ECSA Report:

Data Analysis and manipulation outcome in Industrial Engineering.

Student number: 26162466

Date: 22 October 2025

Contents

Introduction

This report presents a comprehensive analysis of data-driven decision-making in the context of Industrial Engineering. The objective of this report is to improve Quality Assurance by leveraging data analysis, process optimization, and statistical techniques to aspects such as identifying inefficiencies, optimizing staffing levels, and ensuring consistent service delivery, ultimately improving overall operational quality and profitability. The analysis makes use of various data-driven techniques to assess and enhance business operations, using RStudio and Excel for data analysis.

The report begins with a thorough descriptive data analysis, where data wrangling and statistical summary techniques are applied to explore key trends, distributions, and correlations. This is followed by an exploration of Statistical Process Control (SPC) methods to assess process stability; the Process Capability Indices (Cp and Cpk) are calculated to understand whether the process is operating within the required limits and identify areas for improvement.

Next, the report investigates hypothesis testing, focusing on Type I and Type II errors to evaluate the reliability of the business processes. Profit optimization is also required to multiple scenarios throughout the report as well as applying ANOVA and Two-Way ANOVA to identify whether key factors significantly influence service performance and outcomes.

By applying these various analytical methods, the report aims to provide actionable insights into improving the operational efficiency and profitability of businesses through quality assurance methodology.

Part 1.2: Descriptive statistics

Information gathering

The following excel files of a given company were given to be used as part of the Descriptive statistics:

Excel file name	Column titles	Basic explanation of data file
Product_data	Product ID Category Description Selling Price Markup	This dataset provides detailed information on the company's product offerings, including variations and their corresponding market value.
Customer_data	Customer ID Gender Age	This dataset offers valuable insights into the demographics and diversity

	Income City	of the customer base, helping the company understand its market segment.
Product_headoffice	Product ID Category Description Selling Price Markup	This dataset provides the official product data from the head office, detailing product variations and associated pricing and markup information, which is essential for evaluating market value consistency across different sources.
Sales2022&2023	Customer ID Product ID Quantity Order Time Order Day Order Month Order Year Picking Hours Delivery Hours	This dataset is key for understanding company profitability, examining sales trends, and assessing customer service performance through operational metrics such as order processing and delivery times.

As can be viewed from the table above, three integrated datasets were given for analysis: sales transactions, product information, and customer profiles. The datasets were linked using common identifiers (ProductID and CustomerID) to form a consolidated view of business performance. This combined dataset allows us to evaluate not only sales activity, but also how customer demographics and product characteristics influence purchasing behaviour.

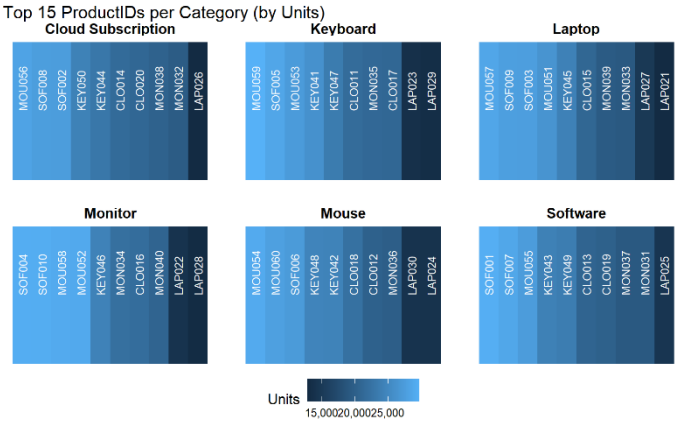
Pre-processing

The analysis began with data preparation and cleaning. Excel sheets were imported into RStudio, where column names were standardised, missing values were addressed, and data types were converted (e.g., characters to numeric where appropriate).

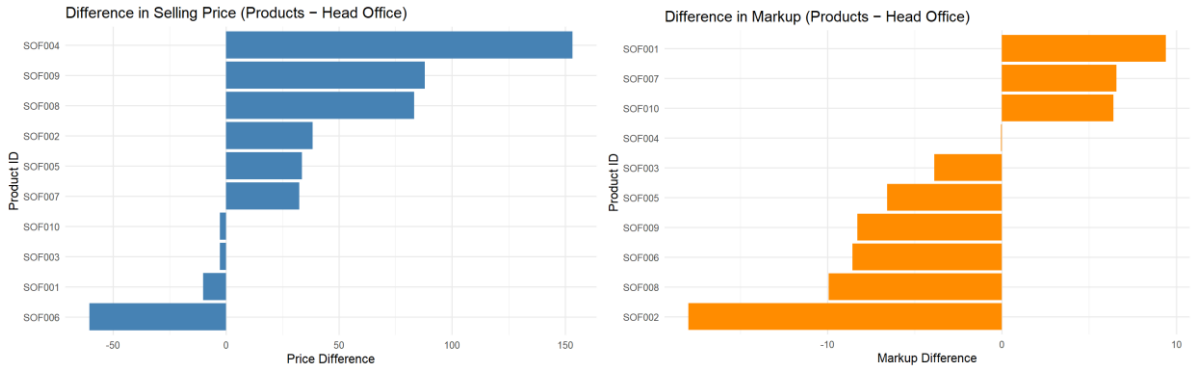
Graphs and tables are used to examine distributions, identify interesting correlations, and highlight any significant differences within the data.

Data exploration

Potential errors in data analysis due to data

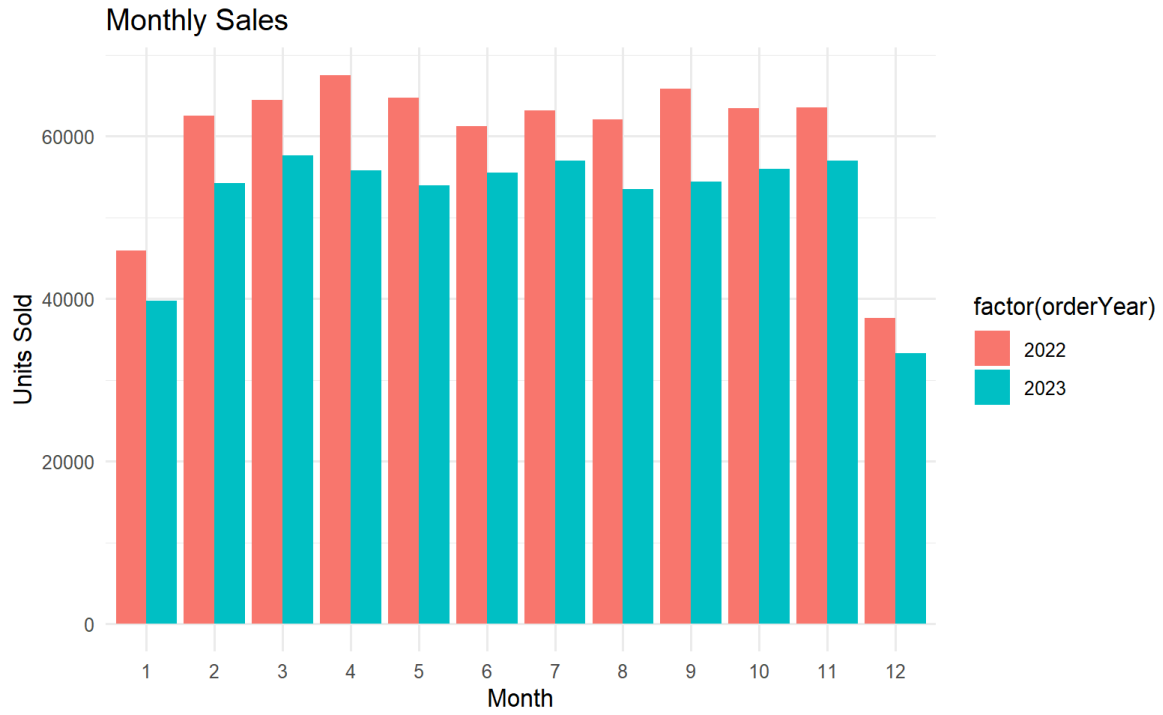


As can be see on the graphs to the left, there seems to be errors between correlating Product ID's to the correct categories. This can be extremely detrimental to the company as it tracks incorrect information between product types and ID's. In this specific example it shows how incorrect sale units are produced under each category due to this error.

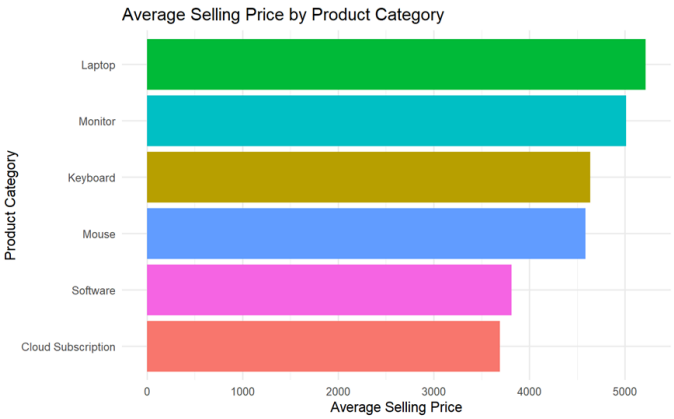


Another error to be aware of when analysing the data is the discrepancies between the "head office" data file given compared to the "product" data for some product IDs. This indicates that there is a difference between the above products in selling price and therefore mark-up as well. This will influence how the companies growth and financial come across.

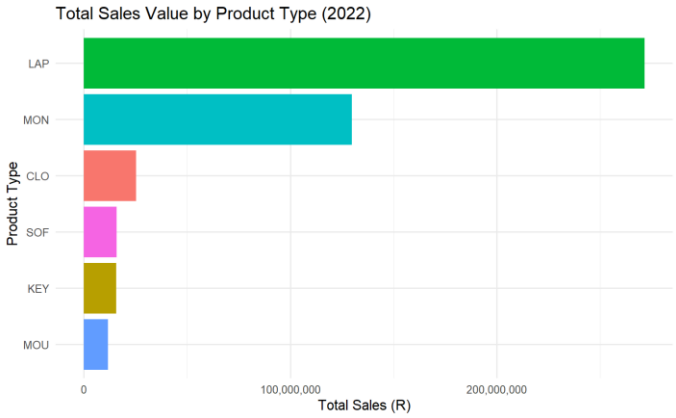
Sales

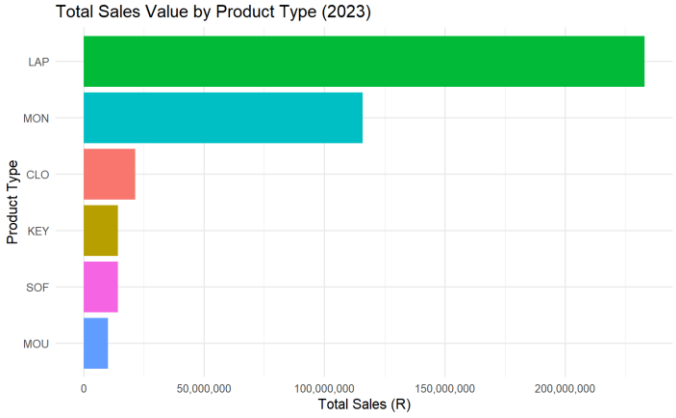


The above graph visually represents the difference between monthly sales from 2022 and 2023. An overall uniform distribution is found for both 2022 and 2023 units sold. It is interesting to note the decrease in sales in 2022 and investigate why this may have occurred in more detail to help improve the future profit of the company. In this graph it is also important to note how January and December months seem to have a significant decrease in sales. This could be due to promotional decrease or seasonal factors such as the holiday season etc, this should be considered when forecasting demand for future sales.

