1/10/2025

ECSA Report

Data Analytics Report

Belfi, DRK, Mr [26945231@sun.ac.za]

Table of Contents

Table of Figures

Lists of Tables

Part 1

1.2 Data Loading and Inspection

The required csv files customer data, monthly sales, product data, sales data and top products were given to us to load and use. Customer_data has 5 features containing 5000 instances where 4 are categorical and 1 is continuous. Monthly_sales has 3 features containing 24 instances where 2 are categorical and 1 is continuous. Product_data contains 5 features containing 60 instances where 2 categorical, 2 continuous and 1 textual. ProductHeadOffice_data contains 5 features containing 360 instances where 2 categorical, 2 continuous and 1 textual Sales_data has 9 features containing 100000 instances 7 categorical and 2 continuous. Below is a ABT that was created merging all the datasets together to perform the analysis.

	CustomerID	ProductID	Quantity	orderTime	orderDay	orderMonth	orderYear	pickingHours	deliveryHours	Date	Gender	Age	Income	City
1	CUST001	MOU056	4	13	5	5	2022	15.7216667	24.5440	2022-05-05	Male	16	65000	New York
2	CUST001	KEY045	1	17	26	8	2022	13.7216667	25.0440	2022-08-26	Male	16	65000	New York
3	CUST001	MOU055	2	19	13	4	2022	12.3883333	11.0440	2022-04-13	Male	16	65000	New York
4	CUST001	SOF004	43	14	16	2	2023	0.7149444	0.5523	2023-02-16	Male	16	65000	New York
5	CUST001	CLO017	2	11	1	1	2023	12.3908333	10.5460	2023-01-01	Male	16	65000	New York
6	CUST001	MOU059	2	5	2	8	2023	18.7241667	28.0460	2023-08-02	Male	16	65000	New York
7	CUST001	SOF005	13	22	16	11	2023	0.9816111	1.0773	2023-11-16	Male	16	65000	New York
8	CUST001	CLO011	28	17	1	9	2022	11.0550000	17.5440	2022-09-01	Male	16	65000	New York
9	CUST001	MOU055	4	16	18	5	2023	12.7241667	25.5460	2023-05-18	Male	16	65000	New York
10	CUST001	KEY045	20	12	30	8	2022	11.7216667	28.0440	2022-08-30	Male	16	65000	New York
11	CUST001	CLO020	1	12	29	5	2023	9.7241667	13.5460	2023-05-29	Male	16	65000	New York
12	CUST001	SOF001	36	9	23	6	2023	0.6705000	0.7523	2023-06-23	Male	16	65000	New York
13	CUST001	SOF002	32	12	30	3	2023	0.7371667	1.2773	2023-03-30	Male	16	65000	New York
14	CUST001	MOU053	2	18	30	11	2022	15.7216667	25.5440	2022-11-30	Male	16	65000	New York
15	CUST001	MOU051	18	10	10	4	2023	10.3908333	10.0460	2023-04-10	Male	16	65000	New York

Table 1: Dataset

1.2.1 Summary Statistics

Sales Data

Feature	Min	1st Qu.	Median	Mean	3rd Qu.	Max
CustomerID	1	3	6	13.5	23	50
ProductID	1	3	6	13.5	23	50
Quantity	1	6	6	13.5	23	50
orderTime	1.00	4.25	13.00	13.5	17.00	23.00
orderDay	1.00	8.00	15.00	15.5	23.00	50.00
orderMonth	1.00	4.00	6.00	6.00	9.00	12.00
orderYear	2022	2022	2022	2023	2023	2023
pickingHours	0.4259	9.3908	14.6955	17.4765	18.7217	45.0575
deliveryHours	0.2772	11.5460	15.0400	18.0440	25.0400	38.0460

Table 2: Sales Data Statistics

information shown in table above clearly suggests that most orders tend to be for a small number of items, with occasional larger orders. Also, the time spent picking and delivering vary – highlighting that some orders require more effort or are more complex. Furthermore, order timing is spread throughout the day with no major peaks.

Product Data

Feature	Min	1st Qu.	Median	Mean	3rd Qu.	Max
SellingPrice	350.4	512.2	794.2	4493.6	6416.7	19725.2
Markup	10.13	16.14	20.34	20.46	25.71	29.84

Table 3: Product Data Statistics

There is a very large range for selling price yet the majority of products are cheaper – there is an offering of both budget and premium products. The markup distribution shows most products have a consistent profit margin.

Customer Data

Feature	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 4:Customer Data Statistics

Age has very uniform distribution. Whereas income has a left-skewed distribution, with more people earning at the lower end of the scale.

1.2.3 Handling Missing Values

There is no missing in any of the datasets.

