ECSA Project Data Analysis Report

Kate Wannenburgh | 27593231

Stellenbosch University – Quality Assurance 344

Table of Contents

Abstract

This report presents a comprehensive analysis of a company's customer base, product portfolio, and sales performance to derive strategic business insights. Utilizing K-means clustering, the customer base was segmented into six distinct groups, identifying high-income earners (Cluster 3) and the largest segment of younger customers (Cluster 1) as primary targets for marketing. Product analysis revealed three clusters, highlighting a strategic divergence between high-value, low-margin hardware and affordable, high-margin digital products like software and cloud subscriptions. Sales trend analysis identified a significant seasonal dip during summer months, crucial for inventory planning. Furthermore, customers over 60 were found to be the most valuable demographic, and Los Angeles was the top revenue-generating city. A Statistical Process Control (SPC) analysis of delivery times indicated that no product category currently meets the Voice of the Customer (VOC) capability target (Cpk ≥ 1.33), revealing a critical area for operational improvement. This analysis provides a data-driven foundation for optimizing marketing, sales, and logistics strategies.

ECSA Project Data Analysis Report

Basic Data Analysis: Descriptive Statistics

Firstly, customers and products will be analyzed individually, then sales will be researched and finally customers per product and vice versa will be discussed. It was found that no missing values were present in any of the data sources before starting with the analysis.

Customers

The company has a wide range of customers from age 16 to 105 with incomes ranging from 5000 to 140 000 with an IQR of 50 000. To understand the customers better K-means clustering algorithm has been run which results is six groups of customers as shown in figure 1. Cluster 3 stands out as an important cluster for marketing with the highest average income of just over 122 000, second largest group and a mean age of 50.

The older customers are split into clusters 4 and 6 which represent an average income of around 46 000 and 100 000 respectively. Cluster 5 consists of young customers with the lowest average income of 33 000 while cluster 1 also consists of young customers but with a higher average income of 84 000. In addition to this, cluster 1 is our biggest group of customers so marketing should target this group. Develop high-value financial products and premium loyalty programs targeted at Cluster 3 to leverage their high income, while creating accessible, entry-level offerings and referral campaigns for the large, young audience in Cluster 1 to secure future loyalty.

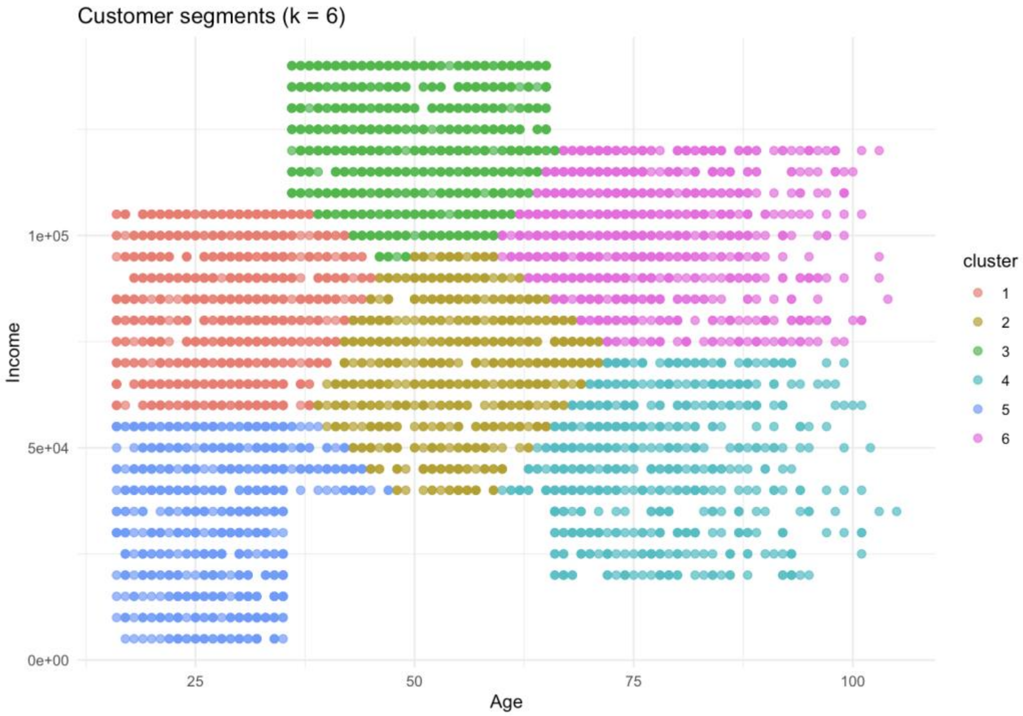


Figure 1 Customer clusters

The following bar chart shows us that customers are evenly spread across cities with San Fransico being slightly more common and Miami and Seattle being less popular.

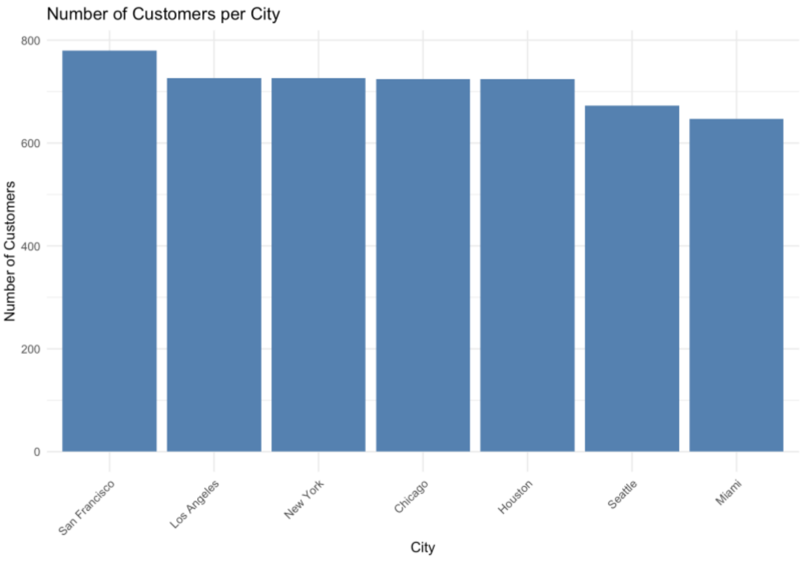


Figure 2 Customers per city

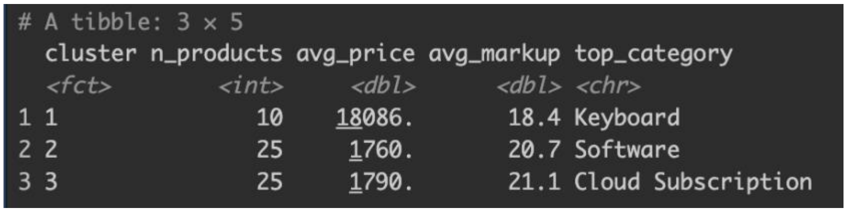
Products

The company currently offers 60 products, divided evenly across six categories with ten products each. Selling prices span a wide range, from around 350 at the lower end to almost 20 000 for premium items, while markup percentages fall between 10.1% and 29.8%. This wide spread highlights a clear distinction between high-value, low-margin goods and more affordable, higher-margin offerings.

To explore these patterns further, k-means clustering was applied using product category, selling price, and markup as the defining features. The analysis identified three main product groups. Cluster 1 is dominated by expensive items that are characterized by relatively low markup percentages, consistent with many industries where high ticket products rely on volume or exclusivity rather than margin to drive profitability (see Table 1).

Clusters 2 and 3, in contrast, group products with significantly lower selling prices but consistently higher markup percentages. Interestingly, these clusters are led by Software and Cloud subscription products respectively, reinforcing the trend that the highest markups are realised from intangible, service-based offerings. This not only reflects the scalability and lower production costs of digital products but also suggests that the company's profitability is more reliant on subscription-type revenue streams than on one-off sales of high-value goods. Moreover, this clustering result indicates opportunities for the business: while premium items anchor the company's reputation and generate substantial revenue, the growth of high-margin intangible products could play a strategic role in improving overall profitability.

Table 1 Product clusters



```
# A tibble: 3 × 5
  cluster n_products avg_price avg_markup top_category
  <fct>        <int>     <dbl>      <dbl> <chr>
1 1               10    18086.       18.4 Keyboard
2 2               25     1760.       20.7 Software
3 3               25     1790.       21.1 Cloud Subscription
```

Sales

The number of units sold per month over the two years of data available shows us that there is a huge dip in sales over summer from November to January as seen in figure 3. This is vital information for demand forecasting. It is important to prepare for this and not have excess inventory building up. This time of year we can expect higher service levels because we have less orders to meet.