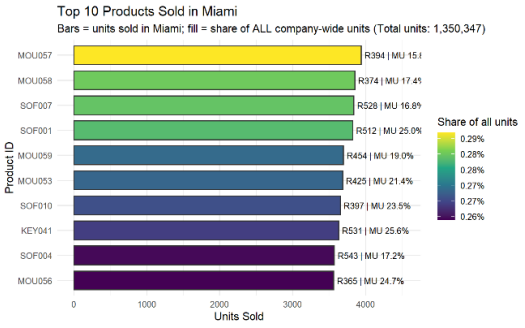


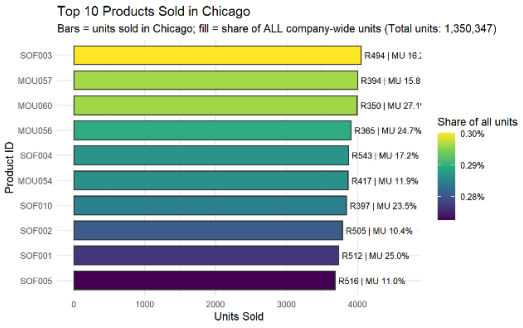
Here it shows that Laptops and Monitors are sold at the highest price however this is not enough to indicate that they bring in the highest amount of profit for that we will have to look at the mark-up as well as the amount sold per year this would give a better view of incoming revenue. The following graphs take this into account:

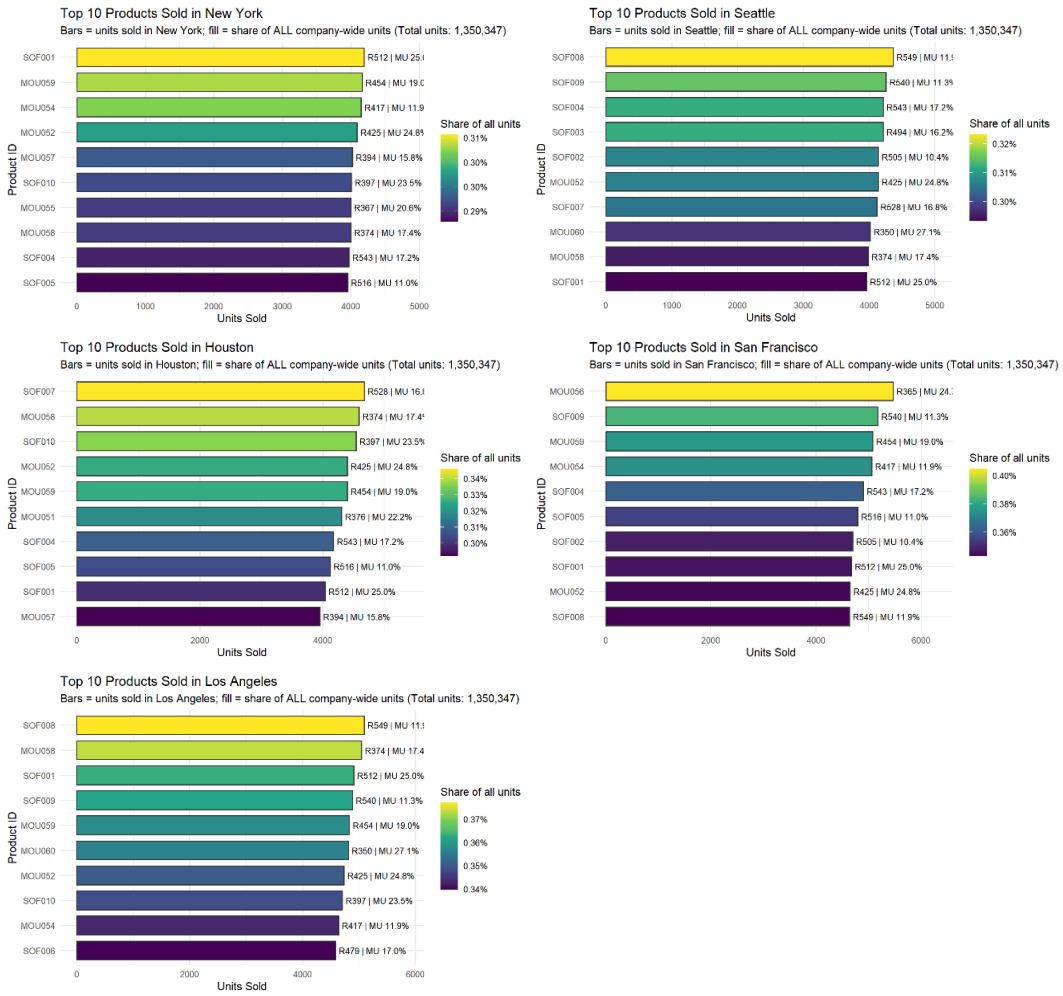




The graph above highlights the significant differences in sales performance across various product types for 2022. Laptops (LAP) are by far the highest revenue-generating product, accounting for more than double the sales of the second-highest category, Monitors (MON). In fact, laptops bring in almost five times the revenue of the remaining product categories combined. This sharp contrast in sales highlights the importance of prioritizing laptop products in the company's operations. Given that laptops are the highest-valued product, the company should focus on ensuring that the supply and tracking systems for laptops are optimized to meet demand. The relatively lower sales of other product types, such as Clothing (CLO), Software (SOF), Keyboards (KEY), and Mouses (MOU), should not be ignored, but these products can be managed with relatively less focus and resources compared to laptops, allowing the company to allocate its efforts more efficiently.

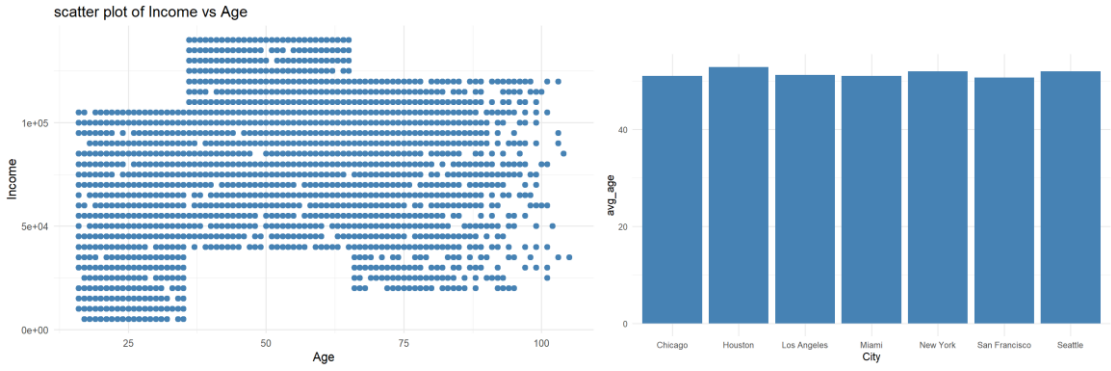






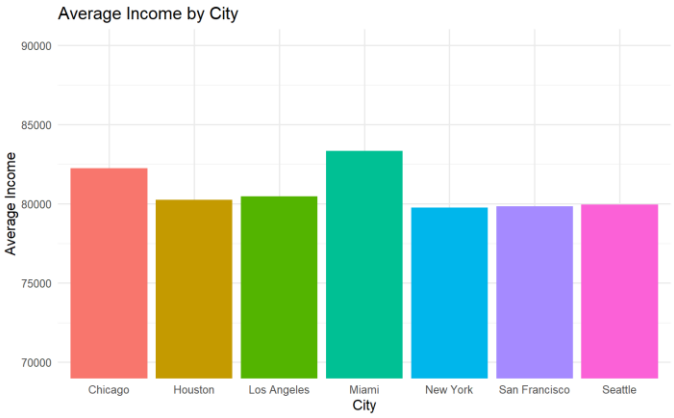
Above is the overall anaylse of what to expect from each city with regards to products sold. It can be interpreted that similar products are sold under the top 10 independent of the city such as SOF (software) and MOU (mouse) products.

Customers



The age analysis above shows a very uniform bar plot distribution therefore does bring important information with regards to what cities age is most important, however for the marketing side of the company the scatterplot above indicates that the highest income comes from the estimated age of 35 to 65, which intuitively makes sense as this is the

age where majority of society have founded a stable job and income for their life where software and technology devices would be the most vital and necessary. This reinforces the idea that this company must highlight marketing and targeting this age range in all locations they are selling their products.



ProductID → Category with sales Miami

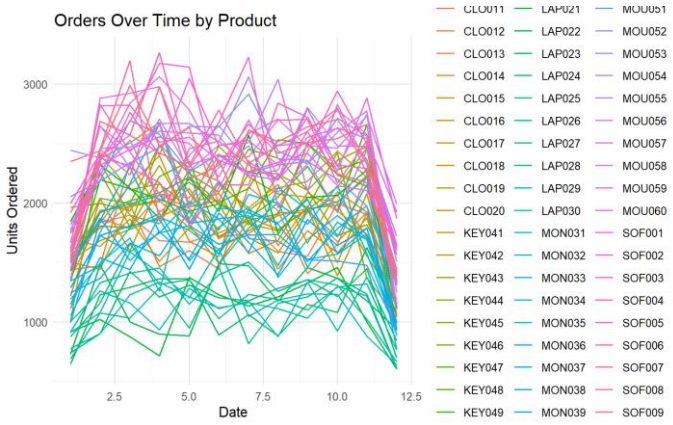
ProductID	Category	Units	Avg price	Avg markup	Share of ALL units
MOU057	Laptop	3942	R394.30	15.8%	0.29%
MOU058	Monitor	3852	R373.82	17.4%	0.29%
SOF007	Software	3841	R527.56	16.8%	0.28%
SOF001	Software	3826	R511.53	25.0%	0.28%
MOU059	Keyboard	3696	R454.04	19.0%	0.27%
MOU053	Keyboard	3689	R424.79	21.4%	0.27%
SOF010	Monitor	3656	R396.72	23.5%	0.27%
KEY041	Keyboard	3631	R530.51	25.6%	0.27%
SOF004	Monitor	3570	R542.56	17.2%	0.26%
MOU056	Cloud Subscription	3564	R364.75	24.7%	0.26%