1.2.4 Data Visualisations

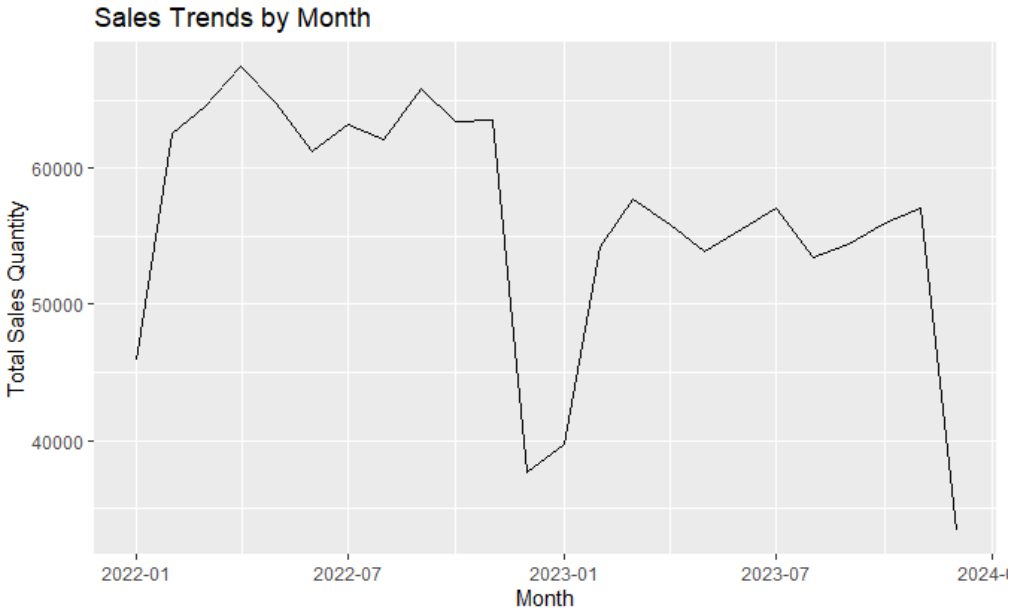

Figure 1: Sales Trend by Month

The sales trend graph suggests to have a seasonal pattern, with strong peak followed by a steady drop, possibly linked to market conditions, promotions or environmental factors.

Product ID	Total Quantity
MOU059	29675
SOF001	29336
SOF004	29219
SOF010	29168
MOU058	28924
MOU054	28875
MOU052	28804
SOF007	28517
MOU057	28423
SOF005	28412

Table 5: ProductID's Total Quantity Sold

Moreover, the products sold indicate that most products are sold in a similar quantity but the most sold is MOU059 and the least sold is 28412.

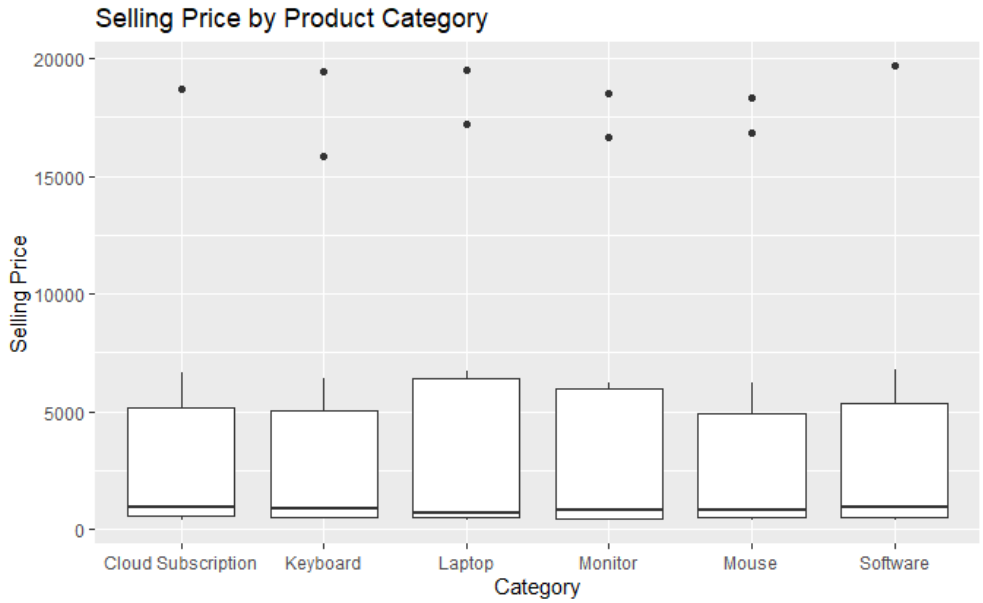

Figure 2: Selling Price by Category Box Plot

Furthermore, as shown above laptop category has the highest selling prices, with some products priced well above the others, likely due to the presence of high-end models. Most of the products fall into a consistent pricing range, but there is a segment of premium products driver in the higher prices.