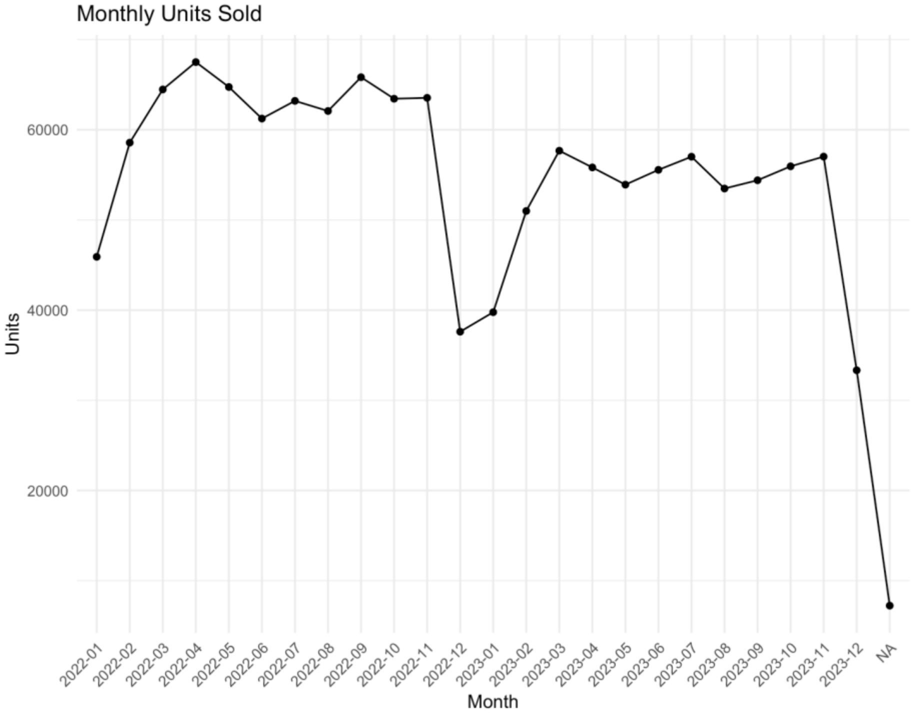
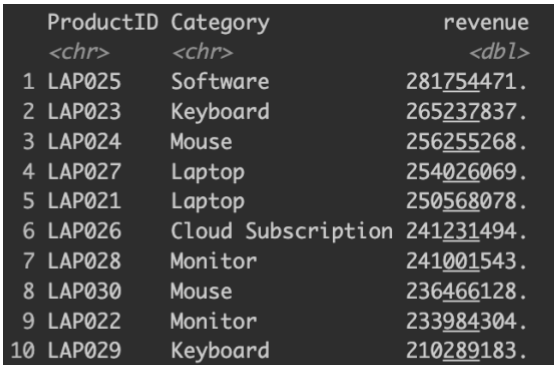


Figure 3 Units sold per month

The top ten of the sixty products by revenue were identified and their respective categories and revenues. These products should be prioritized to increase return on investment percentage.



```
   ProductID Category              revenue
   <chr>     <chr>                   <dbl>
 1 LAP025    Software           281754471.
 2 LAP023    Keyboard           265237837.
 3 LAP024    Mouse              256255268.
 4 LAP027    Laptop             254026069.
 5 LAP021    Laptop             250568078.
 6 LAP026    Cloud Subscription 241231494.
 7 LAP028    Monitor            241001543.
 8 LAP030    Mouse              236466128.
 9 LAP022    Monitor            233984304.
10 LAP029    Keyboard           210289183.
```

Figure 4 Top ten products by revenue

A revenue by category of products bar chart was also generated which shows that no single category dominated the revenue however Laptops and Monitors sales do generate slightly mor revenue than the other product categories. This makes sense because those are the highest value items.

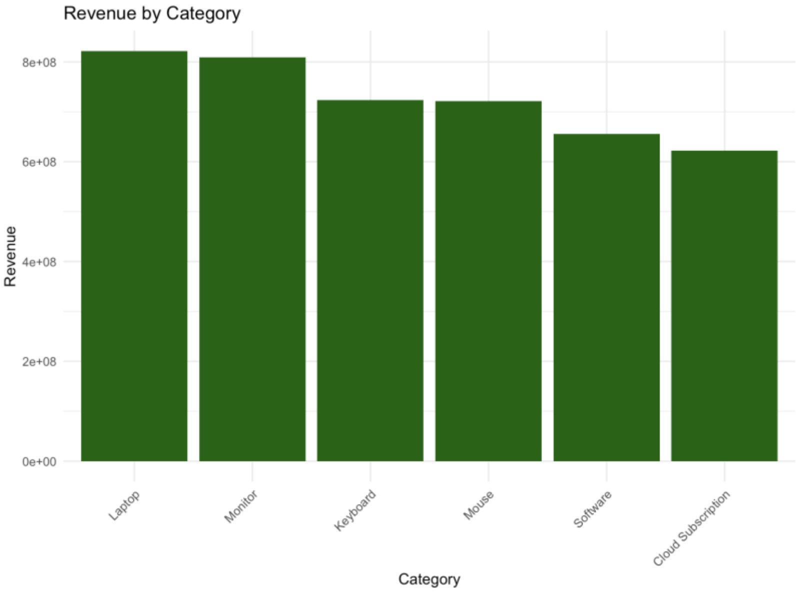


Figure 5 Revenue per category

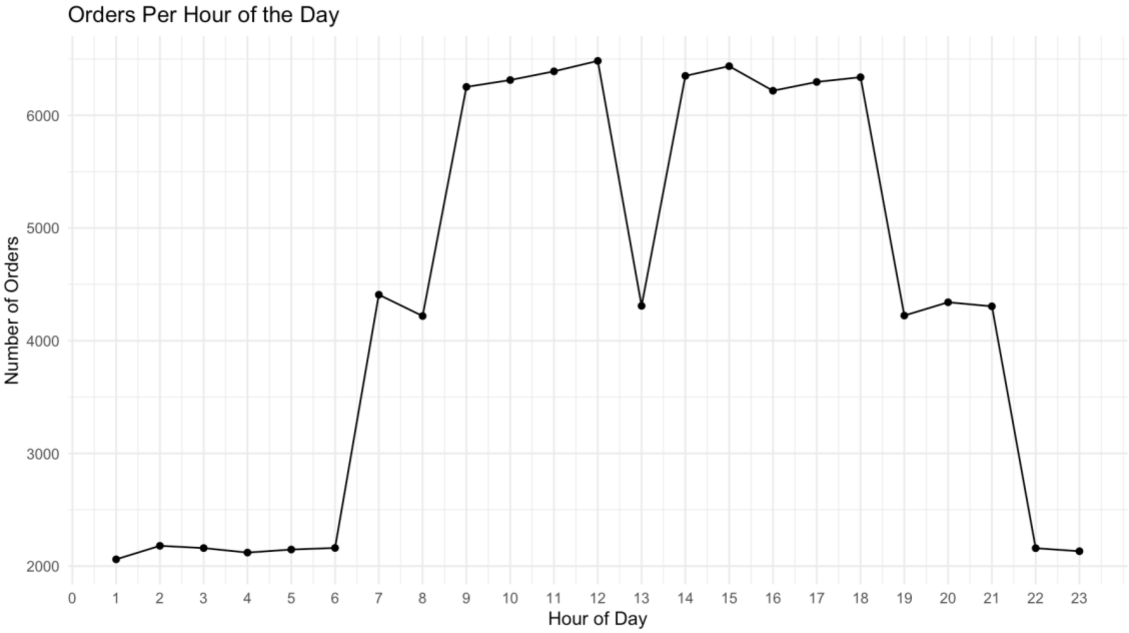


Figure 6 Orders per hour

Figure 6 shows us that the majority of orders are received between 6am and 10PM with a dip at lunch time at 1pm. This suggests the packing staff should be on shift between these hours but that for the first and last hours of this time period, the expected orders drop from approximately 6500 per hour to 4500 per hour so only 70% of the staff are required between 6am-7am and 9-10pm.

Sales with product and customer data

A bar chart of total revenue per age group of customer was created to reveal that sales by customers over the age of 60 dominate the total revenue as seen in the Age Group vs Product Category Sales chart in figure 6. This means that the oldest customers are very valuable to the company. It is visible that laptop and monitors are bringing in the most revenue each age group of sales.

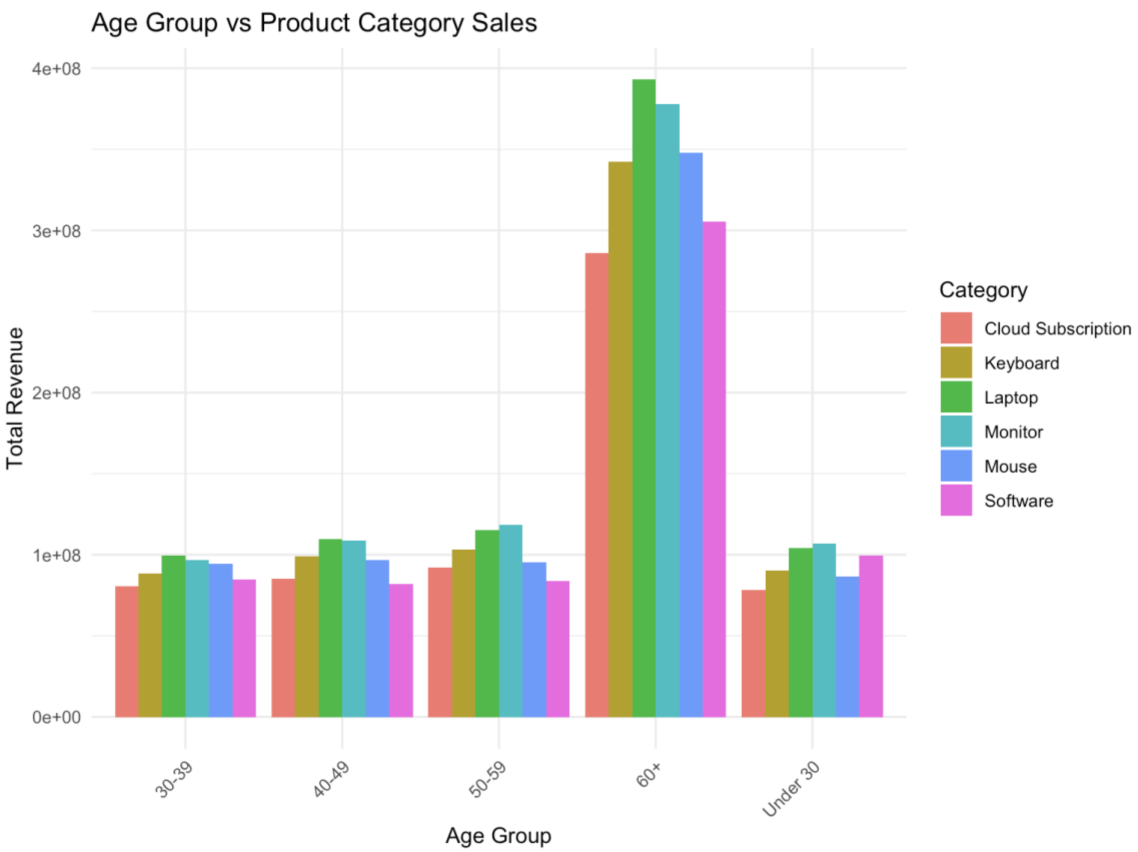
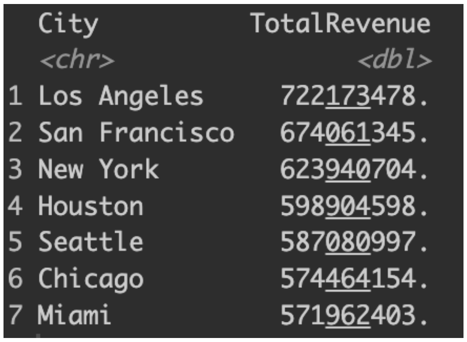


Figure 7 Revenue per age group pet product category

The following table was generated to show that Los Angeles followed by San Francisco brings in the most revenue however the other cities follow closely. Advertising in Los Angeles would be most value for money.