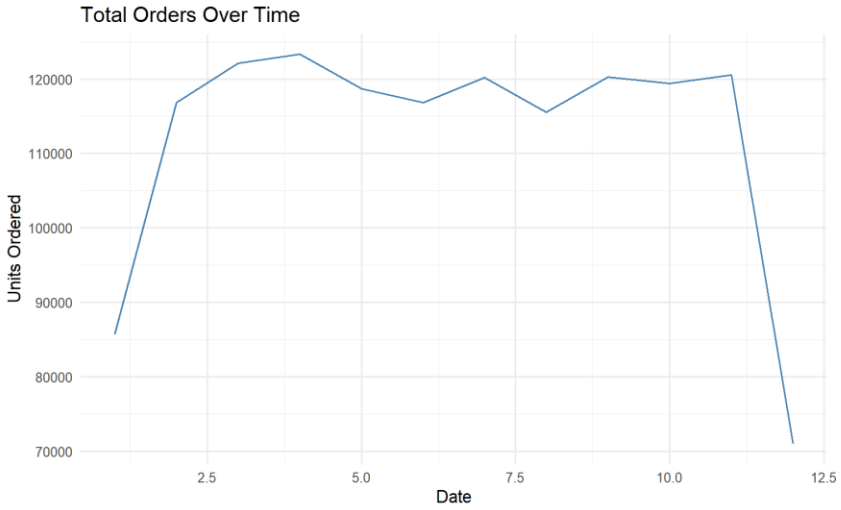
From the uniform distributions to the left, it is interesting to note that Miami and Chicago have the highest income out of all the cities however Miami brings in the least number of customers. This bar plot shows an example of random variation which could be investigated further such as considered variables that could affect a Chicago and Miami to have peaks higher than the rest eg.) average quality of life, salaries etc.

It is also a great example of demonstrating how many factors can play a role for an outcome instead of just 1-1 relationships and therefore taking care of all variable relationships to each other is important for a holistic business analysis.

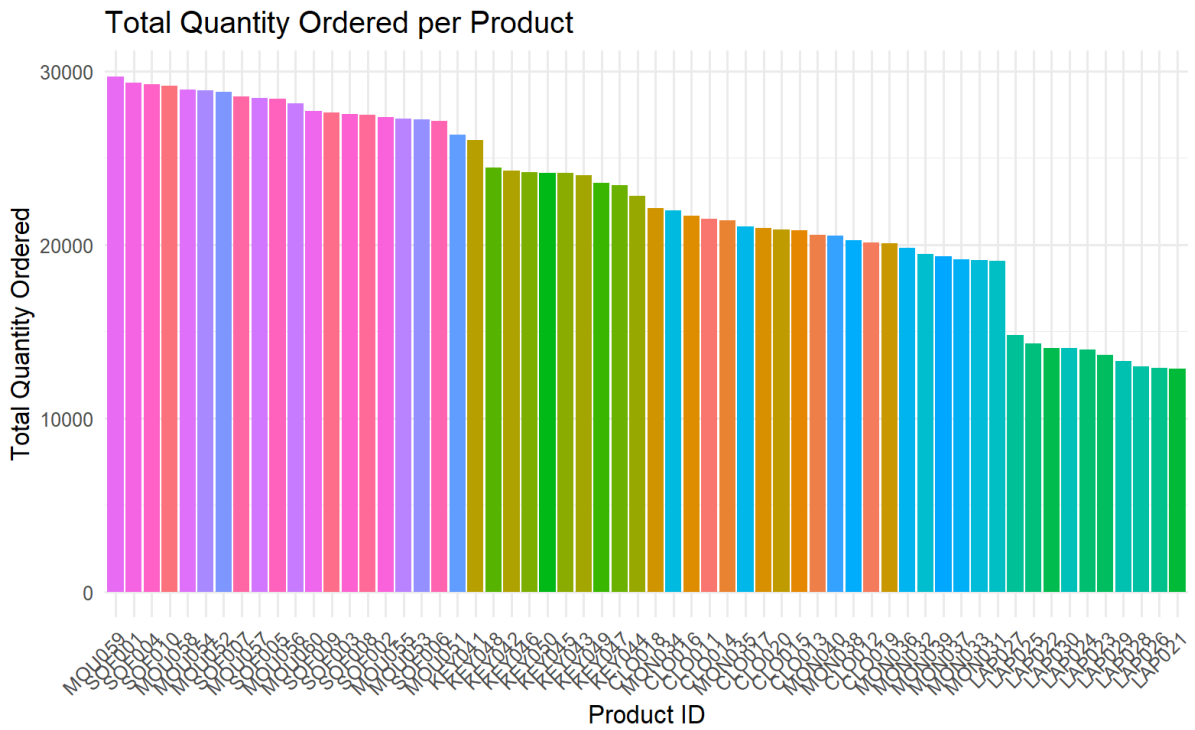
Although there seems to be discrepancies between linking product ID and category, zooming further into Miami's top product performers allows the company to know what to focus on promoting and marketing in Miami, their highest income city with lowest customer engagement. The MOU057 (Laptop) and MOU058 (Monitor) being at the top of this priority list.

Products



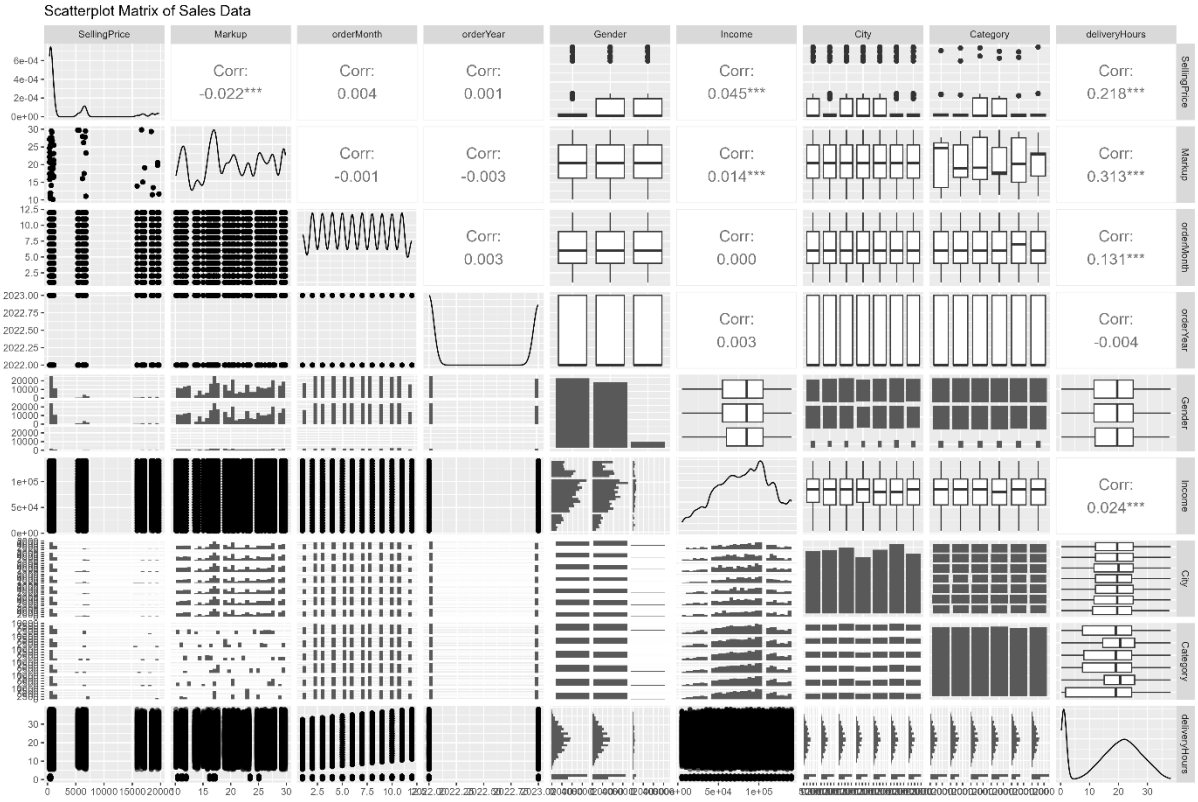


The graphs above correlates well with the monthly sales seen under sales, the dips during Jan and Dec align. This indicates that there is few issues in processing orders to sales, in other words the company is able to meet demand and service level. There is also a clear indication through looking at the variation of colours from the first graph, *Orders over time by product,* distinct separation between quantity and type of product which is useful and is investigated further in the graph below, *total quantity ordered per product:*



The bar chart above displays the total quantity ordered per product, which serves as an essential tool for the company's demand forecasting and inventory management. It helps assess which products need to be ordered in larger quantities from manufacturers and suppliers while also considering their profitability. As observed, although laptops generate more income due to their higher selling price and value price difference, the quantity ordered for Mouse and Software products is higher. This is because these products have greater demand despite their lower price points, which drives up their order volume. Therefore, the company should prioritize inventory levels for Mouse and Software products, as they consistently show higher demand. While laptops contribute significantly to revenue, the relatively lower quantity ordered suggests they may require smaller restocking quantities compared to Mouse and Software. The company should adjust its ordering strategy based on demand volume rather than just profitability, ordering more of the lower-priced, high-demand products and ensuring that inventory levels for these products match their sales velocity. This approach would help meet customer demand efficiently while optimizing profitability.

Insight



The scatterplot matrix above provides a comprehensive view of the relationships between multiple variables in the sales dataset, offering valuable insights into correlations, distributions, and variability.

Significant Correlations: One of the most notable correlations is between Selling Price and Category, with a positive correlation of 0.313, suggesting that different product categories have a notable relationship with their selling prices. Additionally, Delivery Hours and Selling Price show a moderate correlation (0.218), indicating that higher prices are generally associated with longer delivery times. The Income variable also correlates weakly but positively with Selling Price (0.045), suggesting that income could influence purchasing behaviour.

Distributions and Variability: The Selling Price variable shows significant variability, with a wide range of values, indicated by the dispersed distribution in the lower triangle. Similarly, Markup displays a skewed distribution, with several products clustering around certain price points. Other variables like Order Year and City show more discrete, categorical distributions, with Order Year having a clear yearly trend. There is minimal variation in Gender and Category, showing that these variables may not introduce much variability into the data.

Similarities and Differences: The scatterplot matrix highlights how certain variables, like Selling Price and Markup, exhibit similar patterns of distribution, with both having a broad spread. On the other hand, Order Month and Order Year show time-related

trends, with Order Year showing clear yearly cycles, indicating seasonal variations. Interestingly, Category and City variables appear to be relatively constant across the data, with minimal spread, indicating less impact on the outcome.

The scatterplot matrix provides an in-depth view of the relationships, variability, and trends across different product-related variables. The analysis shows that while some variables like Selling Price and Delivery Hours exhibit moderate to strong correlations, others like Gender and City contribute less to the variations in sales patterns. Understanding these relationships is crucial for developing targeted strategies in sales and inventory management, and can inform decisions regarding product pricing, customer demographics, and delivery efficiency.