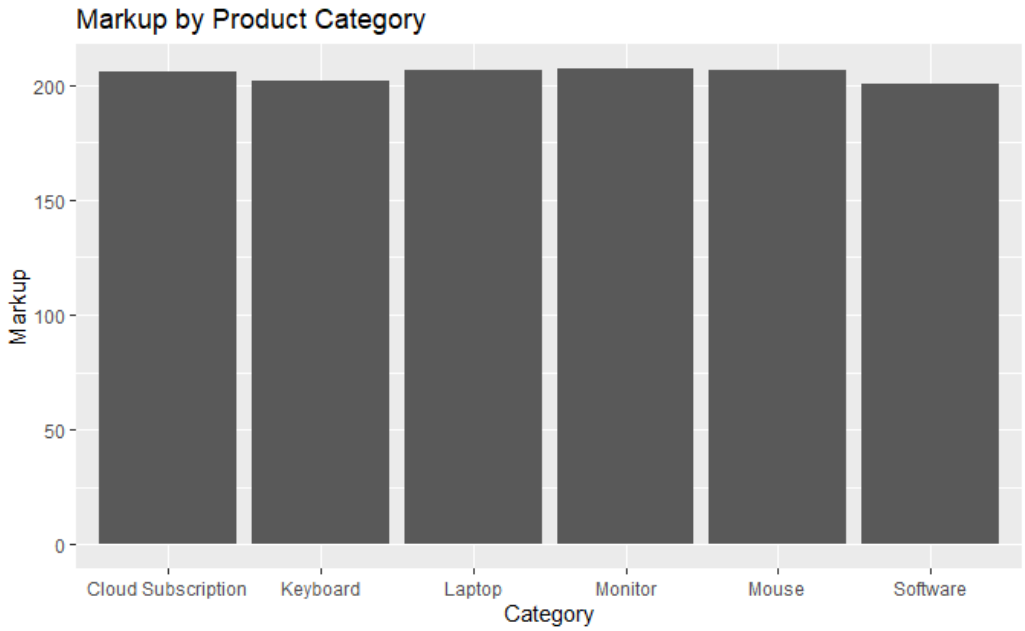

Figure 3: Markup by Product Category Bar Chart

In addition, the uniformity in the markup suggests that a company has a consistent profit margin strategy across all product categories. The pricing strategy is probably more focused on maintaining a set standard markup rates then adjusting for each category of product as shown in the bar plot above.

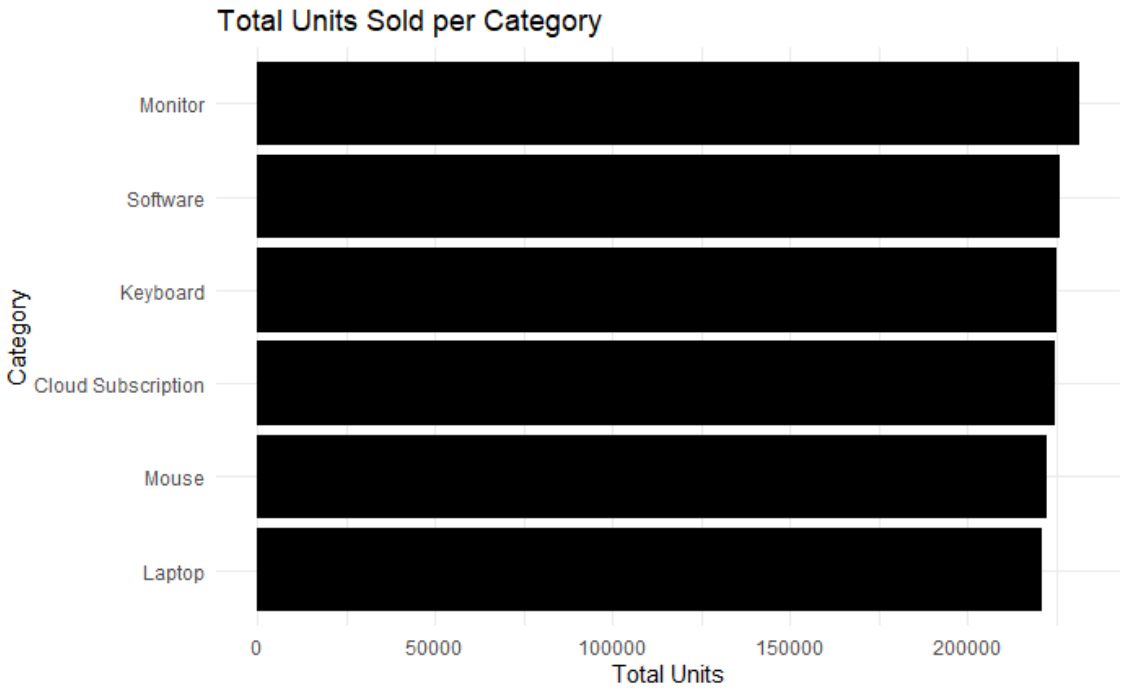

Figure 4: Total Units Sold per Category

Moreover, the total units sold per a category suggests that sales were distributed quite evenly across all categories with slight variations. Monitors recorded the highest number f units sold, followed closely by all other products. This consistency suggests balanced demand across all different product types – reflecting a well-diversified product portfolio and stable customer demand following in all categories.