Table 2 Revenue per city



```
  City          TotalRevenue
  <chr>                <dbl>
1 Los Angeles     722173478.
2 San Francisco   674061345.
3 New York        623940704.
4 Houston         598904598.
5 Seattle         587080997.
6 Chicago         574464154.
7 Miami           571962403.
```

Statistical Process Control Analysis

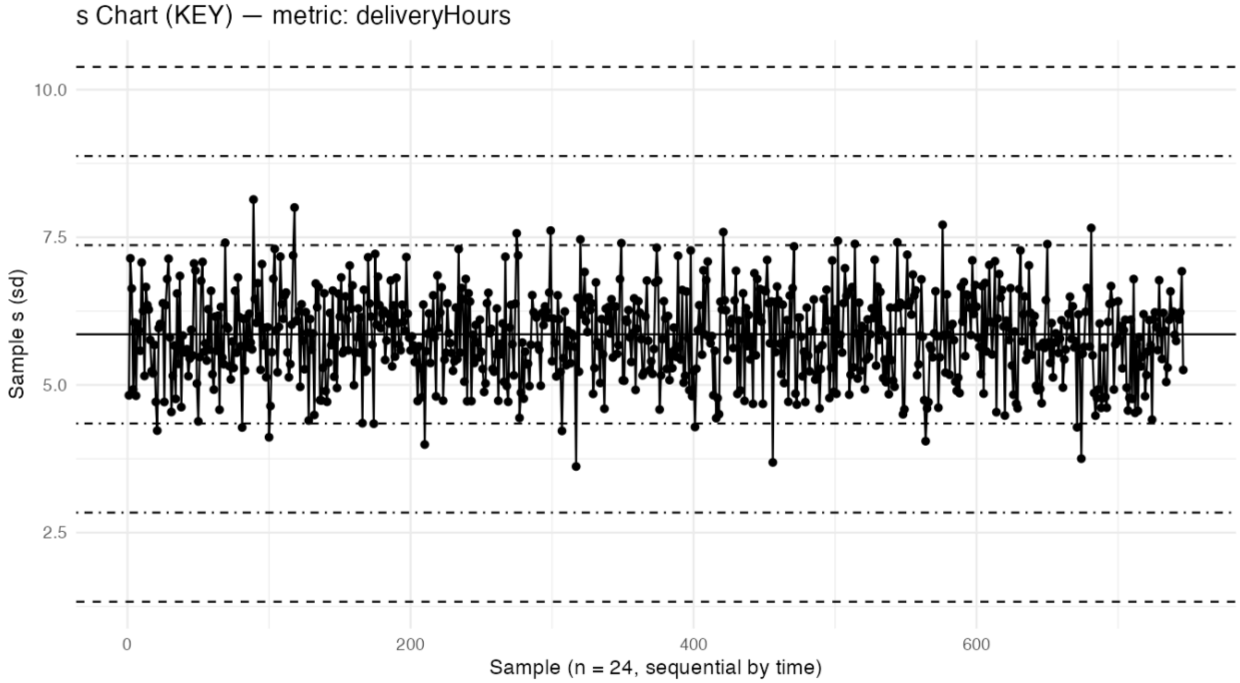
Overview

The data consists of 60 different products with a range of 964 to 2119 sales per product totally 100 000 sales.

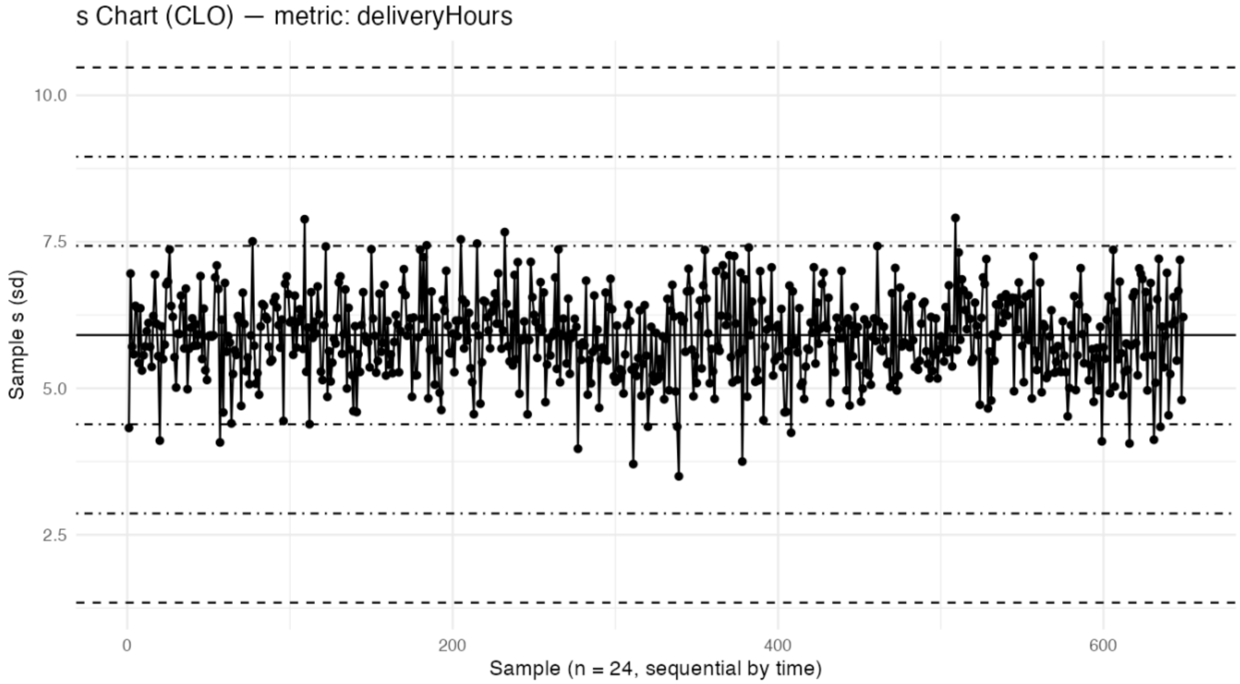
X-bar and S charts

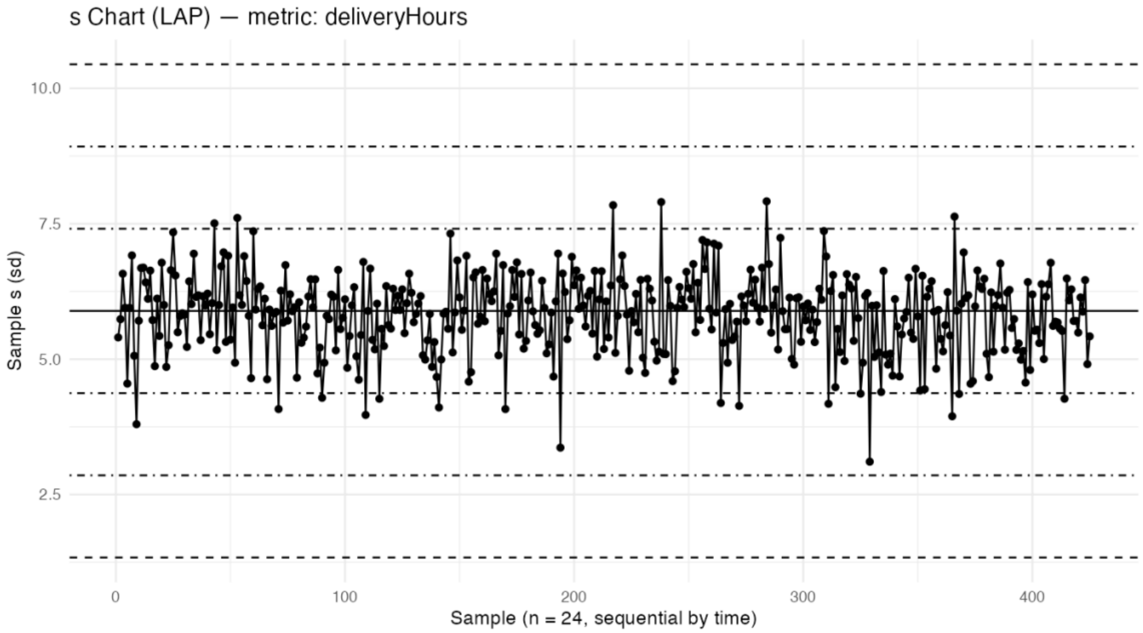
The following six X-bar and six S chart were generated per product category using the first 30 samples of 24 observations each to estimate centerlines and 1σ/2σ/3σ (outer) limits.