Overall, this integrated data exploration provides a foundation for deeper business intelligence. By combining transactional, product, and customer data, we are able to evaluate not just what products are being sold, but also to whom, where, and under what conditions. This positions the organisation to make data-driven decisions regarding product strategy, pricing, and regional targeting.

Part 3: Statistical Process Control

3.1 S-charts and X-charts

For the following section the excel file “sales2026and2027futur” were given to help identify when each product manager should go adjust or check their process control using s-charts and x-charts potential issues in service delivery detectors.

Statistical Process Control (SPC) is used to monitor and maintain the stability of a process over time by distinguishing between *natural (common-cause)* variation and *unusual (special-cause)* variation.
In this case, the “process” is the delivery performance of sales orders (e.g., delivery times per product type).

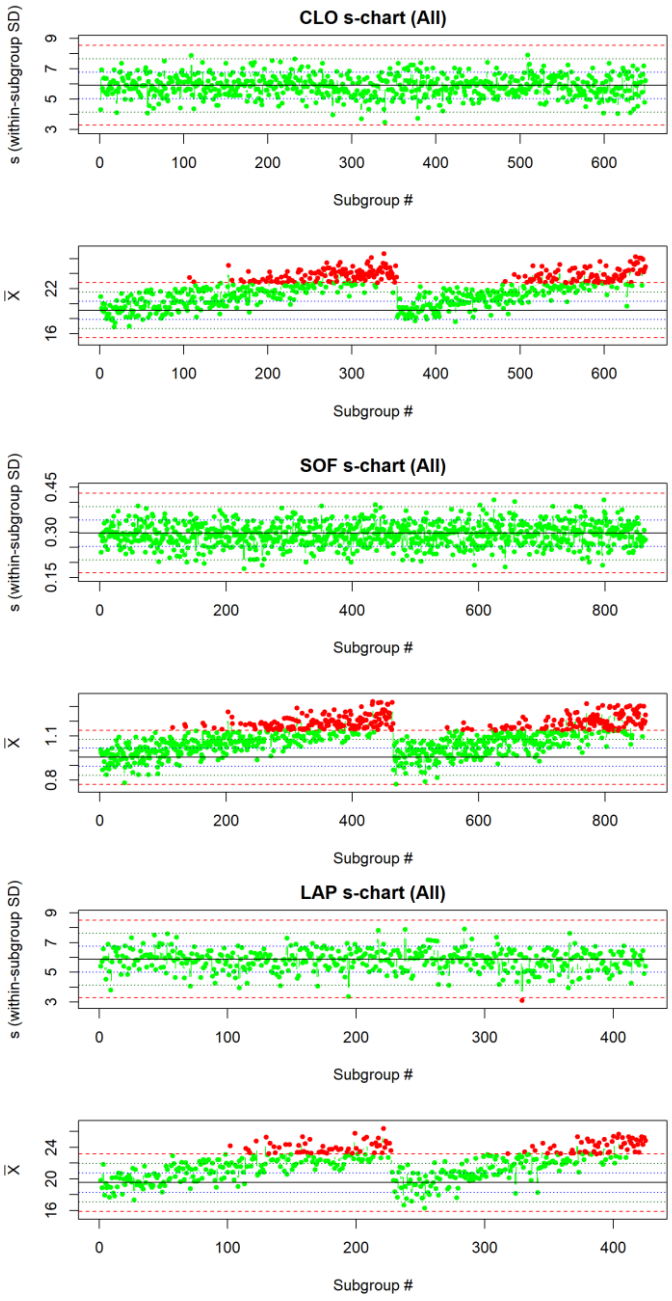
The SPC $\bar{X}$–s chart analysis provided a quantitative method to assess the delivery performance of the sales process over time.
By using subgroup samples of 24 deliveries and the first 30 samples to set statistical control limits, we established a baseline for expected process variation.
The analysis showed that while the variation (s-chart) remained relatively stable, several product types experienced shifts in mean delivery time ($\bar{X}$-chart), indicating special causes affecting delivery performance.
Capability indices (Cp, Cpk) highlighted which product types consistently met the VOC (Voice of the Customer) expectation of under 32 hours.
Overall, SPC enables proactive management , identifying instability early, reducing delays, and ensuring predictable delivery service quality.

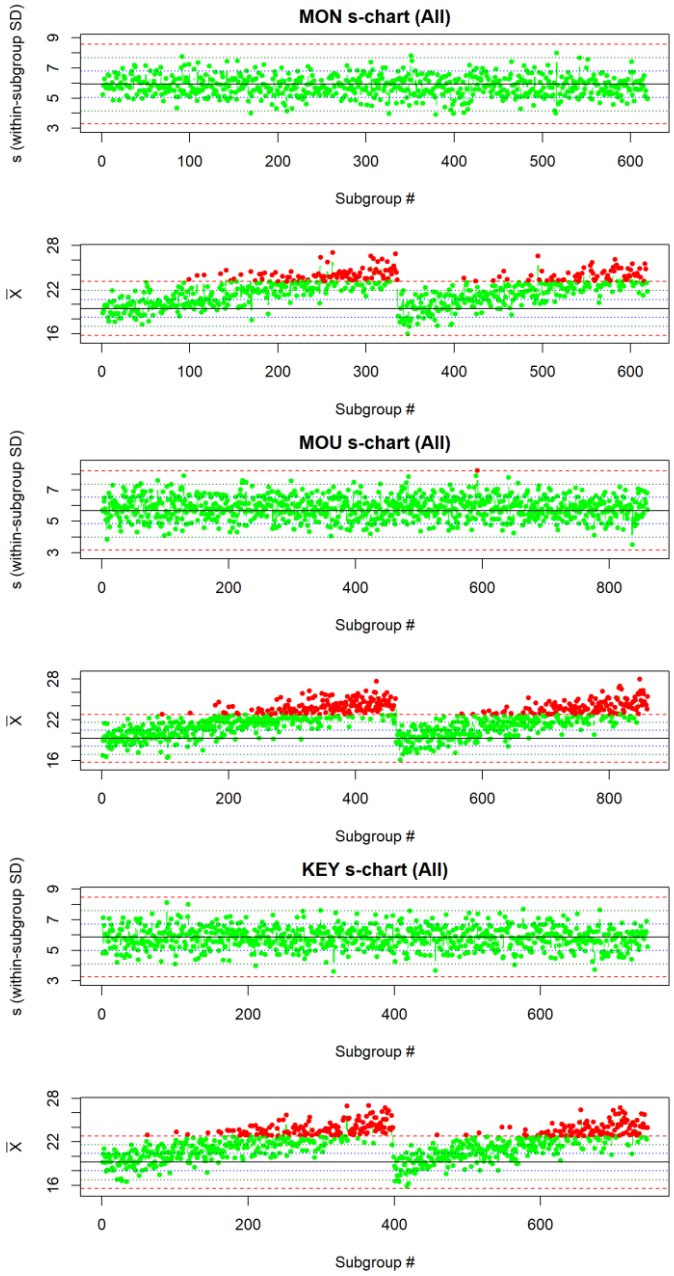
Data preparation

```
   CLO    KEY    LAP    MON    MOU    SOF 
15598  17920  10207  14864  20662  20749 
```

Before applying SPC, code was implemented to verify that each product type had enough observations (≥ 30×24 = 720 deliveries) to initialise control limits reliably. All product types met this requirement, confirming sufficient sampling to represent stable process behaviour before monitoring ongoing deliveries.

Visual representation and discussion





The s-chart monitors variability within each subgroup and is critical for ensuring process consistency before interpreting mean trends.
In most product types, the s-chart indicated stable variation, with points largely contained within the ±3σ limits.
Only a few isolated outliers were detected (Rule A), representing sporadic spikes in variability due to potential factors such as irregular supplier performance, temporary workforce inconsistency, or fluctuating order volumes during high-demand periods.
Because most s-charts were stable, the processes can be considered under control in terms of variation, allowing reliable interpretation of the $\bar{X}$-charts.
Where Rule A triggered, those points should be investigated individually, corrected, and excluded before recalculating updated limits.

The $\bar{X}$-chart tracks average delivery time per subgroup, helping detect systematic shifts in process performance.
All product types showed visible upward shifts across 2026–2027, with multiple subgroups breaching ±2σ or ±3σ limits.
This behaviour aligns with Rule C violations (four or more consecutive points beyond +2σ), suggesting gradual deterioration in average delivery speed.
These shifts may be attributed to seasonal or forecasting inaccuracies, introduction of new delivery routes or distribution centres, or delayed logistics due to product-specific demand surges.

3.3 Calculating Process Capabilities Indices Cp, Cpu, Cpl, Cpk

Capability index	Description	Equation used
Cp (Process Capability Index):	Use when you want to evaluate process potential assuming it is centered	$Cp = \frac{USL - LSL}{6\sigma}$
Cpk (Process Capability Index considering shift)	Use when the process may be off-center, and you want to evaluate the process capability considering the actual location of the mean.	$Cpk = \min\left(\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right)$
Cpu (Upper Capability index	Use when the upper specification limit (USL) is more critical and you want to measure the ability to meet that upper limit.	$Cpu = \frac{USL - \mu}{3\sigma}$
Cpl (Lower Capability Index):	Use when the lower specification limit (LSL) is more critical and you want to measure the ability to meet that lower limit.	$Cpl = \frac{\mu - LSL}{3\sigma}$

ProductType <chr>	Cp <dbl>	Cpu <dbl>	Cpl <dbl>	Cpk <dbl>
MOU	0.9151848	0.7265710	1.103799	0.7265710
KEY	0.9171375	0.7293536	1.104921	0.7293536
SOF	18.1352369	35.1876018	1.082872	1.0828720
CLO	0.8977458	0.7167378	1.078754	0.7167378
LAP	0.8987816	0.6962187	1.101345	0.6962187
MON	0.8890490	0.6995705	1.078528	0.6995705

The data shows that Software (SOF) is the most capable product, with both Cp and Cpk values well above 1, indicating a well-controlled process that meets both the upper and

lower specification limits. In contrast, products like Mouse (MOU), Keyboard (KEY), Cloud Subscription (CLO), Laptop (LAP), and Monitor (MON) show poor process capability, particularly in meeting the upper specification limit (USL), as indicated by low Cpu and Cpk values. The focus should be on improving the upper control limit for these products, as the primary limitation is typically the upper bound. For example, Laptops and Monitors could benefit from tighter control over the USL to improve their Cpu and Cpk scores. Thus, the company should prioritize enhancing the upper control for products with lower Cp and Cpk values, particularly those struggling with the upper limit, such as Mouse, Keyboard, and Laptop.

Linking this to the provided example, where LSL = 0, USL = 32, and capN = 1000, we see that delivery times should ideally never exceed 32 hours. Since delivery time cannot be negative or zero, the LSL of 0 is a theoretical lower bound, ensuring that delivery times are always positive. In practice, however, the focus should be on managing the upper limit—32 hours—to maintain operational efficiency and meet customer expectations. Similarly, just as the upper specification limit is critical in process control for products like Mouse and Laptop, the upper limit of 32 hours for delivery is crucial for ensuring that processes are running efficiently and that exceeding this threshold could signal underlying process issues that negatively impact the company's reputation and profitability.

