	25000	Chicago	Mouse	azure silk	963.14	10.13
6	CUST2766	MON035	32	21	24	12	2022	21.0550000	24.0440	Female	46	120000	Miami	Keyboard	chocolate silk	6396.18	26.20
7	CUST4454	MOU052	29	5	23	1	2022	12.3883333	25.5440	Male	84	115000	Houston	Monitor	azure matt	425.14	24.84
8	CUST582	MON032	1	19	9	6	2023	17.0575000	22.0460	Male	33	45000	New York	Cloud Subscription	cornflowerblue marble	6634.13	27.80
9	CUST3343	MON040	10	19	13	12	2023	24.0575000	24.0460	Female	60	110000	Houston	Monitor	blueviolet matt	5346.14	29.74
10	CUST4331	KEY049	1	18	30	4	2022	15.3883333	20.0440	Female	76	120000	Houston	Software	azure silk	752.75	29.11
11	CUST1628	CLO015	5	10	9	8	2023	13.7241667	14.0460	Female	60	140000	Miami	Laptop	azure silk	728.26	27.70
12	CUST1501	CLO019	6	9	23	10	2022	20.3883333	13.0440	Male	53	70000	New York	Software	aliceblue silk	1092.07	23.14
13	CUST3625	MON033	1	9	24	3	2022	23.0550000	17.5440	Female	68	75000	Los Angeles	Laptop	burlywood sandpaper	5572.82	29.72
14	CUST574	MOU051	3	3	26	6	2022	9.0550000	22.0440	Female	17	85000	Los Angeles	Laptop	blueviolet matt	375.59	22.22
15	CUST4713	KEY043	6	9	30	9	2023	15.0575000	30.5460	Female	72	120000	Houston	Software	cornflowerblue marble	516.41	22.83
16	CUST4488	CLO019	19	10	8	2	2022	15.7216667	21.0440	Female	80	80000	Miami	Software	aliceblue silk	1092.07	23.14
17	CUST2018	KEY049	2	15	24	4	2022	11.3883333	27.0440	Female	39	85000	Houston	Software	azure silk	752.75	29.11
18	CUST3847	CLO015	1	15	28	7	2023	11.3908333	30.5460	Female	88	70000	Houston	Laptop	azure silk	728.26	27.70
19	CUST4073	SOF002	1	7	29	2	2022	0.9814444	1.3522	Female	84	120000	Seattle	Cloud Subscription	cyan silk	505.26	10.43
20	CUST2948	MOU053	8	11	16	1	2022	14.3883333	16.5440	Female	64	75000	San Francisco	Keyboard	cornflowerblue silk	424.79	21.36

Table 1: Data loading

1.2 Summary Statistics

Customer Data Summary:

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Age	16.00	33.00	51.00	51.55	68.00	105.00
Income	5000	55000	85000	80797	105000	140000

Table 2: Customer Data Summary

The table above shows the summary statistics for the customer dataset. The age variable seen above has a minimum age of 16 years and a maximin of 105 years with the median being 51 years. It also has a mean age of 51.55 which is higher than the median representing a slightly older customer base. The income variable has a minimum of 5000 and a maximin of 140000 with a median income of 85000.

The mean income is 80797. The spread of this variable is large with the third quartile being 105000, this suggests that a large portion of customers have relatively high incomes.

Products Data Summary:

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
SellingPrice	350.4	512.2	794.2	4496.3	6416.7	19725.2
Markup	10.13	16.14	20.34	20.46	25.71	29.84

Table 3: Products Data Summary

The table above shows the summary statistics on the products dataset. The selling price variable has a range between 350.4 to 19725.2, with a median price of 794.2. Its mean is much higher at 4496.3, suggesting that the data is skewed towards higher priced products. The Markup variable has a range between 10.13 to 29.84, with a median of 20.34. The mean has a value of 20.46 indicating that the markup percentage across the products are relatively consistent.