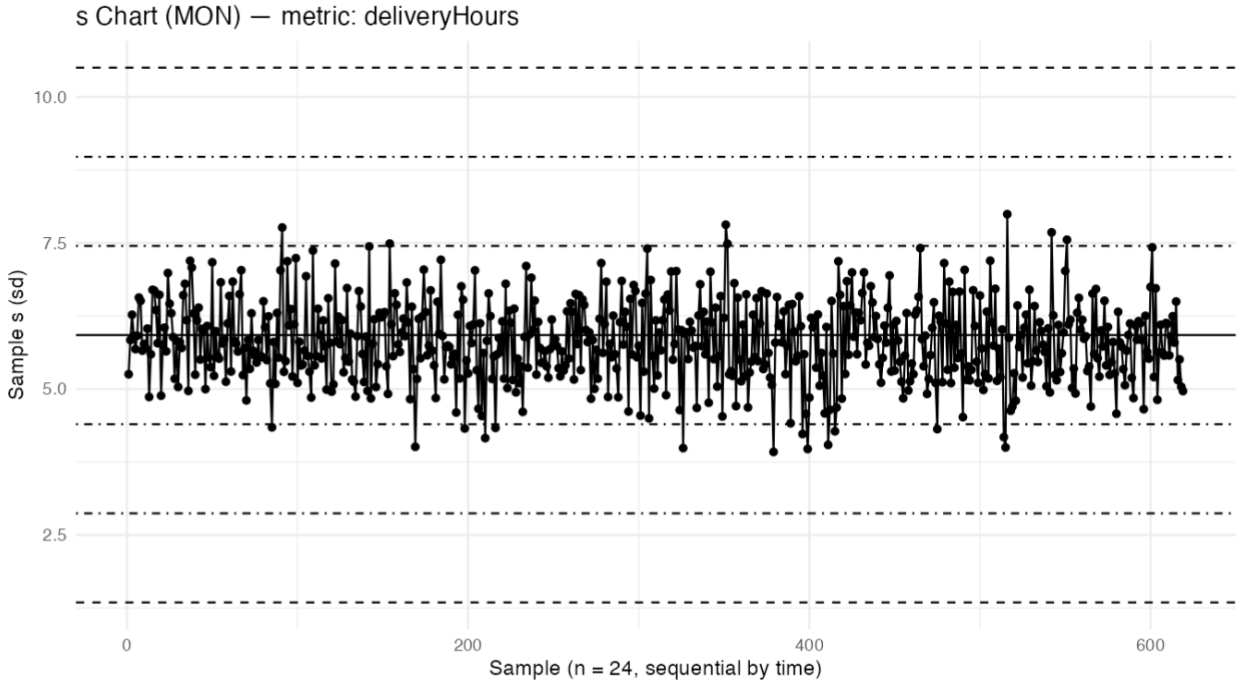
The chart plots the sample standard deviation (s) of delivery Hours for each subgroup (n = 24 samples per subgroup). The solid center line (CL) is the average subgroup standard deviation, representing the long-term process variability. The dashed lines mark the 3-sigma control limits (UCL and LCL), while the dot-dash lines show the ±2σ and ±1σ zones.

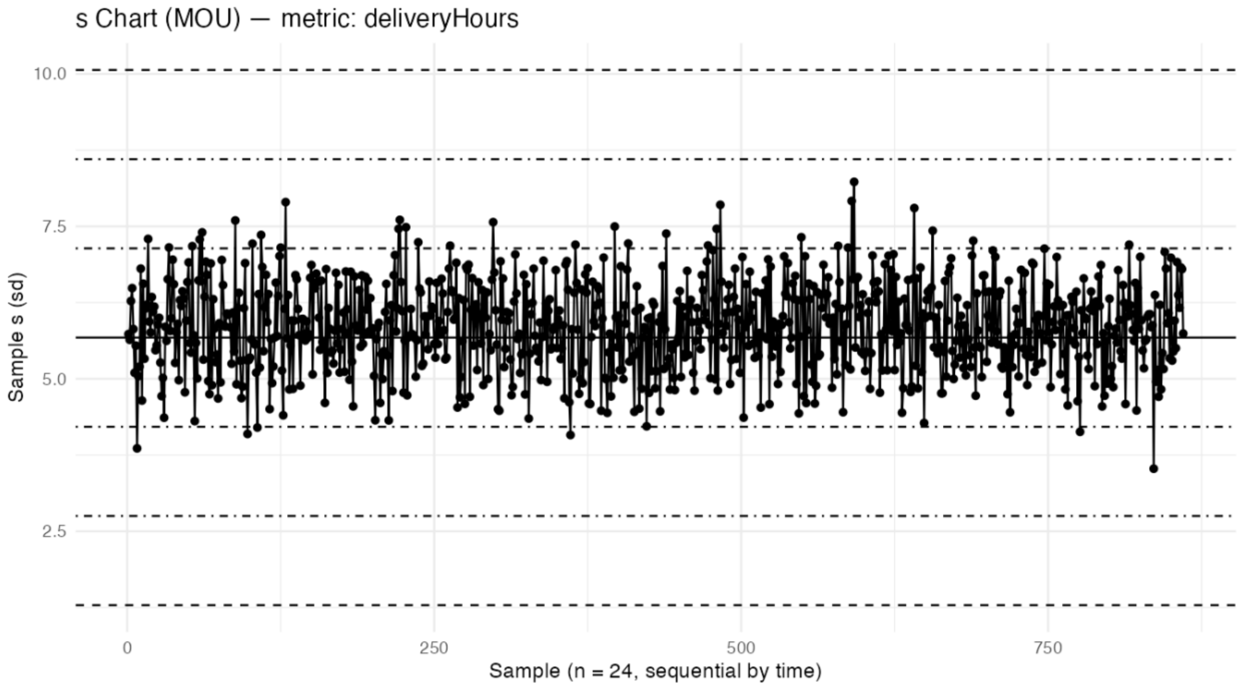


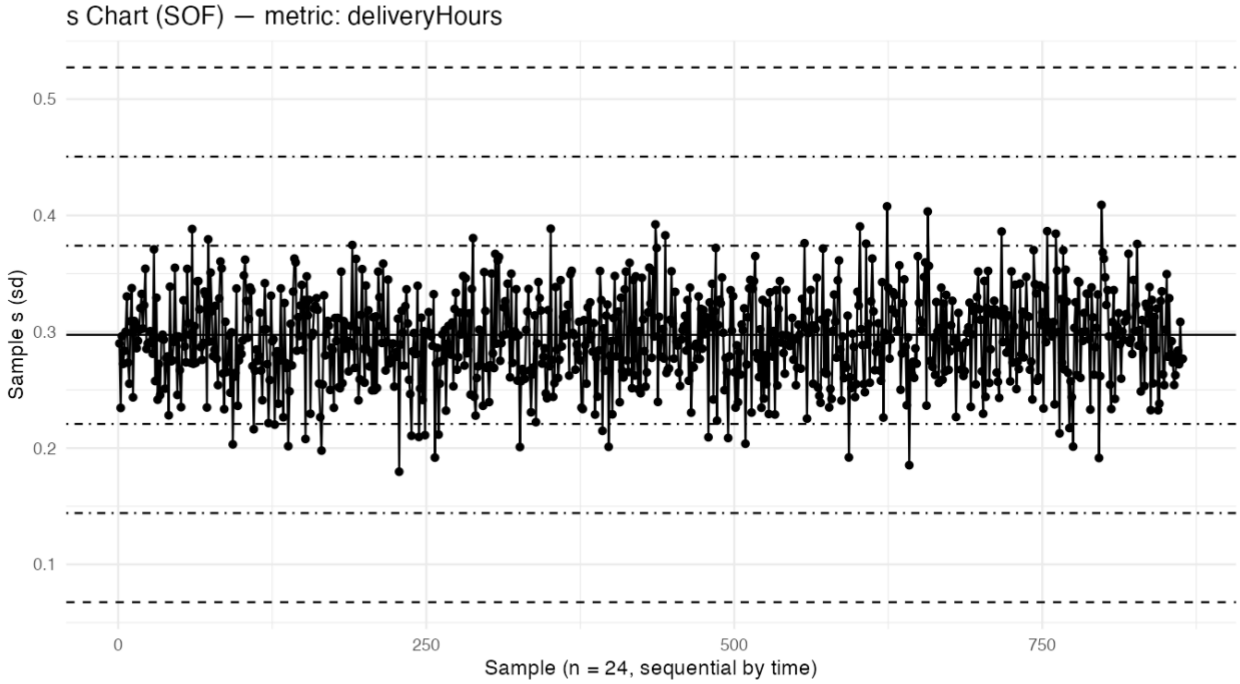
All samples of Keyboards lie within the 3-sigma control limits and there are no visible trends or runs above/below the center line. The points are well distributed about the center line which is typical of noise/common variance and not bias. The fact that there is no clustering suggests good consistency in delivery performance which means over-controlled nor unstable.