3.4 Processing control rules

ProductType	Subgroups	RuleA_total_breaches	RuleA_indices	RuleB_longest_run_len	RuleB_longest_run_starts	RuleC_up_runs_ge4_count	RuleC_up_run_starts
MOU	860	1	592	16	672	23	194 235 280 288 301 315 332 346 359 391 632 661 688 702 ...
KEY	746	0		15	730	25	112 172 187 200 227 239 244 254 262 268 275 281 305 570 ...
SOF	864	0		21	659	25	202 237 244 260 278 304 310 319 325 334 343 353 389 649 ...
CLO	649	0		35	474	20	122 179 192 202 219 235 247 266 284 298 316 350 511 522 ...
LAP	425	1	329	19	116	12	119 130 154 170 191 325 331 337 348 361 374 393
MON	619	0		34	238	23	134 179 190 208 215 222 230 240 248 255 284 301 459 484 ...

Rule A: Any s outside ±3σ — Sparse occurrences across few product types; sporadic special-cause variation requiring investigation.
Rule B: Longest run of s within ±1σ — Detected for most product types; indicates strong local stability and low variability.
Rule C: ≥4 consecutive $\bar{X}$ above/below ±2σ —Highest consecutive pattern in KEYBOARD and SOFTWARE and MONITOR; systematic drift likely due to seasonal fluctuation.

Overall

Variation control: Most s-charts are stable, confirming consistent short-term performance.
Mean control: Several $\bar{X}$-charts indicate mean shifts, revealing systematic delays that must be addressed operationally.
Capability: Only certain product types meet customer expectations; others require corrective action to enhance delivery reliability.
Action:

- Investigate and document causes of Rule A violations.
- Stabilize variation before attempting to adjust mean performance.
- For Rule C shifts, review forecasting models, logistics scheduling, and distribution routing.
- Maintain continuous SPC monitoring to verify that corrective measures return the process to control.

Part 4.1and 4.2: Risk

Assume H_0: the process is in control and centred on the baseline (limits set from the first 30 subgroups). Under H_0 the charted statistics follow their in-control distributions, so the probabilities below are theoretical (data-independent):

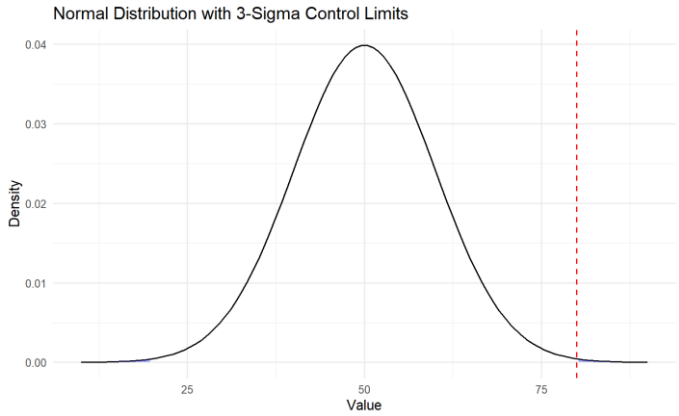
	Decision	
True State of Nature	**Reject H0**	**Do Not Reject H0**
H0 is True	Type I Error	Correct Decision
H0 is False	Correct Decision	Type II Error

Risk correlates to calculating Type 1 and Type 2 errors made throughout a process. This includes understanding whether the business is correctly identifying service issues and how the processes are performing. Overall Calculating Type I and Type II errors is important for understanding the risk of making incorrect decisions based on your statistical test. Type I errors lead to false positives, which can result in unnecessary actions or investments, while Type II errors lead to false negatives, where opportunities for improvement are missed. By carefully balancing these errors and adjusting your significance level and sample size, you can optimize the process to minimize both errors and make better data-driven decisions.

Before continuing it is worth noting, The Central Limit Theorem states that: If you have a large enough sample size, the distribution of sample means will approach a normal distribution, regardless of the shape of the original population distribution. So, even if the underlying data (i.e., the population distribution) is not normal, the distribution of the sample means (i.e., the X-bar values) will approach normality as the sample size increases. This is why the following section will be making use of normal distribution formulas to calculate probabilities as it is assumed the sample size of 24 was large enough to apply CTL, this is also why the central limit theorem was applicable for part 3 previously.

4.1 Type I (Manufacturer's) Error — False-alarm risk under an in-control process

Rule A: Samples outside of the +3 sigma-control limits of all product types from the s-chart.



$$\alpha \approx P(Z > 3) = 0.00135$$

Interpretation of results:

Rule A in SPC refers to monitoring the standard deviation (s) and signal an out-of-control process when s exceeds the upper control limit (UCL), which is typically set at +3σ.

The Type I error in this case is the probability of rejecting the null hypothesis (i.e., concluding the process is out of control) when the process is actually in control. The probability of a Type I error under Rule A is 0.00135 or 0.135%. This means there's a very low chance of observing a sample point beyond the UCL purely by chance, assuming the process is in control.

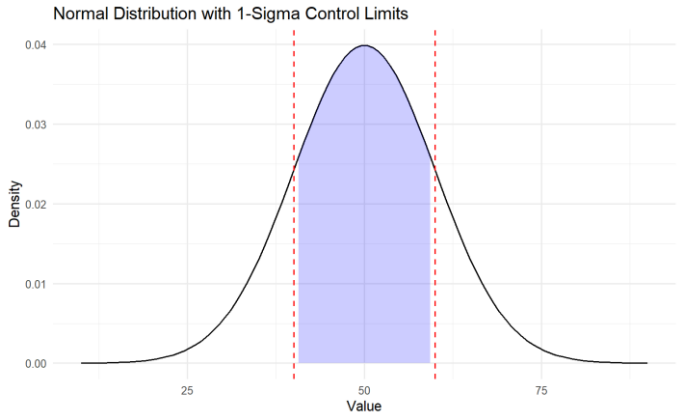
Type I errors result in unnecessary costs, such as:

Wasting time and resources on investigating false alarms.

Potential disruptions in the process due to unnecessary changes being implemented.

Rule A is particularly important when a specific process step contributes heavily to costs and disruptions, especially if this step causes increased variability. By focusing on the UCL (upper control limit), Rule A helps identify larger shifts in variability and signals when corrective actions are needed to maintain process stability.

Rule B (s-chart within ±1σ, longest run).



$$P(|Z| \leq 1) = 0.6827$$

$$\alpha = \mathbf{0.6827}^{k}$$

This is a stability indicator, not a trigger. For reference, under H_0, the chance of k consecutive points within ±1σ is 0.6827^k (e.g., k=10 ⇒ 1.8%). We use it to confirm good control, not to signal out-of-control.

When used for part 3's SPC results the following type 1 errors would be produced per product type:

Table producing Rule B type 1 probabilites using Part 3 results.

Product Type	**Rule B longest run (K)**	**type 1 error probability.**
Mouse	16	0.002227
Keyboard	15	0.003262
Software	21	0.000330
Cloud subscription	35	0.000001
Laptop	19	0.000709
Monitor	34	0.000002

Interpretation of results:

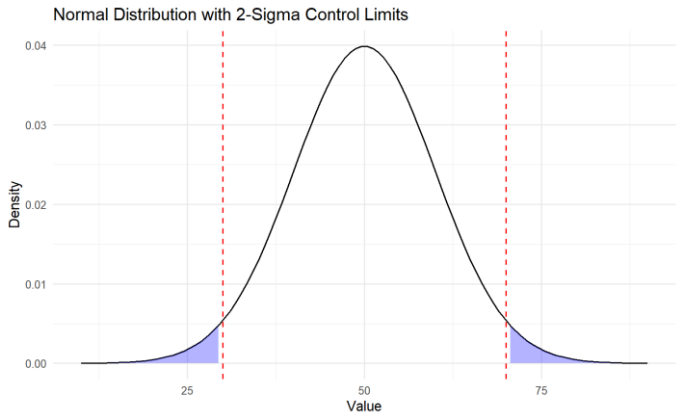
Rule B refers to the longest run rule, used in SPC to monitor whether the process is in control. As the number of consecutive points (k) increases, the probability of Type I error decreases exponentially:

$$\alpha = 0.6827^{k}$$

This means that long consecutive runs (e.g., 20 points) become rare events under the assumption of a stable process, but they may indicate a larger shift in the process, which could signal a problem. Long runs reduce the likelihood of a false alarm (Type I error), but they also increase the likelihood of identifying out-of-control conditions.

Therefore, while longer runs reduce Type I errors, they indicate larger deviations from the process mean, which could signal actual problems with the process that need to be investigated. As shown in the table for Rule B, long consecutive runs are uncommon in a stable process, and when they occur, the Type I error probability decreases, making it less likely to have a false alarm, but more likely to identify a real problem.

Rule C ($\bar{X}$-chart, ≥4 consecutive beyond the ±2σ line).



$$(|Z| > 2): P = 0.0455$$

$\alpha \approx (0.0455)^4 = \mathbf{4.3 \times 10^{-6}}$ per run of 4.

Interpretation of results:

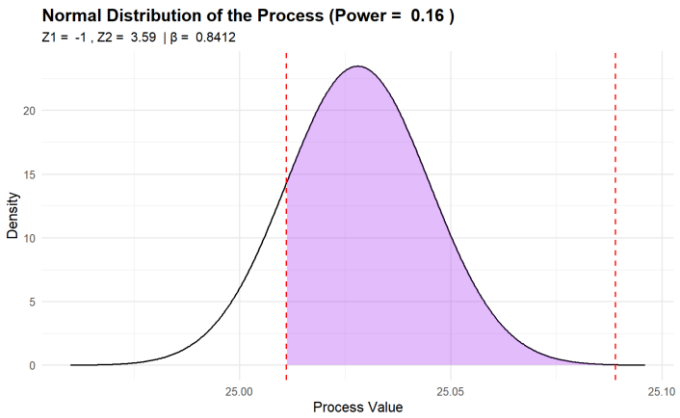
Rule C refers to a monitoring rule in SPC, where we signal an out-of-control process when there are 4 consecutive points beyond the ±2σ control limits on an X-bar chart. The probability of Type I error (false alarm) for this rule is very small: 0.0455^4

This means that the likelihood of seeing 4 consecutive points beyond $\pm 2\sigma$ due to random chance is extremely low, making it a strong signal for process instability.

As the number of consecutive points increases (e.g., 5 or 6 points beyond the control limits), the probability of Type I error decreases exponentially, reducing the chance of a false alarm. Therefore, long runs of points beyond the control limits require strong evidence before triggering corrective actions, minimizing unnecessary disruptions and resource expenditures.

4.2 Type II (Consumer's) Error — Missed detection for the bottle-filling example

Tasked with calculating the Type II error (β) : the probability that we fail to reject $H_0$when $H_a$is actually true (i.e., the process has shifted to 25.028 liters). This problem involves hypothesis testing with a shift in the process mean and a change in the standard deviation.