Sales Data Summary:

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Quantity	1.0	3.0	6.0	12.93	23.0	50.0
orderTime	1.0	9.0	13.0	12.93	19.0	23.0
orderDay	1.0	8.0	15.5	15.5	23.0	30.0
orderMonth	1.0	4.0	6.0	6.44	9.0	12.0
orderYear	2022	2022	2022	2022	2023	2023
pickingHours	0.4259	3.3908	14.0550	14.6955	18.7217	45.0575
deliveryHours	0.2772	11.5460	19.5460	17.4762	25.0440	38.0460

Table 4: Sales Data Summary

The table above shows the summary statistics for multiple variables regarding the sales data. The quantity sold has a range between 1 and 50 with a median of 6 and a mean of 12.93. This indicates that smaller quantities are much more common. Order time ranges from 1 to 23 with a median of 15.5 and mean of 15.5. This suggests that the sales data is balanced across days. The order month variable ranges from 1 to 12 with a median of 6.0 and mean of 6.44, this indicates that there is possible seasonality.

1.3 Handling Missing Values

After inspecting the given data sets there were no missing values found, all variables had complete data. Therefore, no further action is required as the datasets were ready for analysis to occur.

1.4 Data Visualization

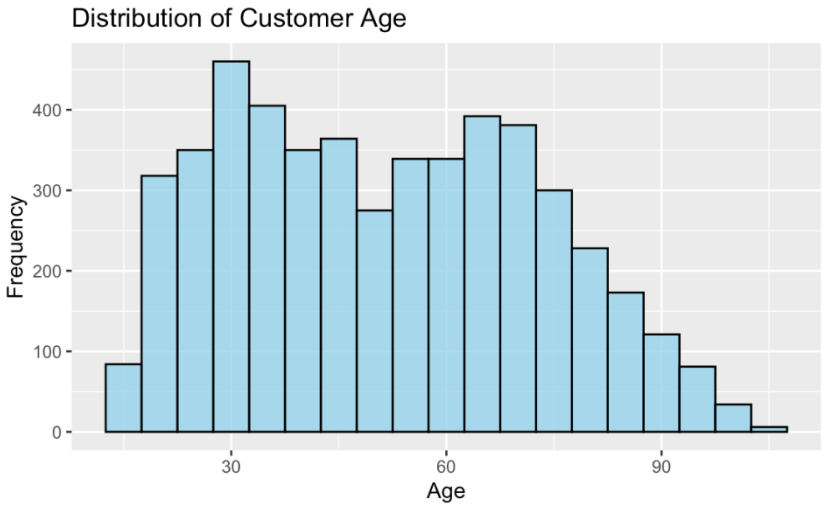

Figure 1: Distribution of Customer Age

The above histogram shows the distribution of the customers ages in the dataset. As seen above majority of the customers are aged between 25 and 70, with peaks noticeable in around the 30 and 60 age groups. The distribution is quite evenly spread but it can be seen that there are hardly any customers above 90 years old. The histogram indicates that the customer base is diverse in age, while having a very strong concentration of middle aged people. The information gathered from this visualization can be valuable in helping tailoring marketing strategies to certain age groups which make up a large portion of the data set.

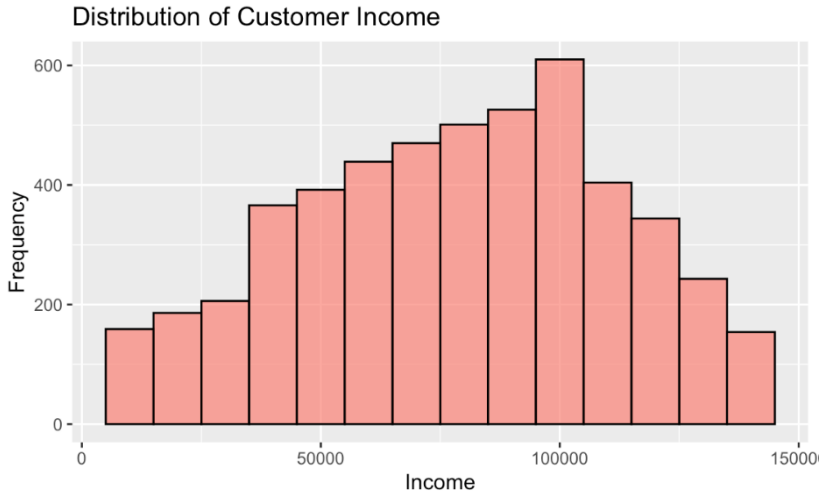

Figure 2: Distribution of Customer Income

The visualization above is a histogram which shows the distribution of the customers income in the data set. The majority of customers have an income between 60000 and 100000, with a clear peak around the 100000 mark. The distribution indicates that most customers fall into the middle to upper class income bracket. Fewer customers fall into the extremely low or high ends of the income distribution, as a very small portion have an income above 130000. The visualization is useful as the company is able to see what customers can afford as well as predict the demand for certain product categories.