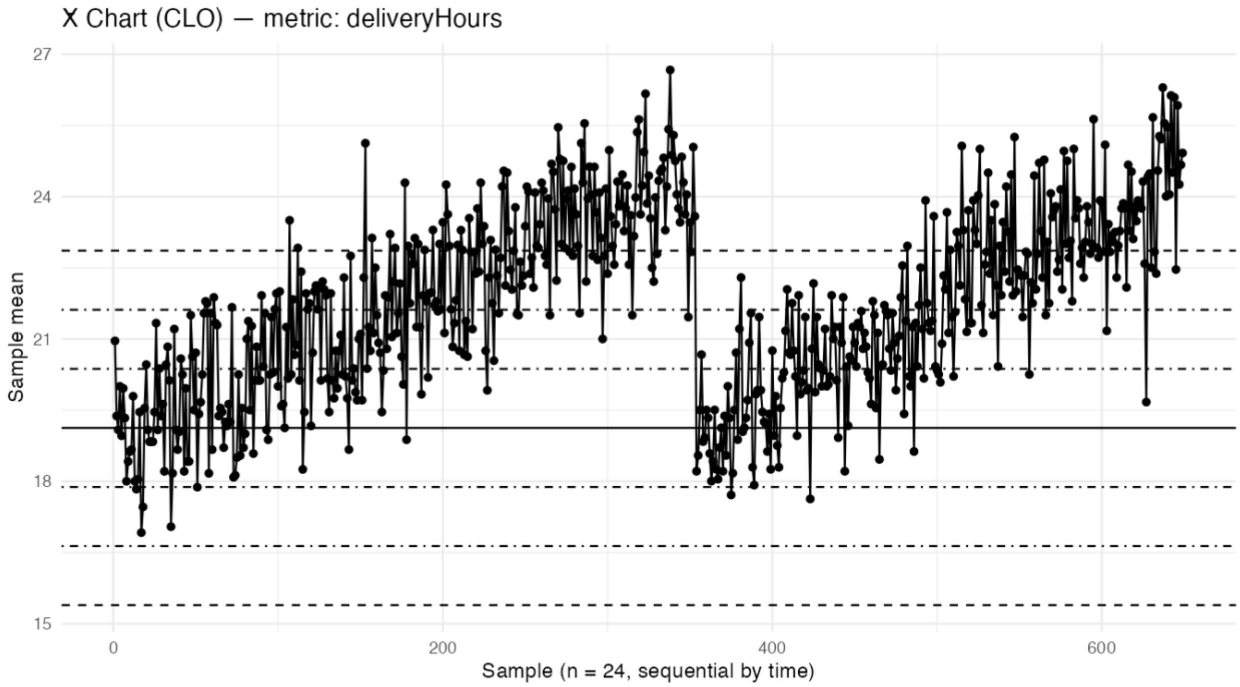



s Chart (LAP) — metric: deliveryHours
Sample s (sd)
10.0
7.5
5.0
2.5
0
100
200
300
400
Sample (n = 24, sequential by time)

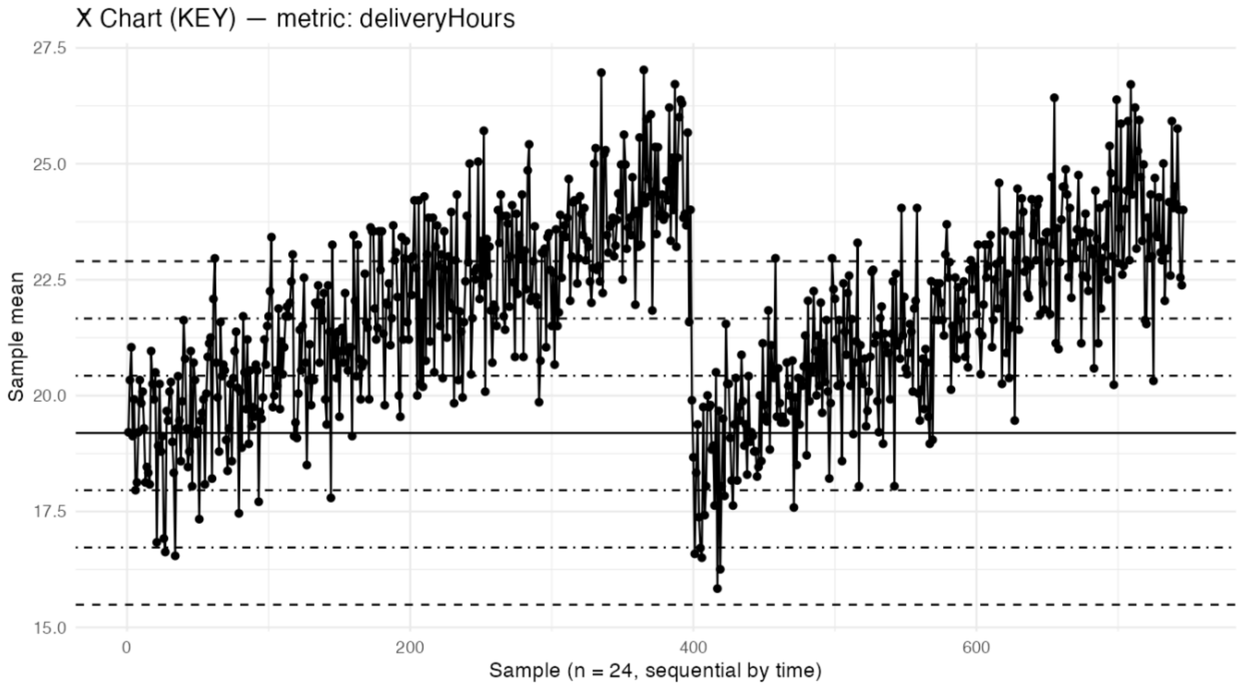

s Chart (MON) — metric: deliveryHours
Sample s (sd)
10.0
7.5
5.0
2.5
0
200
400
600
Sample (n = 24, sequential by time)

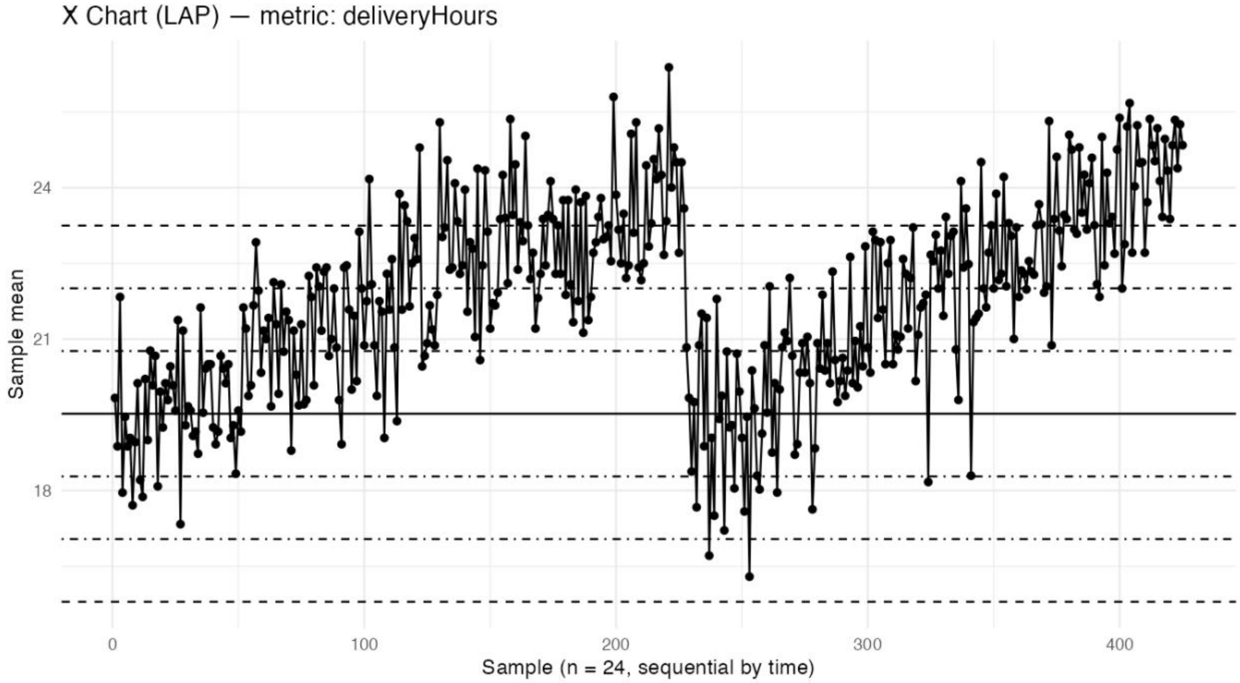

s Chart (MOU) — metric: deliveryHours
10.0
7.5
5.0
2.5
Sample s (sd)
0
250
500
750
Sample (n = 24, sequential by time)

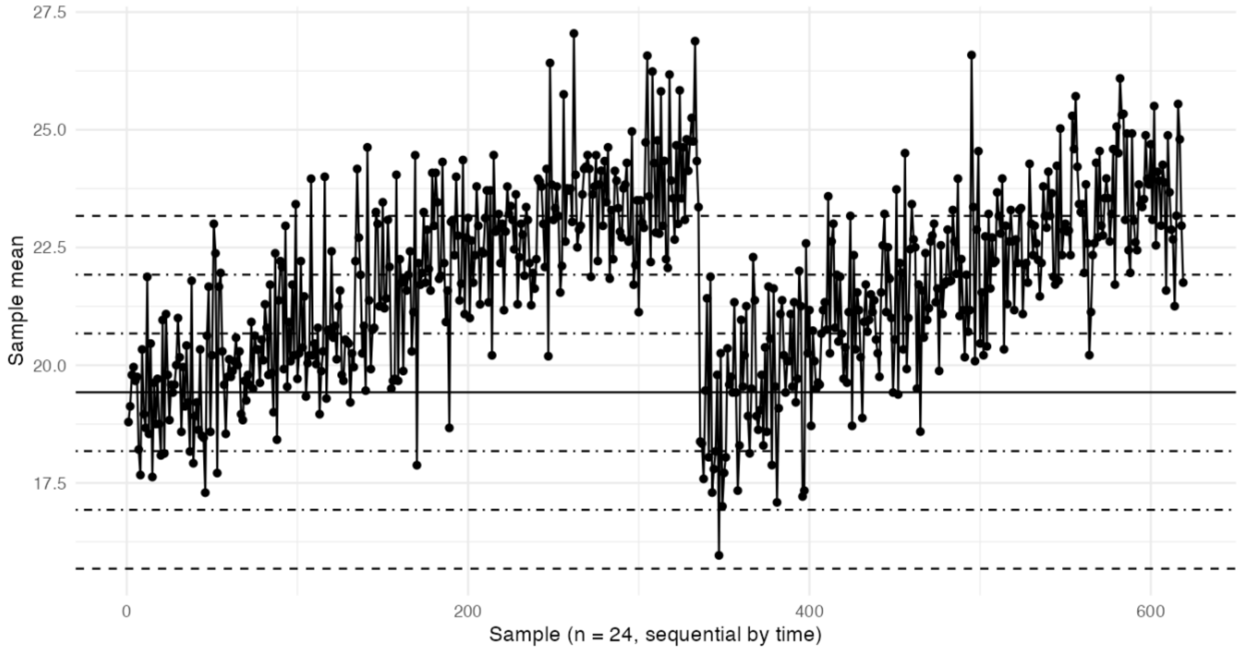

s Chart (SOF) — metric: deliveryHours
0.5
0.4
0.3
0.2
0.1
Sample s (sd)
0
250
500
750
Sample (n = 24, sequential by time)