Type II error β = P(LCL ≤ $\bar{X}$ ≤ UCL | μ=25.028, σ=0.017).

Null Hypothesis H_0: The process is in control and centered at 25.05 liters (this is the baseline value).

Alternative Hypothesis H_a: The process has shifted to an average of 25.028 liters (the new process mean).

The control limits for the X-bar chart are:

Upper Control Limit (UCL) = 25.089 liters

Lower Control Limit (LCL) = 25.011 liters

The true process mean has shifted to 25.028 liters, and the standard deviation has increased from 0.013 liters to 0.017 liters.

calculate Z-scores for LCL and UCL with the new process parameters.

The Z-score is calculated as:

$$Z = \frac{\text{Value} - \mu}{\sigma_{\bar{X}}}$$

LCL Z-score:

$$Z_{\text{LCL}} = \frac{25.011 - 25.028}{0.017} = \frac{-0.017}{0.017} = -1.00$$

UCL Z-score:

$$Z_{\text{UCL}} = \frac{25.089 - 25.028}{0.017} = \frac{0.061}{0.017} = 3.59$$

Step 2: Calculate the Probability of Type II Error (β)

To calculate the Type II error probability (β), we need to find the probability that the sample mean $\bar{X}$falls between the LCL and UCL when the true process mean is 25.028 liters.

The probability of Type II error is the area under the normal curve between these two Z-scores, using the cumulative distribution function (CDF) for the standard normal distribution.

The probability of failing to reject H_0(i.e., Type II error) is:

$$\beta = P(Z_{\text{LCL}} \leq Z \leq Z_{\text{UCL}}) = \Phi(3.59) - \Phi(-1.00)$$

$\Phi(3.59)$= 0.9998 (the probability that $Z$is less than 3.59),

$\Phi(-1.00)$= 0.1587 (the probability that $Z$is less than -1.00).

$$\beta = 0.9998 - 0.1587 = 0.8411$$

This means the probability of making a Type II error is approximately 84.1%.

Power is the complement of Type II error:

$$\text{Power} = 1 - \beta = 1 - 0.8411 = 0.1589$$

Part 4.3: Data correction

The following data was corrected from Part 1 between Products and Products_headoffice.

Process

Objective: produce corrected files products_Headoffice2025.csv and products_data2025.csv that align IDs, categories, and price/markup patterns.

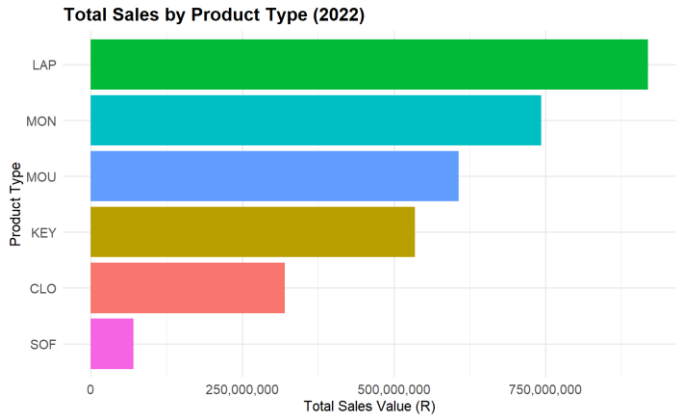
ProductID prefix fixes: Any erroneous/“NA” prefixes were replaced by the first three characters of ProductID (e.g., SOF, KEY, MOU), ensuring consistent type coding.

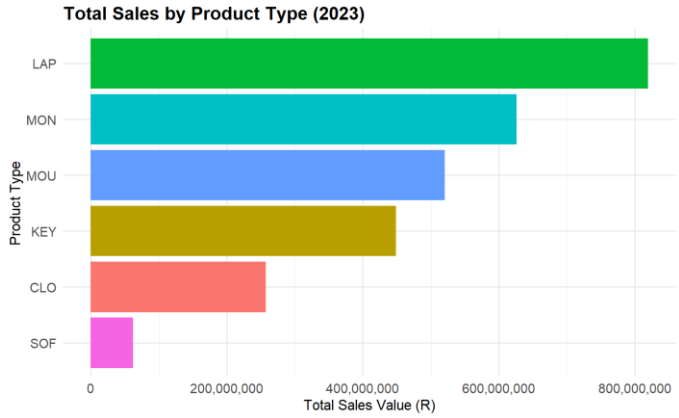
Repeating price/markup for items 11–60: For each type, entries 1–10 are canonical. Items 11–60 are corrected by repeating the price and markup of items 1–10 in 10-item cycles (e.g., 11→1, 12→2, …, 20→10, 21→1, etc.), matching the brand/model pattern.

Category alignment in products_data2025.csv: The Category column was updated to correspond deterministically to the ProductID prefix (e.g., SOF→Software, KEY→Keyboard, etc.), removing mismatches and enabling consistent aggregation and margin analysis.

Interpretation- Specifically refer to Part 1.2: Potential errors in data analysis due to data and sales

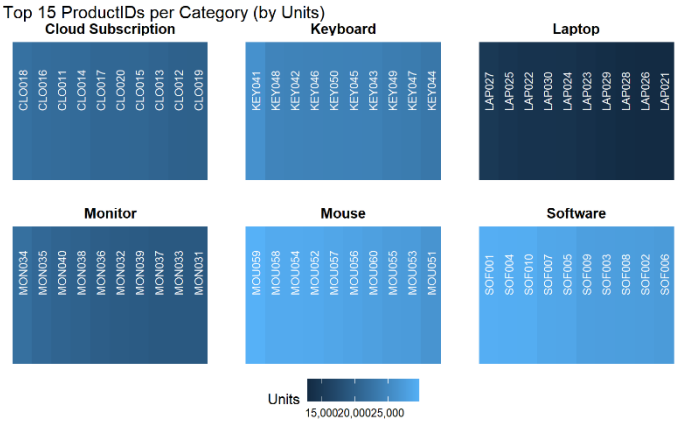
Sales



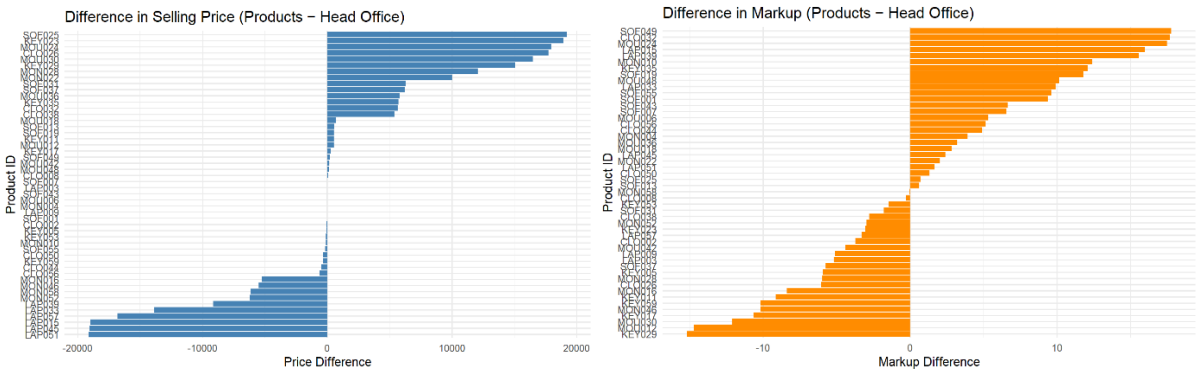


The graphs to the left are much more evenly distributed, although still extremely skewed towards laptops. **Laptops (LAP)** still dominate the sales, with a **slightly lower relative contribution** than in the previous graph. Between 2022 and 2023 it seems as though same products remained constant in ranking however the corrected data now indicates that mouse products bring a lot more sales than previously analysed whereas software brings in the least number of sales unlike previously understood by misaligned data in part 1.2.

Products



As can be seen on the graphs to the left, there seems to be errors between correlating Product ID's to the correct categories have been altered to represent true. This allows a true reflection of units per category sold to be interpreted. For example, now it is seen than laptops bring in the least amount of units but still obtain the most amount of profit for the company. Therefore, viewed as very high valued products.



If rearranged the graph above would represent a normally distributed bar plot. The difference in selling price and markup for product data file and head office data file is the most interesting to note as more discrepancies were found with the corrected data than the incorrect data. This is mainly because the correct data set has increased tremendously in size due to naming the NA files which were outright ignored before as well as filling the head office data set with many more product instances. However, this does indicate that further investigation and corrections between these two data sets should be prioritized for the company.

Insight

It is crucial to correct type codes, prices, markups, and categories to avoid margin leakage and misallocation in dashboards, ensure capability/variance analysis is by the right groupings, and support reliable profit optimisation in later steps, as can be seen in the corrections made above, large changes in sales and product interpretations were corrected, therefore supporting the fact that it can cost a business plenty if not correctly stored and analysed.

Part 5: Optimization (Barista shop)

The analysis explores the balance between staffing costs and the number of customers served, ultimately aiming to maximize profitability. To optimize performance versus efficiency in the coffee shops while staying competitive, the focus should be on balancing speed, customer service quality, and operational cost management.

Makes use of two excel files:

shop 1: timetoserve.csv

Shop 2: timetoserve2.csv

Both these files include number of baristas working and the average time it took for each barista to serve a customer that day.

The following assumptions were made to produce the expected outcome:

Baristas work 8 hours shift per day

There is a proportional relationship between the time taken to serve customer and number of customers per day. Eg. Assume that shorter service times automatically mean more customers to serve.

Unreliable service occurs when barista takes more than **300 seconds** to serve a customer.

$$Profit\ per\ Day = (Revenue\ per\ Customer \times Total\ Customers\ per\ Day) - (Staff\ Cost\ per\ Day \times Number\ of\ Baristas)$$

$$Profit\ per\ Day = (R30 \times Total\ Customers\ per\ Day) - (R1000 \times \#\ of\ Baristas)$$

Given information: (important to note in creating optimal model)

Experience problems if there is less than 2 baristas working and have a maximum capacity of 6 baristas.

R30 profit per customer served.

R1000 cost per barista per day.