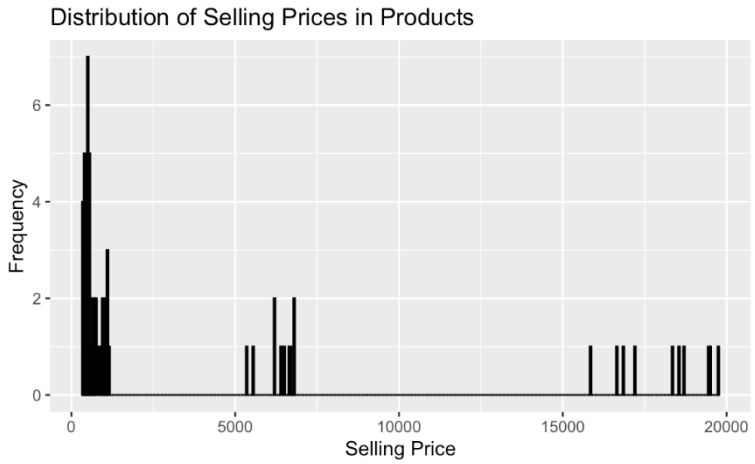

Figure 3: Distribution of Selling Prices in Products

The histogram seen above shows the selling prices of the products that the company offers. The distribution is skewed to the right as majority of the products are priced at the lower end of the pricing scale (under 2000). In contrast there are a small amount of products which are priced at the higher end of the scale, with some selling prices at around 20000. This suggests that most of the company's products are affordable with few higher end (premium) products. The visualization is useful as it provides insights into the product range, showing that the company can cater to both lower and higher earning customers.

Figure 4: Sales Distribution by Product Category

The above bar graph represents the total sales coming from different product categories. The graph indicates that the sales are really evenly distributed across cloud subscription, keyboard, laptop, monitor, mouse and software. There is slight variations in the sales however no single category dominates as each contributes very similar amounts to the total sales of the company. This visualization shows that the company is not dependant on a single product type, indicating that customers purchase a broad range of products.

3. Statistical Process Control (SPC)

3.1 Objective

The main goal of the following section is to evaluate the stability and capability of the delivery times for all the different products in the sales2026and2027 data using SPC. The delivery times were grouped into samples of 24 for each product, with the first 30 samples per product used to set control limits for the X-bar and s charts. These steps are followed to help identify which products processes are capable of meeting the voice of the customer specification limits.

3.2 Methodology

The sales2026and2027 data was organized by order year, order month, order day and order time. Each of the products delivery data was sorted into groups of 24 using the qcc package. Then X-bar charts were made to analyze the changes in the mean delivery time and s charts were made to look into process variability. The first 30 samples per product were used to calculate the center line and control limits for both of the charts. The remaining samples (samples 31 and on) were used to simulate process monitoring. The following X-bar and s graphs represent two different product types. types.