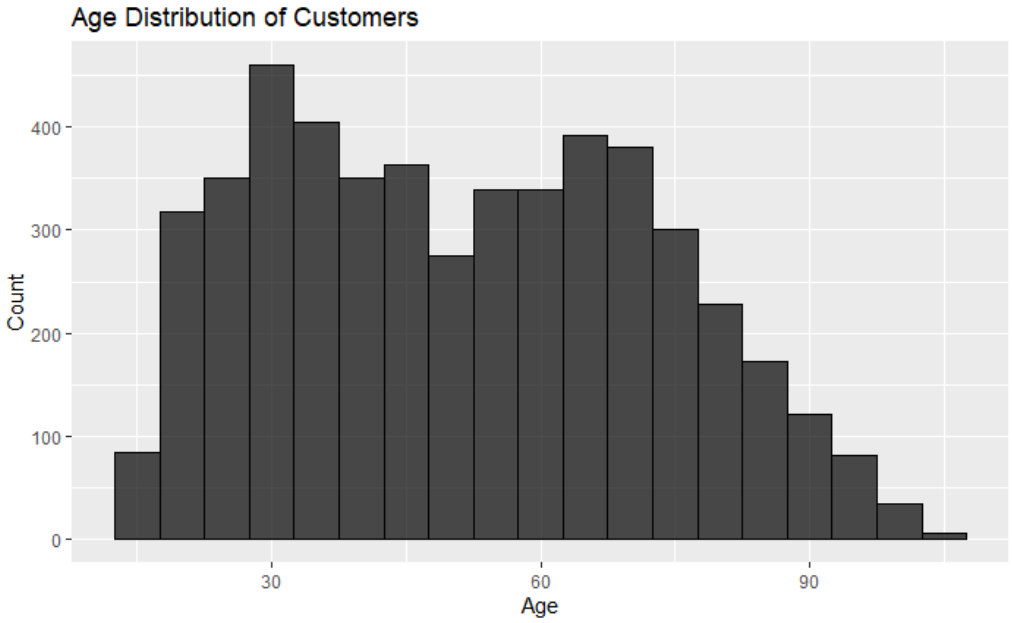

Figure 5: Age Distribution of Customers Histogram

Data is bimodal as there is two peaks, one at 30-40 years and 50-60 years as shown above. There is also an indication that the company has a very large customer base.

Part 3

3.1 Control-Chart Initialisation

The main idea for this section is to determine the stability for the different products delivery times. Delivery-time data for all product types were ordered chronologically with subgroups of 24 consecutive deliveries created for each product with the first 30 samples forming the baseline dataset. Moreover, the average and standard deviation of each subgroup were computed and used to construct the control charts.

3.2 Ongoing Process Monitoring

After establishing baseline limits an additional set sample (31 onwards) were plotted to simulate ongoing monitoring of delivery performance. Also, each new subgroup of 24 deliveries were first assessed on the s-chat that each variation remained in expected limits – which then allowed for the average values to be interpreted.

3.3 Process Capability Indices

For each Product type, the first 1000 deliveries were analysed to determine capability indices with USL = 32 Hours and LSL = 0 Hours. In Addition, majority of products had Cpk values between 1 and 1.33 – indicating marginal to satisfactory capabilities. There are no products with a Cpk greater or equal to 1.33 which if present indicate a stable and capable product delivery. Whereas other products like KEY049 (Cpk<0.53) process can be under control but not meeting the Voice-Of-Customer (VOC) requirement of less or equal to 32 hours.

3.4 Identification of Process-Control Issues

Rule A – no need for adjustment as total number of instances that fall out of the limits of rule A is 0, where it would have instnaces if any of the samples are outside of the upper limit +-3 sigma.

Rule B - The below stated instnaces have the most consectutive samples between -1 and 1 sigma control limits

Group <chr>	longest_run_len <int>
CLO011	26
MON039	23
KEY041	20
LAP026	18
CLO017	17
MOU055	16
SOF002	16
KEY047	15
LAP025	15
MOU056	15

Table 6: Rule B - Longest Run Time

Rule C – The below table shows 10 instances that have the highest level of violations with 4 consecutive upper second control limits**.**

Group <chr>	total_violations <int>	max_consecutive_outs <dbl>
MOU0...	4	7
SOF002	3	9
MOU0...	3	6
SOF004	3	5
SOF008	3	4
MOU0...	2	8
MOU0...	2	8
MOU0...	2	7
MOU0...	2	7
SOF005	2	7

Table 7: Rule C - Violations

Part 4

4.1 Type I (Manufacturer's) Error

A Type I error occurs when the process in in control, but there is a signal that there is something wrong in the SPC chart – known as a false alarm. When the process is perfectly centred on the centreline the probability that any single sample mean lies above the centreline is 0.5, half values are above and the other half are below. Below is the output form the run-length rules.