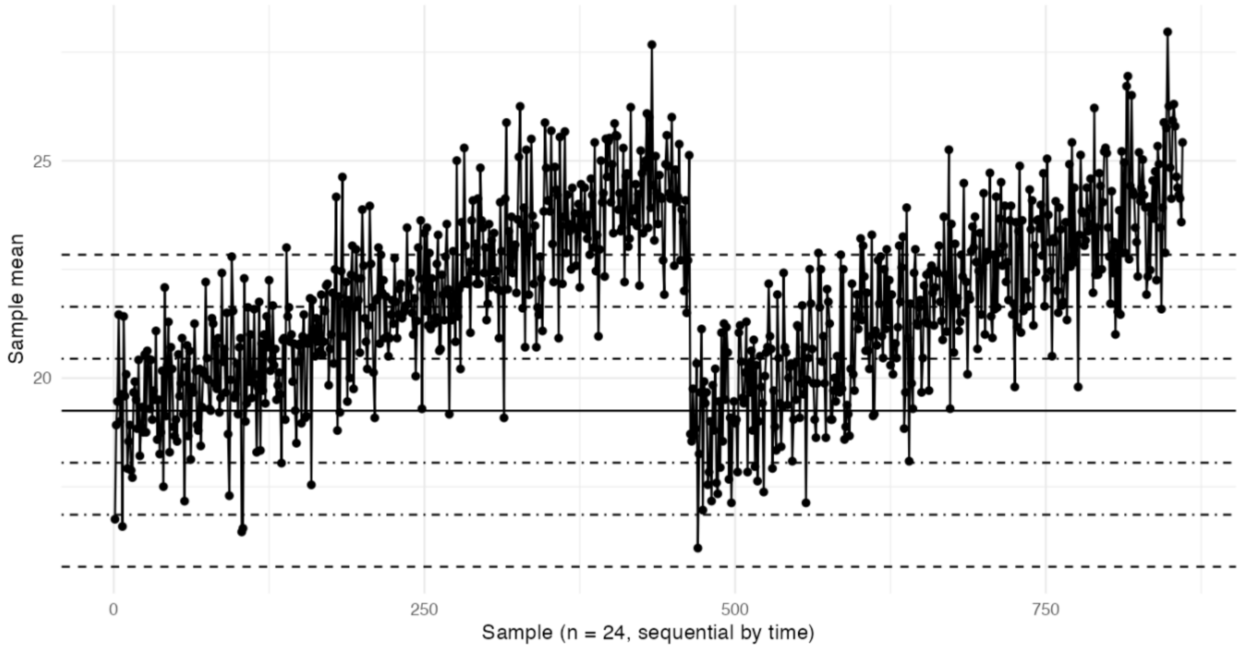
This jump in sample mean visible in all X-bar charts represents the change between 2026 and 2027 where the sample mean days of delivery is significantly reduced. Possible contributing factors include the implementation of new managerial procedures or the resolution of prior backlogged orders, resulting in a reset of delivery operations.


X Chart (KEY) — metric: deliveryHours
Sample mean
27.5
25.0
22.5
20.0
17.5
15.0
0
200
400
600
Sample (n = 24, sequential by time)

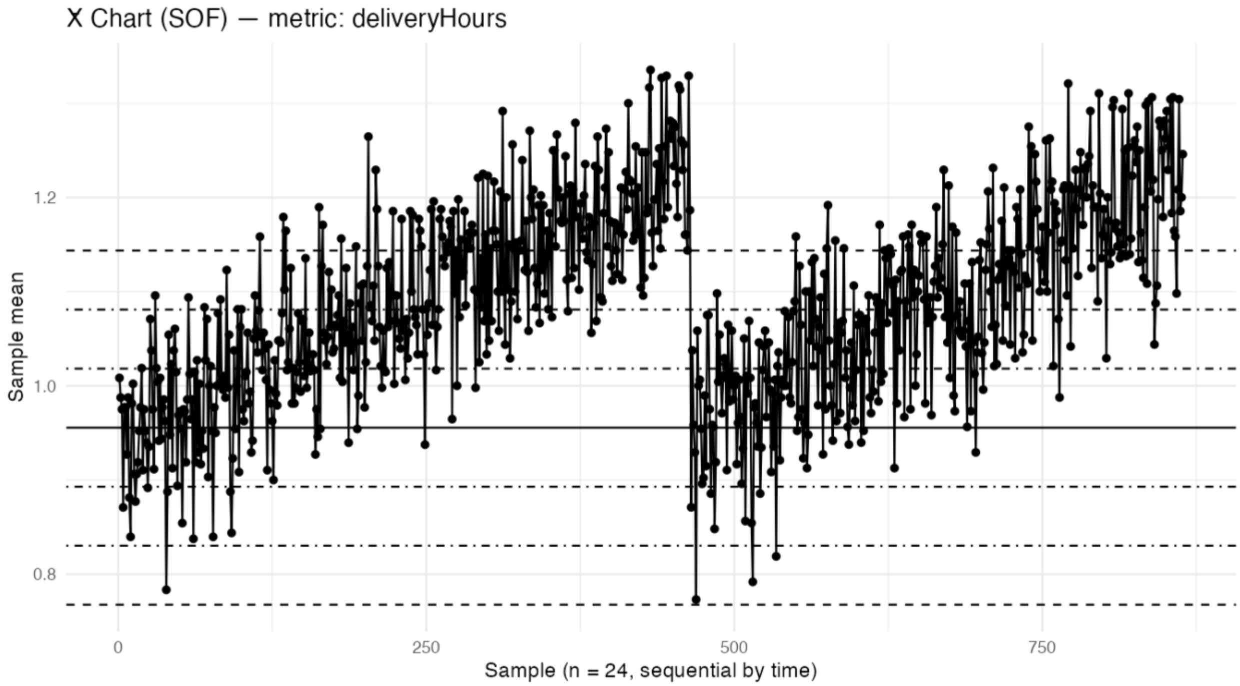

X Chart (LAP) — metric: deliveryHours
Sample mean
24
21
18
0
100
200
300
400
Sample (n = 24, sequential by time)



X Chart (MON) — metric: deliveryHours



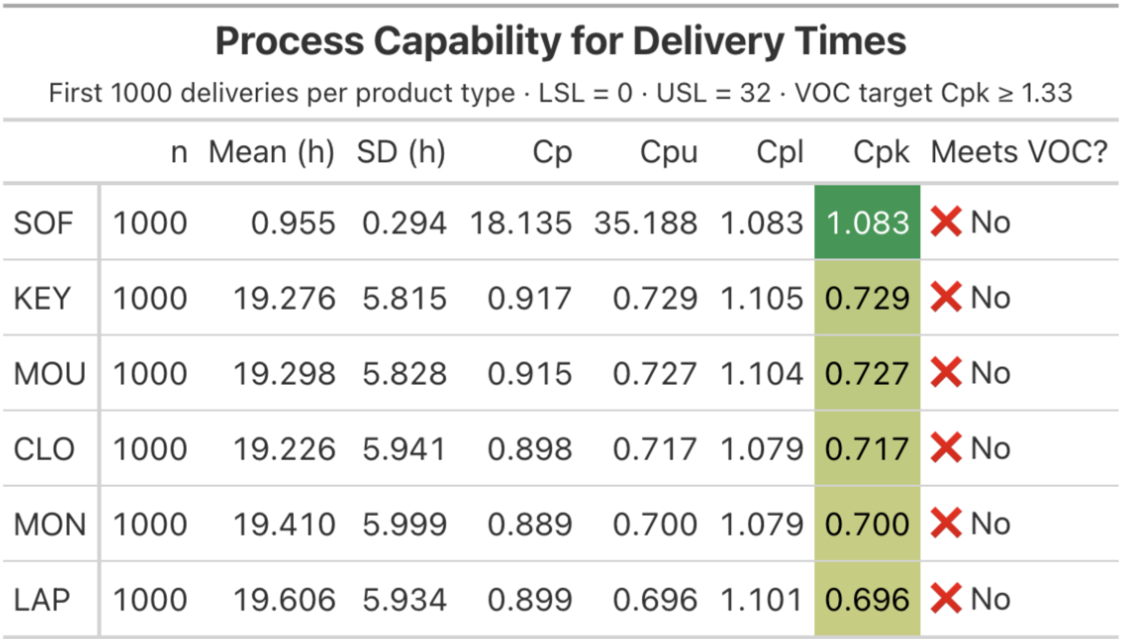
X Chart (MOU) — metric: deliveryHours


X Chart (SOF) — metric: deliveryHours
Sample mean
1.2
1.0
0.8
0
250
500
750
Sample (n = 24, sequential by time)

Process Capabilities Indices

The table was generated by sorting each product type's delivery records by time, taking the first 1 000 deliveries per type, and computing the mean and standard deviation of delivery time to calculate capability indices (Cp, Cpu, Cpl, and Cpk) against LSL = 0 h and USL = 32 h. A product type is considered capable of meeting the VOC (voice of customer) if Cpk ≥ 1.33.

Higher Cpk indicates more centered and consistent delivery performance within the 0–32 h window (assuming a stable, approximately normal process and using overall sigma from those 1 000 observations).