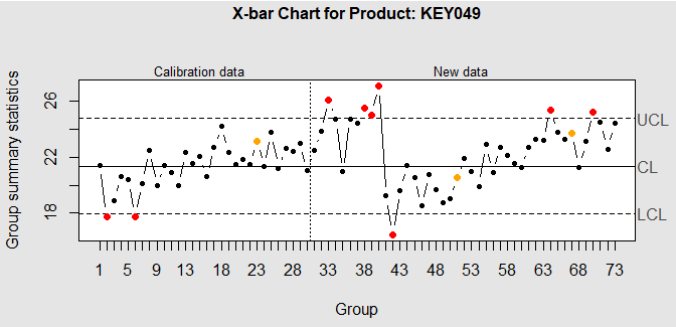

Figure 6: X-bar Chart for Product KEY049

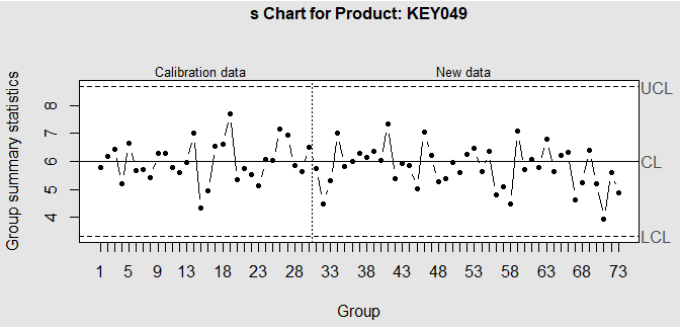

Figure 5: S Chart for Product KEY049

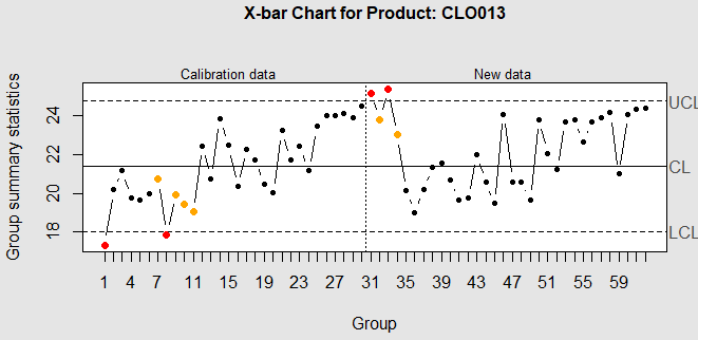

Figure 7: X-bar Chart for Product CLO013

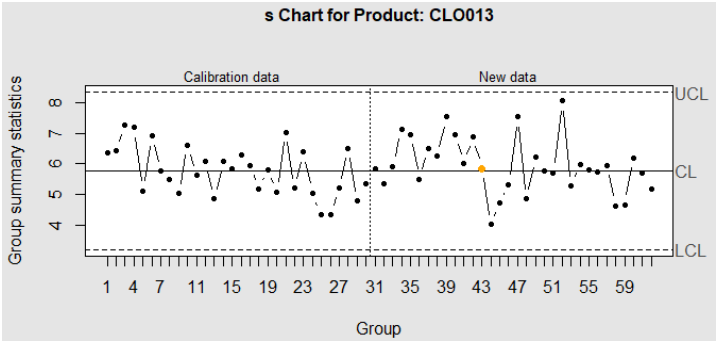

Figure 7: S Chart for Product CLO013

The above graphs for products KEY049 and CLO013 show how the sample means and process variation change over time. Product KEY049 is mostly stable and remains within the control limits, while on the other hand product CLO013 shows more fluctuations and has a couple points exceeding the limits, which suggests minor instability in the process.

3.3 Capability Assessment

The capability indices (Cp, Cpu, Cpl, Cpk) for the first 1000 deliveries per product were calculated using limits of 0 hours (LSL) and 23 hours (USL). Processes are regarded as capable if the Cpk is greater than 1.3, however all the products had Cpk values lower than 1.3, which shows that the variation in the delivery time is too high and the averages are too close to the upper limit. This results in none of the product processes being fully capable of meeting the VOC specification. The table included below is a small example which shows a few different product types and their corresponding indices values.

ProductID <chr>	mu <dbl>	sigma <dbl>	Cp <dbl>	Cpu <dbl>	Cpl <dbl>	Cpk <dbl>	Capable <lgl>
MOU059	21.188940	6.3198958	0.8438958	0.5702130	1.117579	0.5702130	FALSE
KEY049	21.985004	6.3136493	0.8447307	0.5287484	1.160713	0.5287484	FALSE
SOF009	1.085725	0.3050057	17.4860119	33.7854613	1.186563	1.1865625	FALSE
CLO019	21.538764	6.1339829	0.8694731	0.5684852	1.170461	0.5684852	FALSE
KEY045	21.837096	6.2973118	0.8469222	0.5379493	1.155895	0.5379493	FALSE
SOF010	1.069425	0.2964676	17.9895963	34.7767848	1.202408	1.2024078	FALSE
KEY046	21.776264	5.9533842	0.8958490	0.5724327	1.219265	0.5724327	FALSE
CLO012	21.686244	6.1681792	0.8646528	0.5573636	1.171942	0.5573636	FALSE
KEY047	21.498484	6.0944414	0.8751144	0.5743767	1.175852	0.5743767	FALSE
CLO020	20.895046	5.9577793	0.8951881	0.6213139	1.169062	0.6213139	FALSE

Table 5: Capability Indices

3.4 Rule Analysis

Rule A: None of the product types had samples outside of the +3σ limit, this shows that all the processes stayed within the normal variation

Rule B: All the product types showed pretty long consecutive runs within ±1σ range. This indicates very good control and low variability. The product types which had the longest stable runs were CLO011 and SOF009 with runs of 26 samples.