- **Rule A (L = 7): $\alpha_A = 0.5^7 = 0.0078125$**
- **Rule B (L = 8): $\alpha_B = 0.5^8 = 0.00390625$**
- **Rule C (L= 9): $\alpha_C = 0.5^9 = 0.001953125$**

The Above probabilities represent the chances of a false signal when the process is in control. The longer the required run equals the smaller chance of a false alarm. By that, Rule C is the most conservative whilst A is the most sensitive.

4.2 Type II (Consumer's) Error

A Type II error happens when the process has actually shifted, but the control chart fails to detected it. In this scenario the processes should be centred a 25.05L with limits at UCL = 25.089L and LCL = 25.011L. The probability of failing to detect the shift as **$\beta = 0.8412$** – stating that there is about a **84%** that the chart will not signal even though the process mean has moved. The detection power **$(1-\beta) = 0.1588$**, so there is only a **16%** chance that the chart will identify the shift on any of the single subgroups.

The high Type II error rate shows that the chart is insensitive to small mean shifts when the invariability increases.

4.3 Comparison

With the updated head-office data and repeating the Week 1 analysis, there is a clear few difference that emerged between the old and the newly corrected results. In the original dataset (Black Bar Charts), the total units sold and total 2023 values appeared almost identical across the range of different products, which further indicated duplicated prices and incorrect product coding. After the corrections have been made (Blue Bar Charts), the results bow resemble a more realistic variation between product categories.

In conclusion, Software and Cloud Subscriptions sold the highest number of units, while Monitors and laptops were leading the highest number of value due to their higher selling prices. The updated data provides an accurate reflection of the performance - allowing for more reliable conclusions to be made about each product category and how each category drives volume or revenue.