Process Capability for Delivery Times

First 1000 deliveries per product type · LSL = 0 · USL = 32 · VOC target Cpk ≥ 1.33

	n	Mean (h)	SD (h)	Cp	Cpu	Cpl	Cpk	Meets VOC?
SOF	1000	0.955	0.294	18.135	35.188	1.083	1.083	❌ No
KEY	1000	19.276	5.815	0.917	0.729	1.105	0.729	❌ No
MOU	1000	19.298	5.828	0.915	0.727	1.104	0.727	❌ No
CLO	1000	19.226	5.941	0.898	0.717	1.079	0.717	❌ No
MON	1000	19.410	5.999	0.889	0.700	1.079	0.700	❌ No
LAP	1000	19.606	5.934	0.899	0.696	1.101	0.696	❌ No

Type I error rates (per opportunity):

Rule A: 3-sigma single-point rule (Xbar): alpha = 0.0027

The alpha risk of 0.0027 represents the probability of a false alarm - the chance that a point will randomly fall outside the 3-sigma limits when the process is actually in control.

Rule B: Consecutive samples between -1 and +1 sigma

5 consecutive samples: 0.148291

10 consecutive samples: 0.02199

15 consecutive samples: 0.003261

20 consecutive samples: 0.000484

Rule C: 4 consecutive samples outside upper second control limits: alpha = 2.678772e-07

Type II error (β) and power for the specified shift:

β = 0.8412 (missed detection)

This means that about 84 % of the time, the chart will *not detect* this moderate process shift — it will show “in control” even though the process mean has changed.

1-β = 0.1588 (power)

The complement means the probability the chart *does* detect the shift is only about 16 %

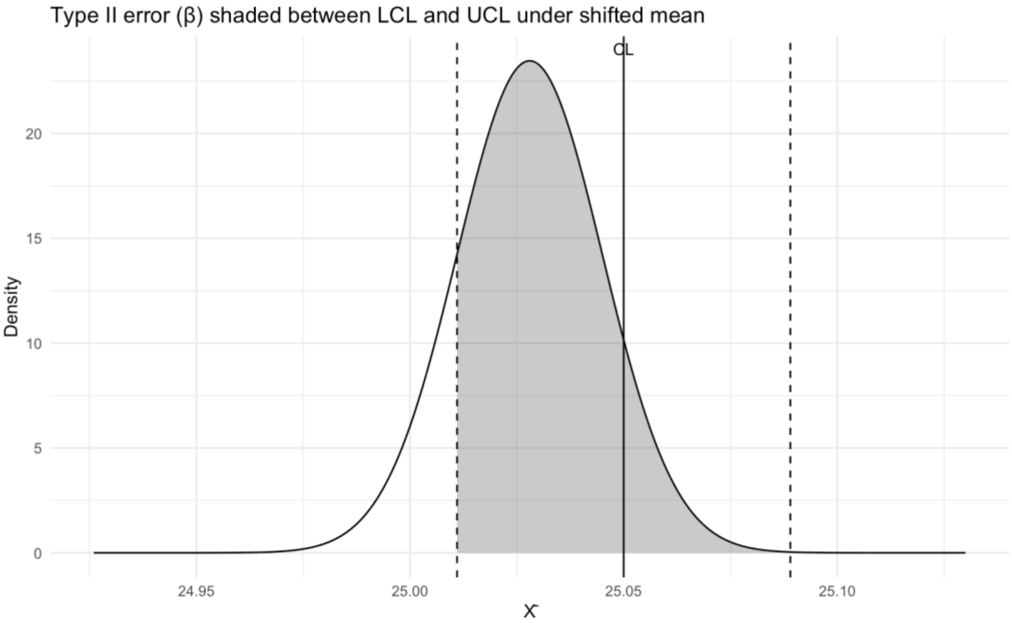


Figure 8 Type II error

The process mean has **shifted** (to 25.028 L), but the X̄ chart still treats anything between 25.011 and 25.089 as “in control.” The shaded area is therefore the probability that, **despite the shift**, a random sample mean still falls between those limits.

Fixing errors

It was discovered that there is a difference between the Head office data naming convention and the sales data which has led to a few issues in our analysis. The `ProductID` for products with a category starting with "NA" are adjusted to have a prefix based on the category. This step corrects issues where the `ProductID` was incorrectly starting with "NA" instead of a valid prefix. Then the Selling Price and Markup Values were updated according to the product data file, the total sales value of all sales that took place in 2023 was calculated as follows:

Category	TotalSalesValue_2023 (Using corrected data)
Laptop	1163889479
Monitor	578385570
Cloud Subscription	98715482
Keyboard	73499067
Software	66468485
Mouse	51219577

Grand total sales value for 2023: **2,032,177,660**

For the 2023 sales before the corrections were made we were only able to calculate total sales for Software due to the different naming conventions between the Head office data and the sales data. The value for total sales of Software is less due to the old uncorrected selling price values.

Category	TotalSalesValue_2023 (Using old data)
Software	61821700

Now with the corrected selling prices we can see that although Laptops and Monitors were appearing to result in highest revenue from calculations with the old data, we can now see how significant this difference in revenue brought in per product type really is.

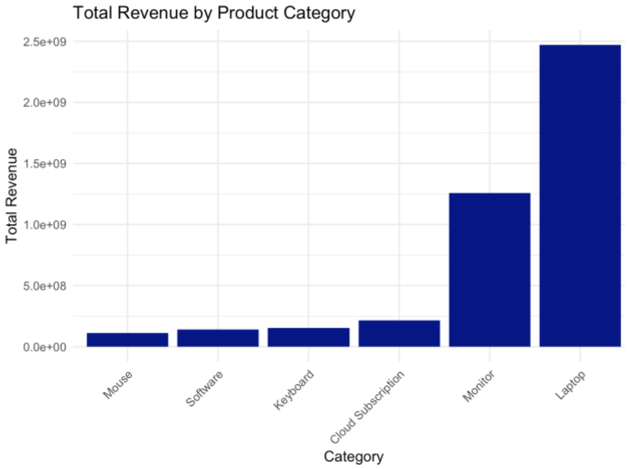


Figure 9 Revenue per category (Corrected data)

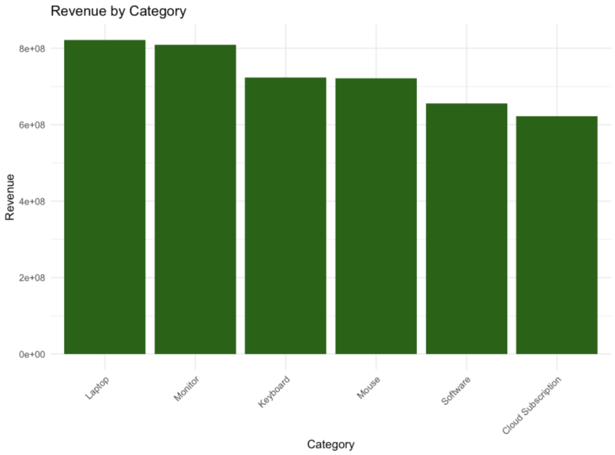


Figure 10 Revenue per category (Old data)

Optimizing profit for coffee shop 1 and 2

Data on the service time and number of baristas working per day for two coffee shops are analyzed below.

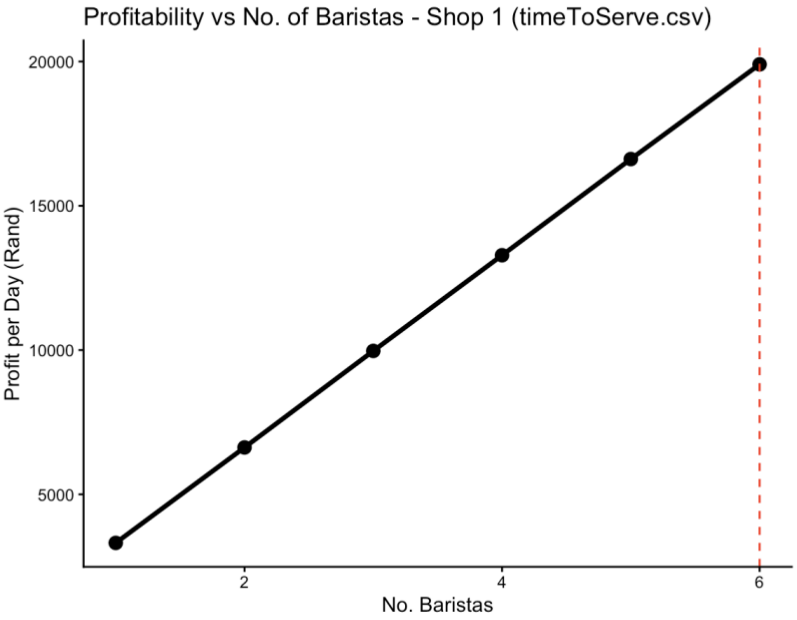


Figure 11 Shop 1 optimal profit:

The plot for Shop 1 (figure 11) shows a clear linear increase in profit as the number of baristas increases from 2 to 6. A maximum daily profit of approximately R20 000 is obtained with 6 baristas. This suggests that increasing the number of baristas directly leads to higher profits, because the number of customers served increases, while the personnel costs grow at a slower rate. This relationship between Service time and No. Baristas is also show in figure 12 below.