Rule C: There were only two product types, product MON035 and CLO013 that had one instance of four consecutive X-bar samples outside the second limit.

Overall, the processes are pretty well controlled and stable with only slight variations observed.

4. Risk, Data Correction and Optimising for Maximum Profit

4.1 Type 1 (Manufactures) Error

Type 1 errors occur when a stable process is misidentified as being out of control. Once the SPC rules were applied the following results were produced.

Rule	Description	Probability (α)
A	1 point above +3σ	0.00135
B	Run rule	0.00781
C	4 above +2σ	2.68×10^{-7}
Within ±1σ	Expected in-control range	0.6827

Table 6: Table of Type 1 and 2 Errors

The above table shows that Rule A produces a false alarm chance of about 0.135% per sample. The 7-run rule produces a probability 0.78% of incorrectly detecting an issue over 7 consecutive samples. Rule c is very strict as it has a probability of almost 0%. When the process is stable it is expected that 68% of all subgroup means fall within ±1σ of the centreline. This result shows that most points will remain within control limits and any errors that emerge from rule A and B are most probably random occurrences rather than real process shifts.

4.2 Type 2 (Consumer's) Error

Type 2 errors occur when a processes mean shifted but the associated control graph fails to show this. The bottle-filling process was designed to be centred on 25.050 L with UCL= 25.089 L and LCL=25.011 L. However, the actual process mean moved to 25.028 L with a subgroup standard deviation of 0.017 L.

The following results were calculated (β= 0.8412, 1−β= 0.1588). This represents that there is an 84% chance that the control chart will miss the shift and only a 16% of detecting it. In order to improve detection larger subgroups sizes could be used to try lower the value of β.

4.3 Data Correction and Re-analysis

The products_Headoffice.csv data file contained errors that were identified, to fix these issues the ProductID for items 11-60 were updated to match with its correct type (SOF, KEY, etc). The Category column in the products_data.csv was also aligned with each ProductID prefix. The corrected file was then used to re-run the basic data analysis done in part 1 to observe any differences.

After the data analysis was completed again using the corrected data file products_headoffice2025.csv, the customer age and selling price distributions remained the same. However, the sale by category distribution changed noticeably.

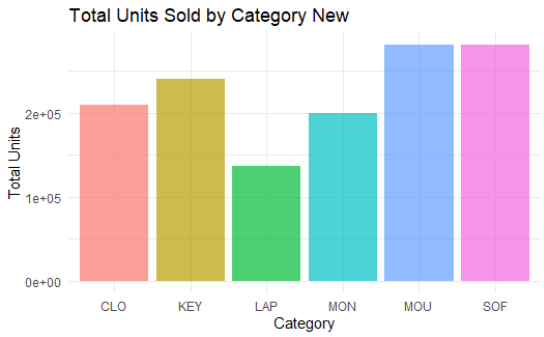

Figure 8: Updated Total Units Sold by Category

Previously the sales by category graph in part 1 showed very similar totals, but now after the corrections were made much clearer variation emerged. The mouse and software category had the highest sales, with keyboard and cloud having the second highest sales. The laptop and monitor category had the lowest sales. Overall, the corrected dataset shows a more accurate representation of the product performance. While being more reliable for further analysis.

5. Optimising Profit for Coffee Shops

The analysis was preformed using two service time datasets, timeToServe.cvs (Shop 1) and timeToServe2.csv (Shop 2). The goal was to find the number of baristas that maximises the profit while still having reliable service.