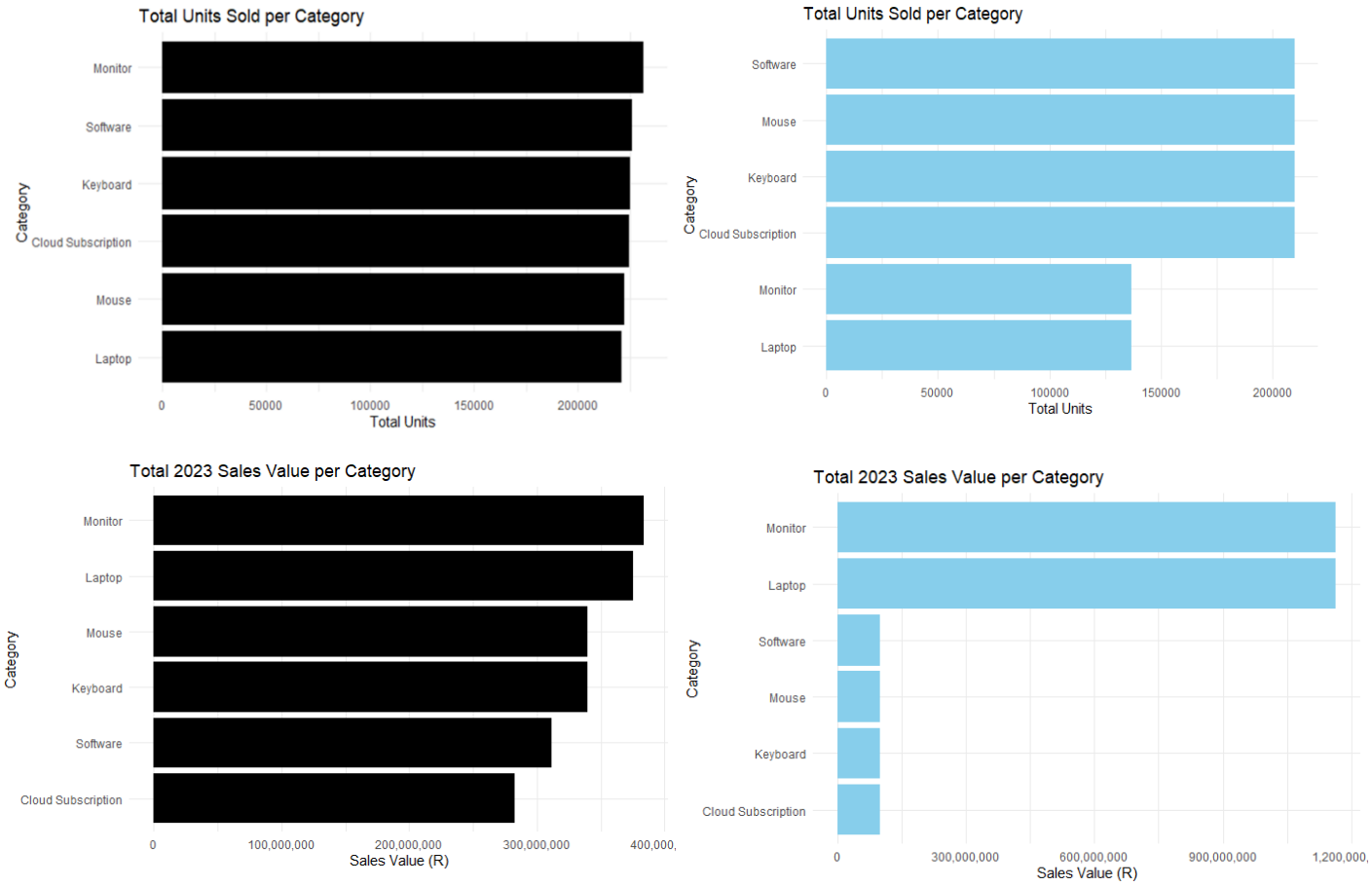

Figure 6: Comparison Between Before and After Data Adjustment

Part 5 – Profit Optimisation

For **Shop 1,** the graph of yearly profit increases linearly with each additional barista, showing no signs of a bad return within the tested range. This indicts that hiring more baristas consistently can increase the overall capacity and revenue faster then the increase of costs. Therefore, the best financial outcome occurs when maximum level of baristas which is 6. At this point, the shop completes around 900 services per an eight-hour day, which adds R27 000 per a day to the revenue of the shop. After subtracting the daily barista costs of R6000 (R1000 per a barista) this leaves the shop with a net profit of roughly R7.7 million per year. Also, Shop 1 provides **99%** of customers with reliable service.

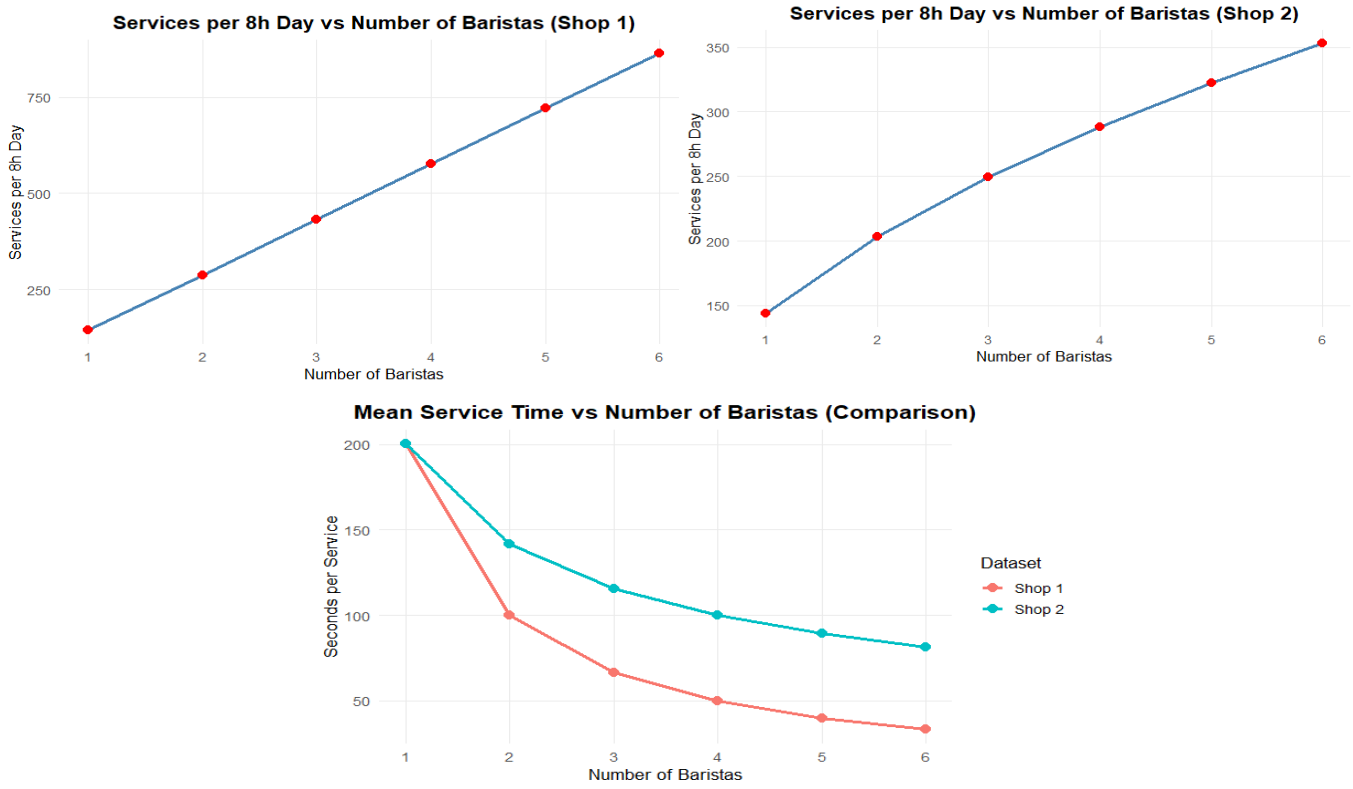

Figure 7: Comparisons Between Shop 1 and Shop 2

For **Shop 2**, displays a diminishing return. Profit rises sharply until five baristas, after which the curve flattens and starts to decline. This change indicates that additional staffing beyond this point only increases costs faster then the gain for service capacity. With five baristas, Shop 2 handles 330 services per a day, generating daily revenue of R9 900. After deducting daily staffing costs of R5000, the estimated annual profit peaks near R1.8 million. Also, Shop 2 provides **95%** of customers with reliable service.