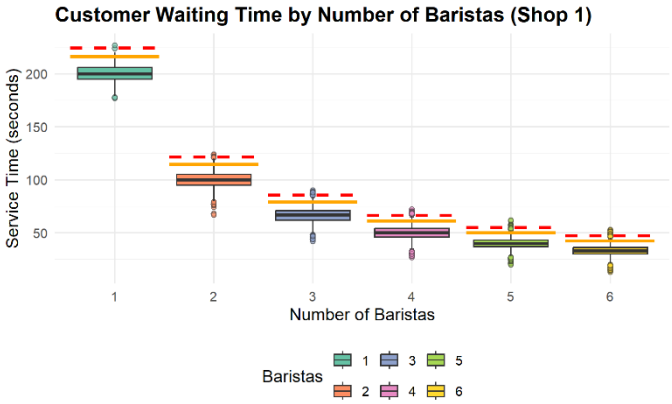
To find total customers per day:

$$customers\ per\ day = \frac{8 \times 60 \times 60(\text{sec})}{avg\ service\ time(sec)}$$

To find average service time:

$$avg\ service\ time(sec) = \frac{sum\ of\ all\ service\ times\ for\ \#\ of\ baristas\ working}{total\ amount\ of\ instance\ \#\ baristas\ were\ working}$$

Shop 1

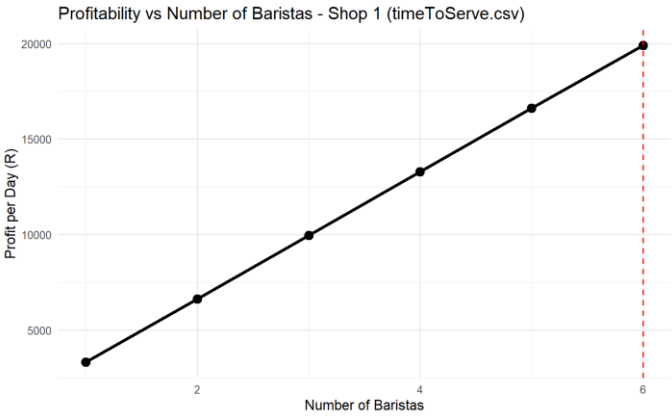


The graph on the left helps visualize the distribution of service time plotted on a box and whisker plot for each number of baristas with the yellow and red lines representing the +2-sigma and +3-sigma SLA levels respectively.

Summary of service time and barista information

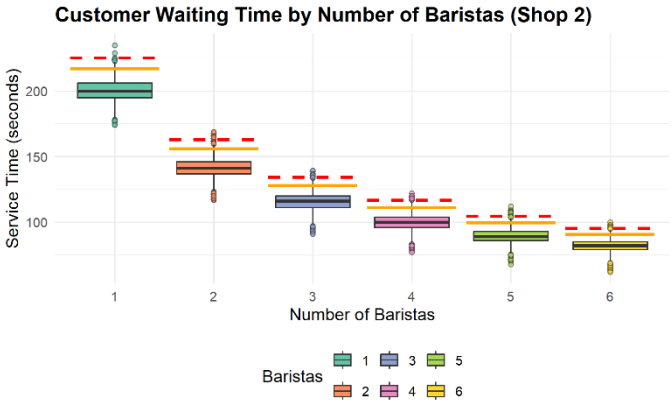
Baristas <int>	avg_service_time <dbl>	reliable_service_percent <dbl>
1	200.15588	100
2	100.17098	100
3	66.61174	100
4	49.98038	100
5	39.96183	100
6	33.35565	100

rows



The outcome of optimizing profit for shop 1 follows a proportional relationship, as number of baristas increase so does profit proportionally. This due to the fact that the mean service time is inversely proportional to the number of baristas; $\frac{200}{number\ of\ baristas}$. Optimal number of baristas was found to be ***6*** with a avg service time of ***33.35565.***

Shop 2

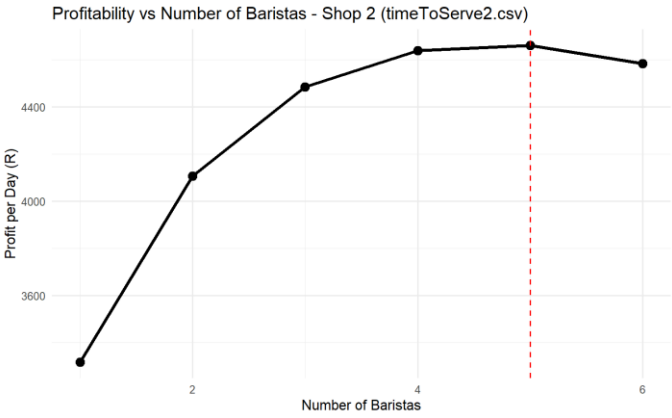


The graph on the left helps visualize the distribution of service time plotted on a box and whisker plot for each number of baristas with the yellow and red lines representing the +2-sigma and +3-sigma SLA levels respectively.

Summary of service time and barista information

Baristas <int>	avg_service_time <dbl>	reliable_service_percent <dbl>
1	200.16894	100
2	141.51462	100
3	115.44091	100
4	100.01527	100
5	89.43597	100
6	81.64272	100

rows



The outcome of optimizing profit for shop 2 shows a concave relationship. Optimal number of baristas was found to be ***5*** with a avg service time of ***89.43597***. This is indicative of diminishing returns: as more baristas are added, the incremental increase in profit becomes smaller. This is due to the fact that the avg service time decreases less as more baristas are added.

Discussion of results

Taguchi loss:

$$L(y) = K \times (y - T)2$$

L(y) is the loss (in profit or efficiency).

y is the actual output (e.g., the number of baristas or service time).

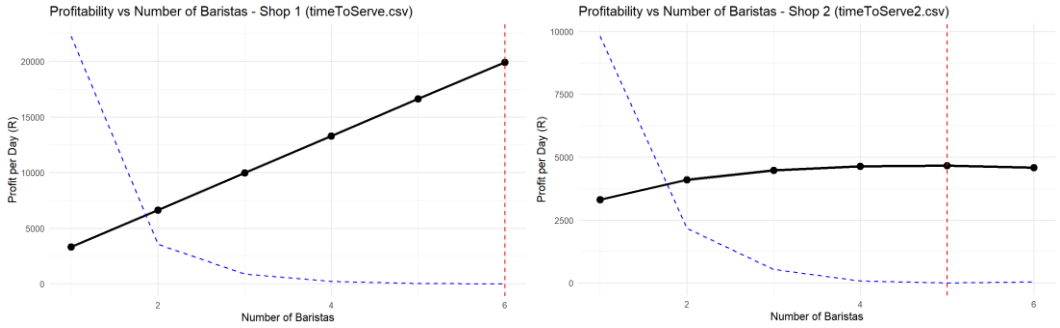
T is the target value (the optimal number of baristas or service time).

K is a constant that scales the loss value.

For ***Shop1*** the target value, ***T***, was chosen as the optimal service time found, ***33.35565*** (6 baristas) and the ***K*** constant of 0.8 was chosen to fit the graph respectively to profit gained.

As can been see from the graph below where the blue dashed line represents Taguchi loss, as the deviance from optimal target value increase the Taguchi loss increases quadratically. The squared term $(y-T)^2$means that the loss grows faster as the actual value (y) deviates from the target (T). This quadratic behaviour makes sense in real-world contexts, where minor inefficiencies might not matter much, but large inefficiencies (e.g., high service time) cause significant losses in profit and service reliability. It is also It's important to note that the value of Kdoesn't just "fit the graph"; it reflects the magnitude of loss that results from deviations from the target.

The same can be said for ***Shop 2*** where target value, ***T,*** was chosen as the optimal service time found***, 89.43597*** (5 baristas) and the ***K*** value of 0.8 was chosen to fit the graph respectively to profit gained.



Shop 2 demonstrates superior process control and lower variability, possibly due to better workflow design or staff training. Both visual and numerical findings confirmed that service time decreases rapidly with added baristas up to a certain point, beyond which the benefit flattens, an essential insight for the forthcoming profitability optimisation model.

Part 6.1 and 6.2: Design of Experiment (ANOVA)

To investigate the design of experiment the following two analysis of variance were conducted:

Type of ANOVA	Description	Equation used
One-Way ANOVA	to check if there's a significant difference in delivery time across years for different products.	$\text{Delivery Time} = \mu + \text{Year} + \epsilon$
Two-Way ANOVA	investigated the combined effect of year and month on delivery time (including the interaction between	$\text{Delivery Time} = \mu + \text{Year} + \text{Month} + (\text{Year} \times \text{Month}) + \epsilon$

	order year and order month.	

Where:

$\boldsymbol{\mu}$ is the overall mean delivery time,

Year is the factor for the year

$\boldsymbol{\epsilon}$ is the residual error

Year and Month are the main effects

Year × Month is the interaction effect

This helps reveal how these factors interact and their impact on delivery times and overall business performance.

Factors of great importance that must be considered before performing ANOVA are to first check for interaction between the two or more variables analysed, if there is interaction then cannot perform ANOVA as it becomes unclear what is influencing the results.

ANOVA assumes that the sample size across groups (e.g., years or months) is roughly equal and that the variances across these groups are homogeneous. If the sample sizes are unequal, or if the variance is not homogeneous, the results of ANOVA may not be reliable. In such cases, adjustments like filtering the data or using alternative tests. As can be seen below

A tibble: 2 × 4

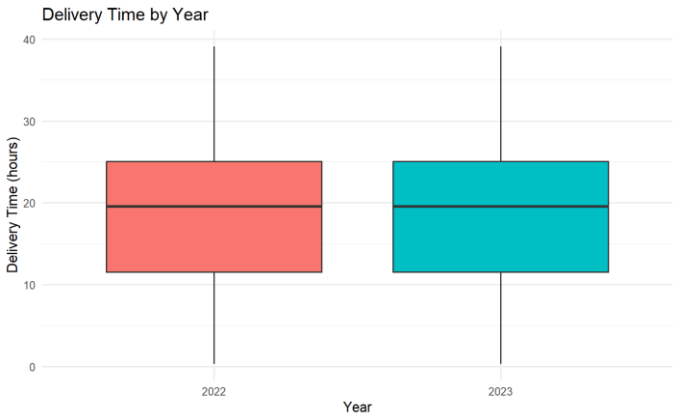
order_year <fctr>	n <int>	mean_pick <dbl>	mean_del <dbl>
2022	9553	13.68465	21.86467
2023	8367	13.72249	21.60562

2 rows

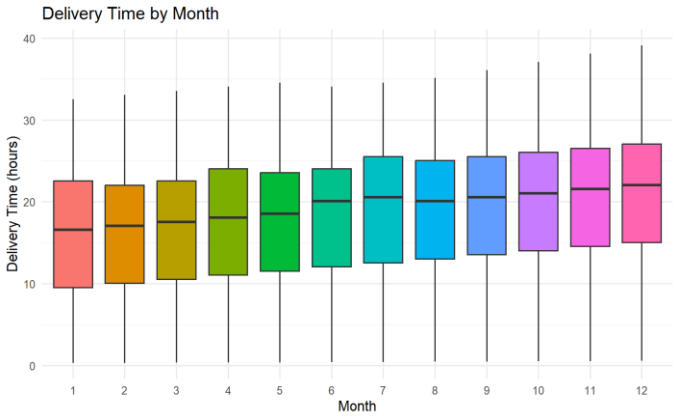
2022 does indeed make use of more samples than 2023 and therefore discrepancies in the results may be due to this fact.