- Profit = 30(No. of customers served per day) – 1000(No. baristas)

Shop 1:

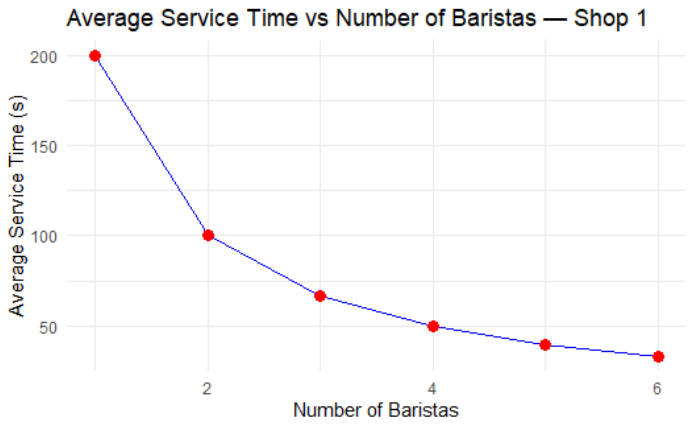

Figure 9: Average Service Time vs Number of Baristas Shop 1

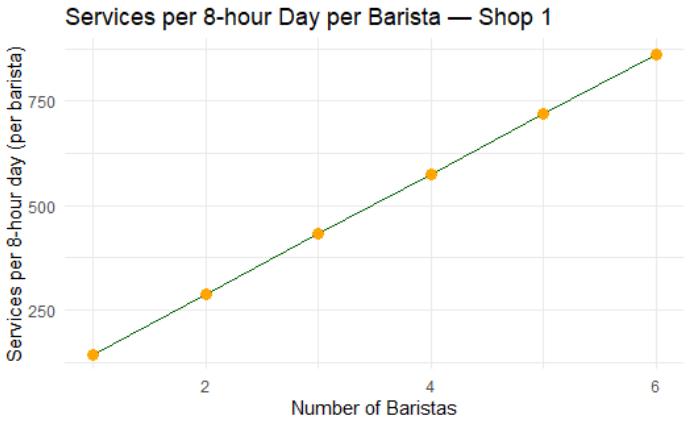

Figure 10: Services per 8-hour Day per Barista Shop 1

The above average service time vs number of baristas graph shows a big drop in the in the service time. The service time drops from about 200 seconds with one barista to around 40 seconds with six, this shows good efficiency improvements with increased baristas. The services per 8-hour day per barista graph has a positive linear gradient, with the graph reaching almost 900 services per barista. Shop 1 also had a service reliability of 99.79%, indicating that almost every customer was served within the target time. This shows that the profit increases significantly with 6 baristas compared to when only 2 are used.

Shop 2:

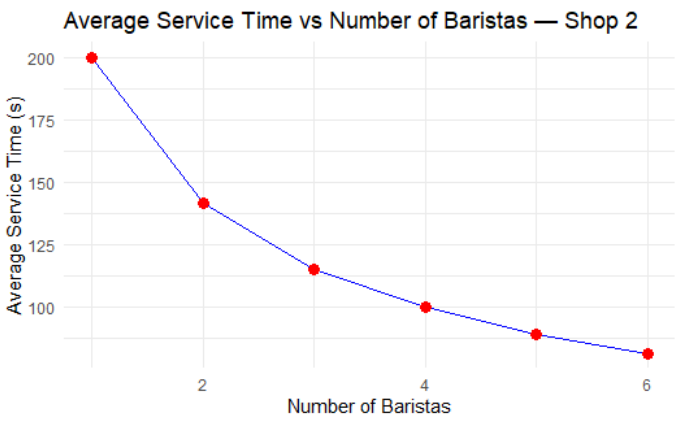

Figure 11: Average Service Time vs Number of Baristas Shop2

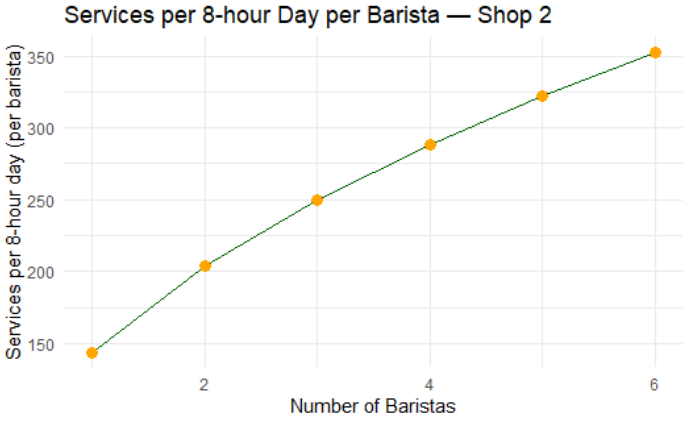

Figure 12: Services per 8-hour Day per Barista Shop 2

The above graphs represent the average service time vs number of baristas and services per 8 hour day per barista for shop 2. The average service time also declined dramatically form 200 seconds with one barista to about 90 seconds with six baristas. The services per barista increased to 350 services per barista (6 baristas), comparing this to only around 200 services per barista when there are 2 baristas. The reliability reached 92.44% which shows solid performance but slightly lower than shop 1. The profit reaches a max when the staffing limit is reached (6 baristas), just like shop 1 does.

Overall, across both of the shops the graphs shows that by adding baristas the speed, reliability and overall profitability increases. This shows that six baristas offer the best balance between cost and efficiency, while delivering the highest profit.

6. DOE and ANOVA

6.1 Setting up the ANOVA test

The sales2022and2023.cvs dataset was used to perform a two-way ANOVA in order to determine if the laptop delivery times differed across months and also between the years 2022 and 2023. The variable that was of focus was delivery time and the analysis aimed to identify if there were and patterns that affected the delivery performance.

6.2 Results

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
orderYear	1	19	18.9	0.548	0.4591
orderMonth	11	24376	2216.0	64.220	<2e-16
orderYear:orderMonth	11	680	61.9	1.793	0.0495
Residuals	10183	351371	34.5		

Table 7: ANOVA results

Table 7 shows the ANOVA results, which indicate that the orderMonth had significant effect on the delivery time (F= 64.22, p< 2e-16). The orderYear showed to have very little effect on the delivery time as it had a F value of 0.548 and a p value of 0.4591. Furthermore, the interaction between the order month and year was marginally significant as it had a F value of 1.79 and a p value of 0.0495. This shows that there are very small differences between the year-to-year monthly delivery trends.

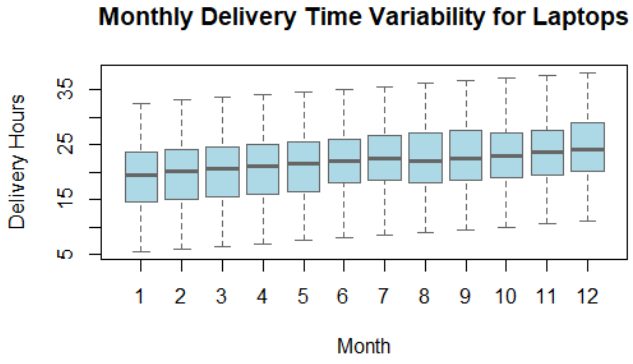

Figure 13: Monthly Delivery Time Variability for Laptops

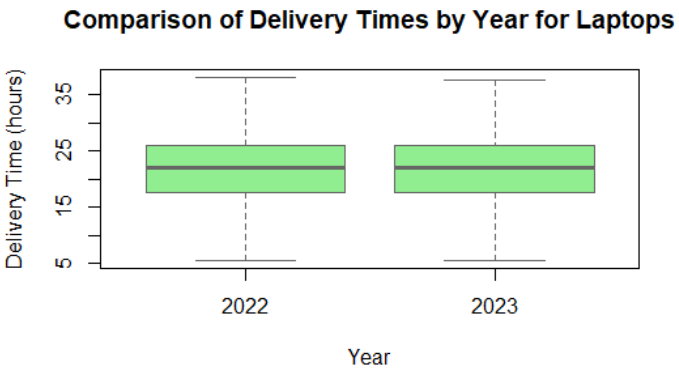

Figure 14: Comparison of Delivery Times by Year for Laptops

The monthly variability boxplot shown in figure 13 indicates a steady increase in the median delivery times form month 1 to 12, with bigger variation being able o be seen towards the end of the year. The yearly comparison plot shown in figure 14 shows that the overall averages for 2022 and 2023 were very similar, indicating that the variation is mostly caused by seasonal factors rather than annual factors.