Part 6 – ANOVA/MANOVA Analysis

The ANOVA and MANOVA analysis aimed to test whether there were major differences in laptop performances between the years 2022 and 2023. The dependent variables chosen were Quantity, Picking Hours and Delivery Hours. The null hypothesis assumed that there would be no meaningful changes in these variables between the years.

From the box plots below, Quantity remained constant, indicating a controlled and steady order quantity. Delivery Hours showed a micro increase in 2023, yet the overlap of the interquartile range indicates nothing statistically significant. in contrast, Picking Hours increased in 2023, with higher medians and a much larger spread of values, significantly towards the end of the year.

The ANOVA results confirmed no significant differences with Quantity and Delivery Hours in the two years ($p > 0.05$) whilst Picking Hours showed a significant difference ($p < 0.05$). In summary, laptop sales volumes and delivery performance were consistent, while warehouse picking operations become slower and showed high variability in 2023.

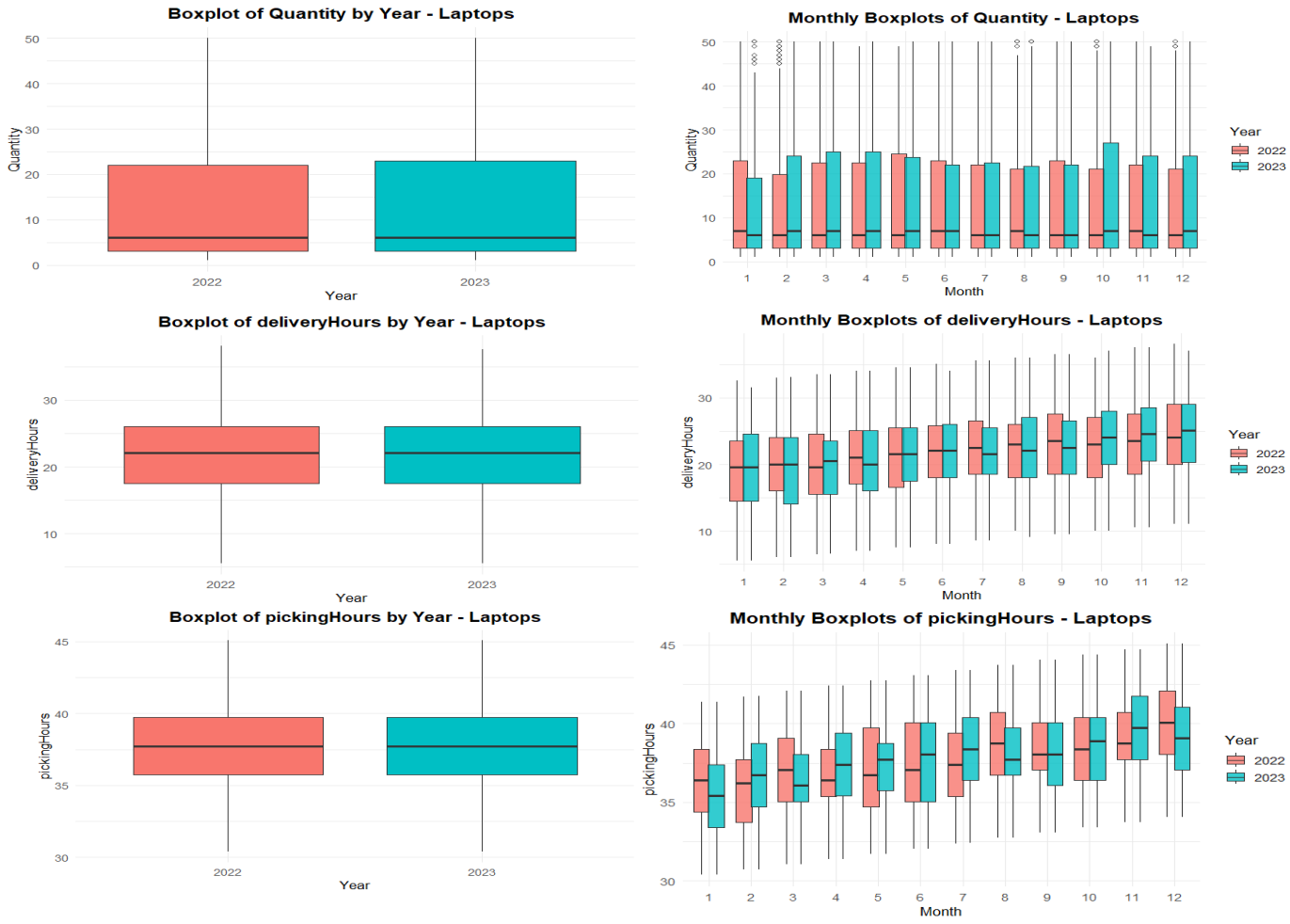

Figure 8: Box Plots Per Category (Yearly and Monthly)

Part 7 – Profit Optimisation

The data shows that, with 16 staff members the agency provides a reliable service on approximately 336 days per year which equates to 92.2% reliability and consistency. When the model was then modelled using a binomial distribution with an estimated available probability of p = 0.9740, the expected number of reliable service days remained consistent – thus suggesting the staffing level is mostly sufficiently but not optimised.