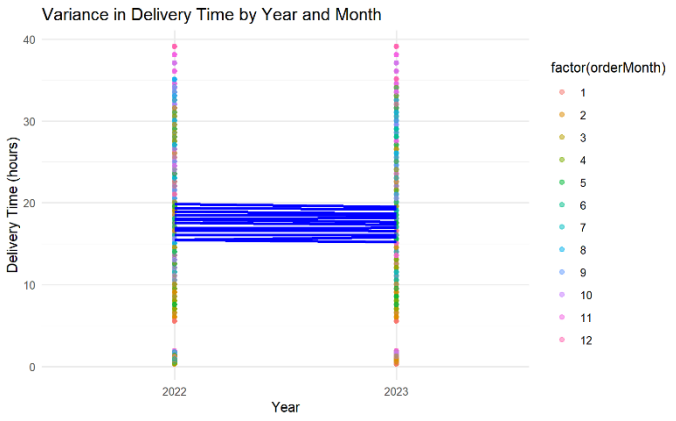
Statistical Significance was assumed that if p_value < 0.05 for a factor (e.g., year or month), we conclude that the factor significantly affects the delivery time.

One-way ANOVA: factor was 2 (the two years)



Two-way_ANOVA: factor was 12 (every month of the year)





One-way-ANOVA

Summary of Results and Findings:

The lack of interaction means the results are reliable for the factors analyzed, and we can confidently conclude that month affects delivery time, without being influenced by the year.

One-Way ANOVA:

Df	Sum_Sq	Mean_Sq	F_value	Pr_F	Significance
1	138.6173	138.6173	1.390657	0.238297	Not Significant

The One-Way ANOVA tested whether there was a significant difference in delivery time between 2022 and 2023. The analysis showed that there is no significant difference in delivery time across these years, as evidenced by the p-value of 0.234, which is greater than the typical significance threshold of 0.05. This finding is consistent with the boxplot showing that the distribution of delivery times for both years is very similar. The boxplot of delivery time by year confirms the ANOVA result, showing similar distributions of delivery times for 2022 and 2023, with no significant difference between the years.

Two-Way ANOVA:

Df	Sum_Sq	Mean_Sq	F_value	Pr_F	Significance
NA	*NA*	*NA*	*NA*	*NA*	*NA*
1	169773.7486	169773.7486	1732.734039	0.0000000	***
1	150.6433	150.6433	1.537486	0.2149953	Not Significant

The Two-Way ANOVA examined the effect of year and month on delivery time, along with their interaction. The results indicated that year has a significant effect on delivery time (p-value = 0.0000000), while month and the interaction between year and month were found to be not significant (p-value = 0.21499953). This suggests that while delivery time varies between 2022 and 2023, the effect of month on delivery time does not significantly change between the two years.

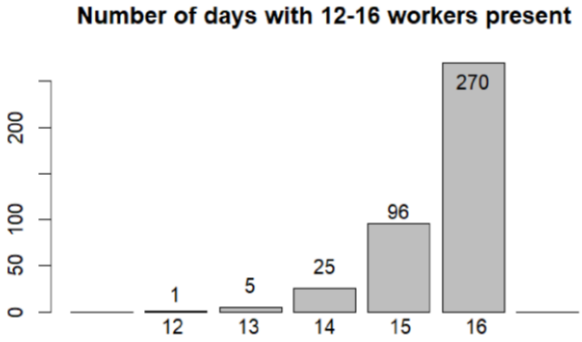
The boxplot of delivery time by month reveals noticeable variation across the months, supporting the significant effect of month observed in the ANOVA. The variance plot further supports the absence of a significant interaction between year and month, showing that the pattern of delivery times across months remains similar for both years.

Overall, the statistical analysis and graphical representations highlight that month significantly affects delivery time, while year does not, and there is no significant interaction between year and month. These findings suggest that the primary driver of delivery time variability in the dataset is month, rather than the year of the order.

Part 7: Optimization (Reliability of Service)

The following analysis incorporates a model for service reliability, estimating the percentage of reliable service days and understanding how staffing levels affect profitability.

The assumption was made that reliable service is considered to occur when there are 15 or more workers present.



7.1 Calculating the p value to estimate the number of reliable days per year for car service

$$\#\ of\ reliable\ days\ per\ year = \frac{\#\ of\ days\ equal\ or\ above\ 15}{total\ number\ of\ days} \times \#\ of\ days\ in\ the\ year$$

$$\#\ of\ reliable\ days\ per\ year = \frac{(96 + 270)}{397} \times 365$$

$$\#\ of\ reliable\ days\ per\ year = 336.5 \approx \mathbf{337\ days}$$

7.2 Optimizing profit for the company

$$Minimize\ costs\ per\ day =$$

$$\#of\ personnel \times cost\ of\ hiring\ \#\ personnel\ perd\ day\ +$$
$$cost\ of\ unreliable\ service\ on\ sales \times\ \%\ service\ not\ reliable\ for\ \#\ of\ personnel$$

$$= \#of\ personnel \times \frac{25000 \times 12}{365} + 20000 \times\ \%\ service\ not\ reliable\ for\ \#\ of\ personnel$$

Information given:

Experience problems if less than **15 people** on duty.

Every day we experience problems reduces profit by **R20 000**.

Appointing a personnel cost **R25 000** per month per person.

To find avg weighted % that service is reliable:

P12: $1/397 = \binom{16}{12}x^{12}(1-x)^{16-12}$

P13: $5/397 = \binom{16}{13}x^{13}(1-x)^{16-13}$

P14: $25/397 = \binom{16}{14}x^{14}(1-x)^{16-14}$

P15: $96/397 = \binom{16}{15}x^{15}(1-x)^{16-15}$

P16: $270/397 = \binom{16}{16}x^{16}(1-x)^{16-16}$

$$avg\ weighted\ \%\ that\ service\ is\ reliable = \frac{1(p12)+5(p13)+25(p14)+96(p15)+270(p16)}{397}$$

$$avg\ weighted\ \%\ that\ service\ is\ reliable = \mathbf{0.974}$$

The following assumption was made t effectively use the binomial distribution function:

The % that is not reliable for # of personnel below 15 is automatically 100%, as it costs the full 20000 for unreliable service on sales revenue, and does not follow the usual binomial distribution function which is used for 15-20 in the following manner:

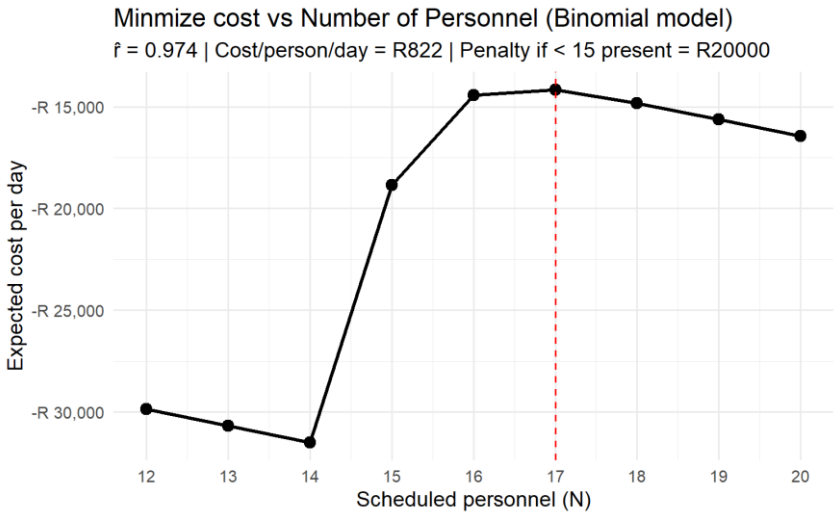
$$Binomial\ distribution\ equation = \binom{n}{x}p^x(1-p)^{n-x}$$

N*= *20 *personnel. The threshold that this equation will be based off is that the max amount of employees to hire is 20.*

***P*=** $avg\ weighted\ \%\ that\ service\ is\ not\ reliable = 1 - 0.974 = \mathbf{0.026}$

X= *range between 15-20.*

N <int>	P(shortage) <dbl>	Reliability <chr>	Staff cost/day <dbl>	Penalty/day <dbl>	Total cost/day <dbl>	Profit/day (rel) <dbl>	Reliable days/yr <dbl>
12	1.0000	0.0%	9863	20000	29863	-29863	0.0
13	1.0000	0.0%	10685	20000	30685	-30685	0.0
14	1.0000	0.0%	11507	20000	31507	-31507	0.0
15	0.3262	67.4%	12329	6524	18852	-18852	245.9
16	0.0636	93.6%	13151	1273	14423	-14423	341.8
17	0.0091	99.1%	13973	181	14154	-14154	361.7
18	0.0010	99.9%	14795	21	14815	-14815	364.6
19	0.0001	100.0%	15616	2	15618	-15618	365.0
20	0.0000	100.0%	16438	0	16439	-16439	365.0



Discussion of results

As can be seen from the results anything below 15 personnel will automatically cost the company the full 20000 however equal to or above 15 will slowly improve reliability and therefore service and sales for the company. The trade-off to be made here is between cost of an extra employee per month and cost of unreliable service. The graph shows a peak minimum cost at 12 personnel count around R14750.

Conclusion

In conclusion, this comprehensive analysis underscores the critical role of data-driven decision-making in elevating Quality Assurance standards within Industrial Engineering operations. By systematically applying advanced statistical and optimization techniques to key operational data, this report has successfully translated raw data into actionable strategic insights.

The descriptive data analysis successfully established a baseline understanding of operational performance, paving the way for the rigorous evaluation of process stability. The application of Statistical Process Control (SPC) and the calculation of Process Capability Indices provided quantifiable metrics demonstrating whether processes are operating within the specified control limits, thus clearly identifying areas requiring immediate improvement.

Furthermore, the investigation into hypothesis testing (with a focus on Type I and Type II errors) quantified the reliability and risk inherent in current business processes, offering a framework for robust quality management. Crucially, the deployment of ANOVA and Two-Way ANOVA isolated and confirmed the statistically significant factors influencing service delivery and operational outcomes. The subsequent profit optimization scenarios, built upon these statistical findings, delivered clear pathways to improved financial performance.

Ultimately, this report fulfils its objective by providing a holistic, data-backed methodology to enhance operational quality, efficiency, and profitability, aligning fully with the high-level ECSA outcomes required for competent Industrial Engineering practice.

References

ChatGPT. (2016). *ChatGPT*. [online] Available at: https://chatgpt.com/g/g-p-68a8a31d05908191bc4d921c000f7ece-laras-chats/c/68fb5caa-b88c-832a-ac81-c74022c062e6 [Accessed 24 Oct. 2025].

Gemini. (2024). *Gemini*. [online] Available at: https://gemini.google.com/app/61f2e8c880df8a2a [Accessed 24 Oct. 2025].

Illowsky, B. and Dean, S. (2022). *Introductory Statistics*. OpenStax.

Sthda.com. (2025). *MANOVA Test in R: Multivariate Analysis of Variance - Easy Guides - Wiki - STHDA*. [online] Available at: https://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance.

Sun.ac.za. (2025). *Log in to the site | stemlearn*. [online] Available at: https://stemlearn.sun.ac.za/pluginfile.php/65516/mod_resource/content/5/QA344%20Statistics.pdf [Accessed 24 Oct. 2025].