After analysis it is clear that the month-to-month variation is the main cause of delivery time differences and indicated that the performance between the years remained consistent.

7. Reliability of Service

7.1 Reliable-service days per year

Given the number on people on duty over the 397-day period it is said that reliable service occurs when 15 or more staff are on duty. The data from the from the given graph shows that there were 96 days where 15 staff were present and 270 days with 16 staff present, which gives a total of 366 days of reliable service. This is about 92.2% out of the total 397-day analysis period and when scaled to a 365-day year its about 336 reliable days of service.

7.2 Optimising company profit

In order to estimate how staffing changes, affect the profit a binomial model was used. The days which showed issues in reliability of service were ones with less than 15 staff and caused around R20 000 in lost sales each day. Each added staff member costs an additional R25 000 per month.

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 8: Table of Net profit Gian with Additional Staff

Table 8 shows the effect of adding one to three additional staff members. If one extra staff member is added six problem days still occur but there is a big reduction in loss of about R459 700 and a net gain of R159 698. However, if more than one additional staff member is added the costs become greater than the savings and a net loss experienced.

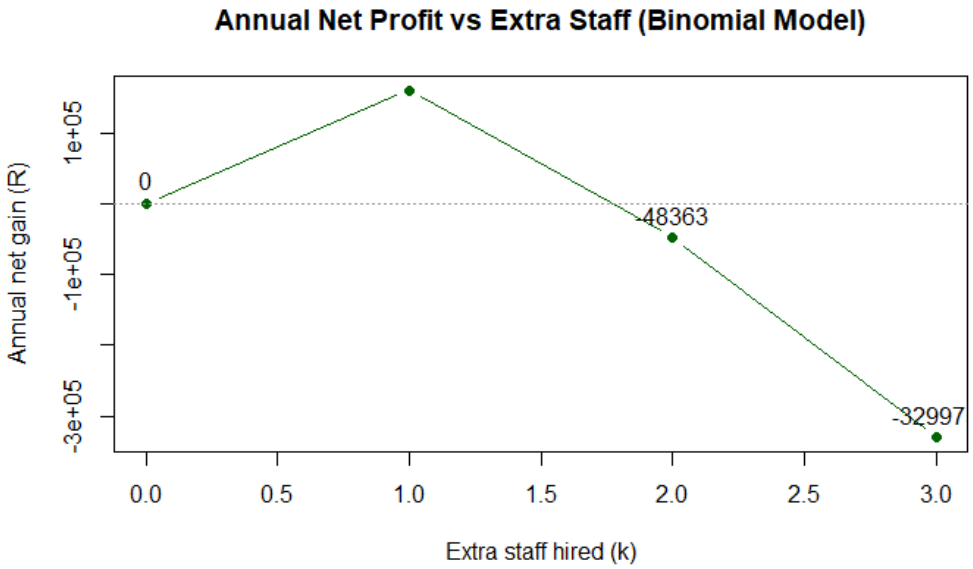

Figure 15: Annual Net Profit vs Extra Staff

As seen above in figure 15 by hiring one additional staff member it gives the highest annual net profit and the service reliability improves to about 360 days of reliable service per year. This is the most optimal balance of adding employees as it reduces the lost sales associated with service problems.

Conclusion

In conclusion the analysis showed that how data driven method can be used to improve process control, reliability and profitability. The statistical process control charts indicated that mostly all of the delivery processes were stable and within the limits. The type 1 and 2 errors that were calculated showed the likelihood that process variation would go undetected. The coffee shops were analysed and profit optimisation was preformed to find the optimal number of baristas. The car rental agency was also analysed in regards to the number of staff, the staffing levels were adjusted until the efficiency and financial performance improved. Finally the ANOVA test was preformed and showed that delivery time varied seasonally and not yearly.

References

Faraway, J. (2002). *Practical Regression and Anova using R*. [online] Available at: https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf?utm_source=chatgpt.com [Accessed 20 Oct. 2025].

Oakland, J. and Oakland, R. (2018). *Statistical Process Control*. *Routledge eBooks*. Informa. doi:https://doi.org/10.4324/9781315160511.

OpenAI (2025). *ChatGPT*. [online] ChatGPT. Available at: https://chatgpt.com/.