When the profit model was applied, with the daily sales loss of R20 000 whenever staffing fall below 15 employees and a monthly staffing cost of R25 000 per each additional employee. The optimisation suggested that one additional staff member yields the best financial outcome – as this reduces the probability for a problem day from 6.4% to 0.9%, additionally it cuts the expected monthly costs by R7 736 per month, which increases the net profit by R158 700 per a year.

In conclusion, the addition of one employee, changing it from the usual 16 to 17 employees achieves a reliability of around 99% where the service days have no problems.

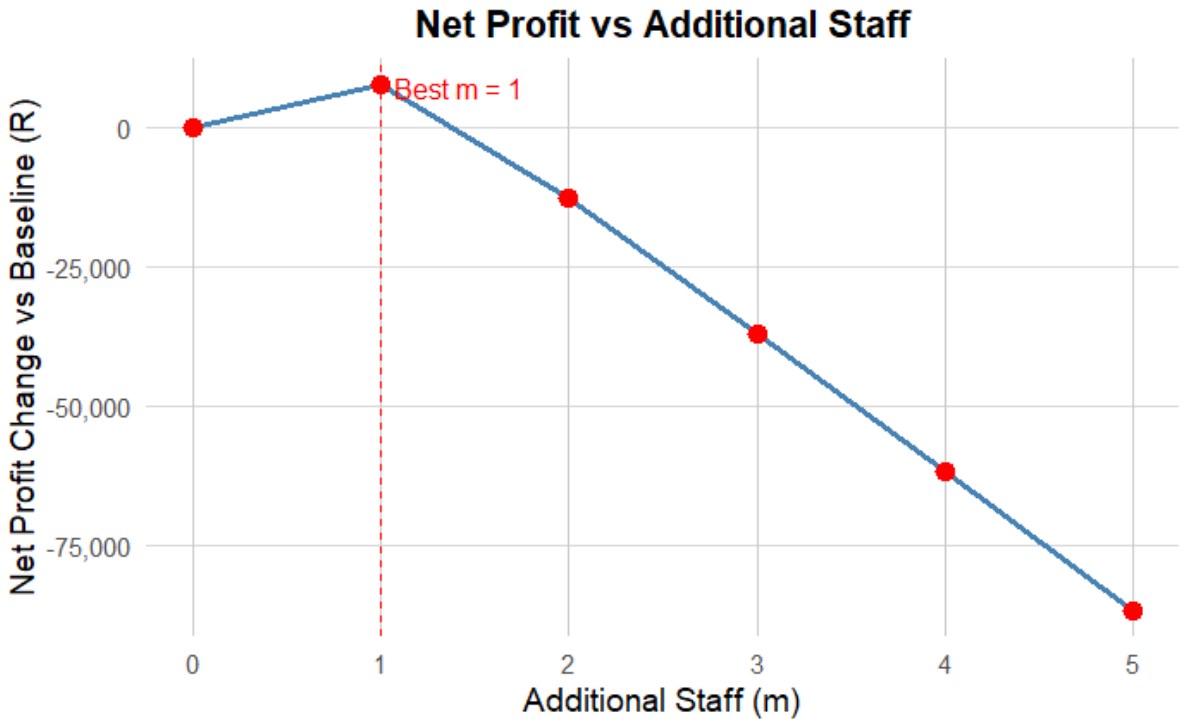

Figure 9: Net Profit with Each Additional Staff Member

Conclusion

This report applied data analytics and statistical; tools to solve real work industrial problems. After correcting the data inconsistencies, the now updated analysis produced accurate sales insights reflecting a true story of category performance. Statistical; process control confirmed stable delivery processes for most produces, whilst capability indices highlight areas in need pf improvement. In addition, the Type I and Type II error evaluations quality assurance balance. Profit optimisation for both coffee shops and the car rental agency demonstrated how the adjustment of staffing directly affects profitability and reliability. ANOVA result's identified increased variability within the warehouse picking times. Finally, the project combined the theoretical knowledge with real time world data, meeting the required ECSA GA4 outcomes through engineering reasoning and evidence-based decision making.

References

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/.

Nist.gov. (2019). *NIST/SEMATECH e-Handbook of Statistical Methods*. [online] Available at: https://www.itl.nist.gov/div898/handbook/.

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

Sthda.com. (2025). *Easy Guides - Wiki - STHDA*. [online] Available at: https://www.sthda.com/english/wiki/wiki.php [Accessed 20 Oct. 2025].