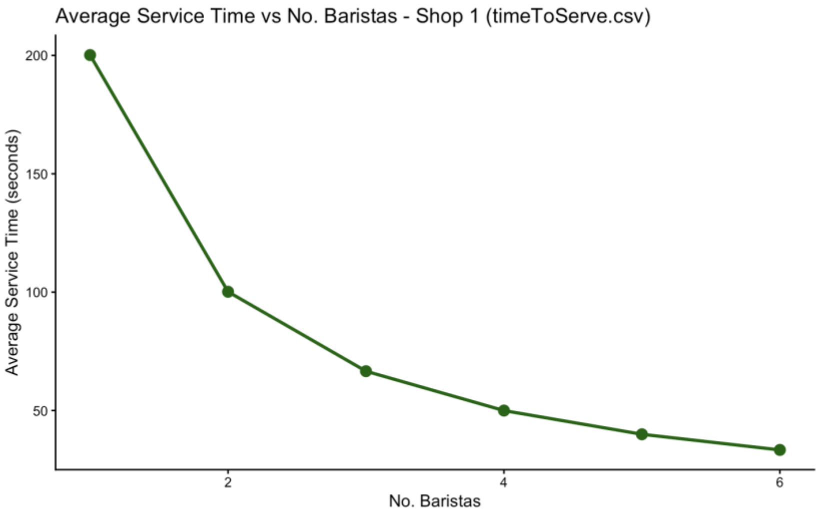


Figure 12 Shop 1 Service time vs no. baristas

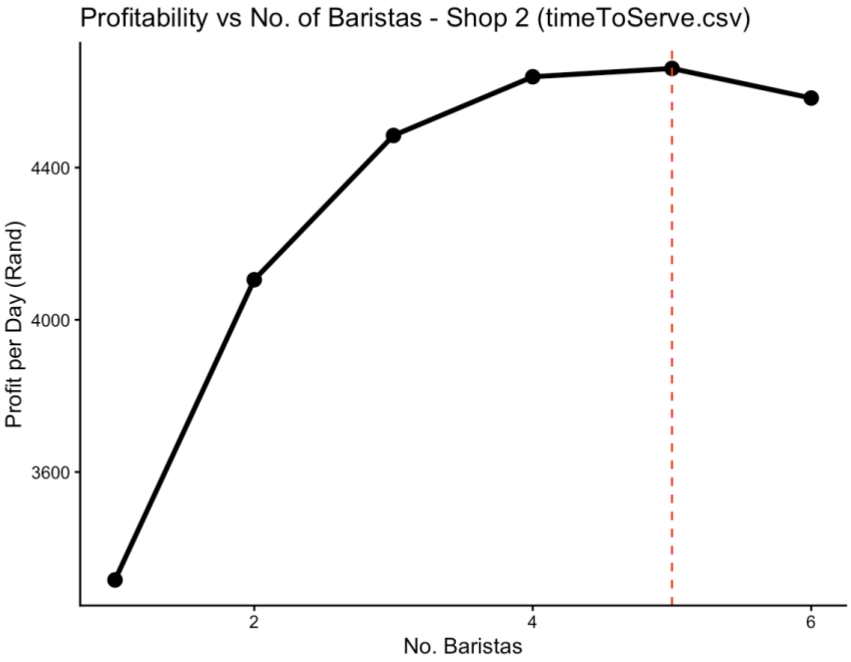


Figure 13 Shop 2 optimal profit:

Figure 12 shows an increasing trend in profit as the no. of barista increase with a peak at 5 baristas with an approximate daily profit of R4600. The red dashed line at 5 baristas marks the optimal number of baristas.

In figure 14 it is visible that the slope of service time vs no. baristas for shop 2 is less steep than shop 1 and that the average service times are longer for shop 2 which explains why the the profit for shop 2 peaks before 6 baristas. This suggests that shop 2 is less productive in terms of service time which is why it is not worth it to employee a sixth barista.

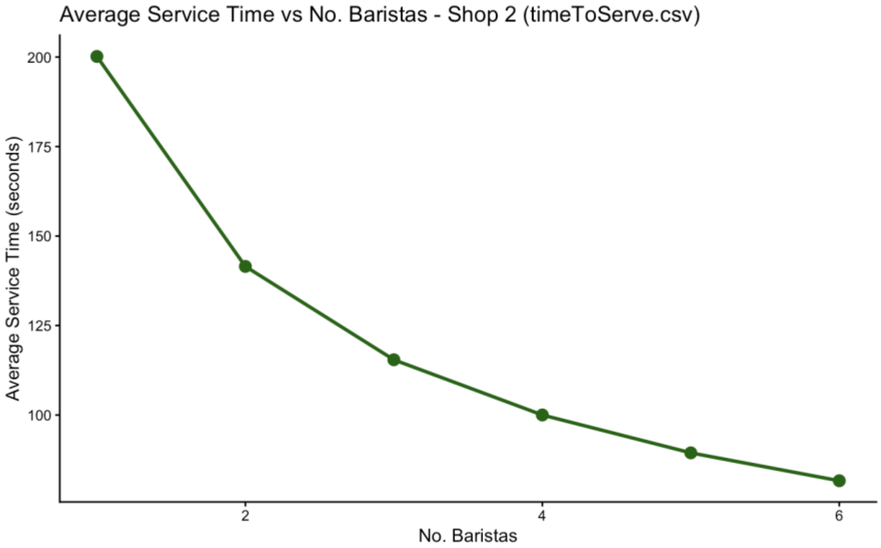


Figure 14 Service time vs no. baristas

Part 6: ANOVA

ANOVA Table

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
orderYear	1	454	454.1	2.398	0.121
Residuals	99998	18933479	189.3		

The p-value of 0.121 in the ANOVA suggests no significant difference between 2022 and 2023. So **fail to reject the null hypothesis**; there's no evidence that `orderYear` affects `Quantity`.

Descriptive Statistics:

orderYear	Quantity Mean	Quantity SD	Quantity N
2022	13.44093	13.74713	53,727
2023	13.57608	13.77500	46,273

These statistics suggest that the average `Quantity` is very similar between 2022 and 2023, with a slight increase in 2023, but this difference is not statistically significant based on the ANOVA results.

Levene's Test:

Source	Df	F value	Pr(>F)
group	1	1.3642	0.2428
Residuals	99,998		

The p-value of **0.2428** suggests that the assumption of equal variances between the two groups is not violated, so you can trust the results of the ANOVA.

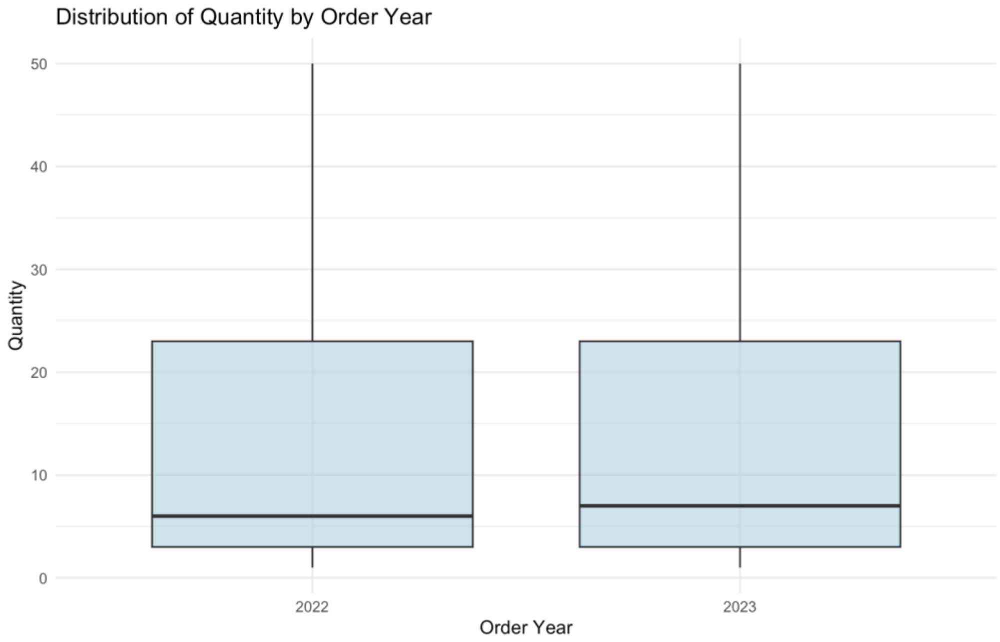


Figure 15 Box plots of product quantities per year

Figure 11 visually represents the fact that year 2022 and 2023 quantity of orders per product are very similar. The range for both years appears the same while the mean quantity for 2023 is approximately 2 units higher.

Part 7: Reliability of service

The reliability and optimal number of workers at a car rental company are discussed.

Estimating Reliable Service Days (7.1):

Reliable service days in the year:	366
Percentage of reliable service days in the year:	92.19144 %

The **92% reliability** in terms of staffing is quite high. The company can expect **366 days** out of the **397 total days** to have reliable service, assuming there are 15 or more workers present. This indicates that, almost every day, the company has a sufficient number of workers to avoid issues related to under-staffing.

Optimizing Profit (7.2):

Total profit loss due to insufficient workers:	-725833.3

This is the financial impact (loss) caused by having insufficient workers on duty (fewer than 15 workers). The profit loss is attributed to the sales decrease and the cost of hiring additional personnel to mitigate the problem. This indicates a significant cost for the company. The company may want to consider reducing this loss by hiring more personnel or optimizing its staffing levels to ensure at least 15 workers are available each day. Reducing days with fewer workers would minimize the sales loss and help increase overall profitability.

Conclusion

This data analysis provides a multi-faceted view of the company's operations, yielding actionable insights across marketing, product strategy, and logistics. The key findings indicate that marketing efforts should be strategically focused on high-income clusters and the large segment of younger customers, while also nurturing the highly valuable over-60 demographic. Geographically, Los Angeles and San Francisco present the highest-return opportunities for targeted advertising.

From a product perspective, the company benefits from a balanced portfolio. High-ticket items like laptops and monitors drive substantial revenue and should be a strategic growth focus. Operationally, the pronounced seasonal sales dip necessitates proactive inventory management to avoid overstock.

A critical conclusion from the SPC analysis is that the current delivery process for all product categories is not capable of consistently meeting customer expectations. The failure to achieve a Cpk of 1.33 across the board signals an urgent need for process improvement in the supply chain to enhance customer satisfaction and reduce operational risk. In summary, by using these data-driven insights and targeting key customer segments, capitalizing on high-margin products, managing seasonal demand, the company can make informed decisions to drive growth, increase profitability, and strengthen its competitive position.

References

InTouch Insight. (2024). *American coffee preferences: a brief insight*. [online] Available at: https://www.intouchinsight.com/blog/american-coffee-preferences-a-brief-insight [Accessed 14 October 2024]

Sigma Zone. (2024). *Calculating Type I Probability*. [online] Available at: https://sigmazone.com/calculating-type-i-probability/ [Accessed 20 October 2024].