Quality Assurance 344

ECSA Final Report 2025

MB Pheiffer, 23772220

Stellenbosch

UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Plagiarism Declaration

I, MB Pheiffer (23772220), declare that:

- I have read and understand the Stellenbosch University Policy on Plagiarism and the definitions of plagiarism and self-plagiarism contained in the Policy [Plagiarism: The use of the ideas or material of others without acknowledgment, or the re-use of one’s own previously evaluated or published material without acknowledgment or indication thereof (self-plagiarism or text recycling)].
- I also understand that direct translations are plagiarism, unless accompanied by an appropriate acknowledgment of the source. I also know that verbatim copy that has not been explicitly indicated as such is plagiarism.
- I know that plagiarism is a punishable offense and may be referred to the University's Central Disciplinary Committee (CDC), which has the authority to expel me for such an offense.
- I know that plagiarism is harmful for the academic environment and that it has a negative impact on any profession.
- Accordingly, all quotations and contributions from any source whatsoever (including the internet) have been cited fully (acknowledged); further, all verbatim copies have been expressly indicated as such (e.g., through quotation marks) and the sources are cited fully.
- Except where a source has been cited, the work contained in this assignment is my own work and that I have not previously (in its entirety or in part) submitted it for grading in this module/assignment or another module/assignment.

	26/10/2025
Signed	Date

AI Use Declaration

I am convinced and can support my claim that my assessment product is an indication of my own learning, knowledge, skills, and understanding.

Where I have used AI tools for enhancing my own creation of ideas and words, I acknowledge that I have to declare it.

Where I have used AI tools for generating new ideas, words, code, image-prompts for other AI Image-generating tools, or structure and even presentations (or other AI tools that can be used as assistants to the knowledge building and representing process), I have declared and documented the use of such tools and I am pre- pared to talk about the process I used and what it contributed to my learning and insights.

I am aware that the lecturer can ask me to demonstrate my learning, for example, through explaining the choices I made in terms of approach, content used, literature selected, conclusions drawn, etc., through an additional assessment like an oral (for example).

I understand that if I am not able to agree to the above points, there is a chance that my academic behavior will be deemed unethical and might lead to a disciplinary case being brought against me on the grounds of cheating or plagiarism, and that the standard procedures for such behavior will be followed.

As per the Disciplinary Code of SU (par. 10.2.1 and 10.2.2), I understand that I take responsibility for the integrity of my work, which includes the obligation to ask for clarification from an academic member of staff if I am unsure of anything, and that I strictly adhered to all instructions received in the course of the academic assessment by relevant and authorized staff (whether the instruction is in oral or written format).

I understand that when I am not able to document and declare my use of AI tools, this behavior will be deemed as cheating in examinations and assessments (Disciplinary Code 1.1 c.), as I referred to “unauthorized notes, books, electronic devices, or other reference material”.

AI Tool	**AI Tool Used For**
ChatGPT	• Coding Errors • Knitting Errors
Grammarly	• Spelling and punctuation

Executive Summary

This report presents a comprehensive statistical data analysis of the sales and operational data from various product categories of a company. The analysis is conducted within the framework provided by the Quality Assurance 344 module, in line with the ECSA graduate attribute 4. This demonstrates the application of industrial engineering analysis tools to ensure process stability, profitability, and reliability.

The report is initiated with the exploration of the provided datasets, namely, sales2022and2023, products_data and customer_data, through applying descriptive statistics in order to identify patterns and trends within the datasets, do interpret and provide insights. The distribution of revenue and selling price was skewed to the right, which reflects the presence of large order quantities and highly priced items. Analysis of the customer demographics determined a wide income spread that follows a normal distribution, which suggests a diverse consumer base. As expected, the correlation analysis revealed positive relationships between selling price and revenue, and between picking and delivery hours.

In part 3 of the project, statistical process control (SPC) was conducted to develop X (bar) and s-charts for each product type, using subgroups of 24 observations as required. Phase I results were used as the baseline control limits, confirming overall process stability, and phase II monitoring identified various extreme control points for variability and mean delivery time. This was seen especially in product types where a high order quantity was recorded. The process capability analysis (Cp, Cpk) concluded that most product types approached a Cpk of smaller than or equal to 1.00, with only a subset achieving the preferred Cpk greater than or equal to 1.33 value, which indicates that there are opportunities to improve in terms of consistency of delivery times.

In part 4 of the analysis, error probabilities and data correction were conducted. Type-I error (manufacturer's error) probabilities for the three SPC rules remained below 1%, which signified a strong control sensitivity, while the Type-II error (consumer's error) probability of approximately 0.14 confirmed detection power deemed as acceptable. Systematic data errors in the products_Headoffice dataset were corrected by repeating verified local pricing patterns and restoring full catalog consistency; hence, the updated sales analysis confirmed accurate total 2023 values and aligned product classifications.

In part 5 of the analysis, a profit optimization model was developed to simulate a coffee-shop scenario based on the timeToserve2 dataset. This was used to simulate across two-to-six baristas, which indicated that four baristas maximized the daily profit of the coffee shop, while maintaining service reliability above 95%, balancing throughput and labour cost.

Part 6 applied ANOVA to test differences in delivery time means for the SOF product category throughout the years. The analysis concluded a significant year effect ($p<0.05$) with homogenous variances, implying measurable performance improvement over time.

Part 7 of the analysis addressed workforce reliability in a car-rental business. A graph representing the number of workers present was provided, and a probabilistic model was developed to estimate that maintaining or more equal to 15 staff on duty ensured reliable service. The optimization model indicated that three permanent employees result in

maximum annual gain, reducing downtime losses while keeping personnel costs at a manageable level.

Overall, the report integrates basic descriptive statistics, statistical control, risk estimation, data preparation and quality review, optimization techniques, and modeling techniques into an in-depth decision-support framework. The findings of the report satisfy the ECSA GA4 outcomes through evidence of competence in data analysis, process improvement, and operational decision-making within an industrial engineering context.

Table of Contents

List of tables:

List of figures:

Nomenclature

Acronym:	Description:
n	Sample size, number of observations
$\bar{X}$	Sample mean of subgroup
s	Sample standard deviation
$\bar{\bar{X}}$	Grand mean of sample means in phase-I
$\bar{s}$	Average sample standard deviations in phase I
CL	Center line on SPC charts
UCL	Upper control limit
LCL	Lower control limit
U1, U2	±1σ and ±2σ upper zone limits
L1, L2	±1σ and ±2σ lower zone limits
σ	Process standard deviation
Cp	Process capability index
Cpk	Process capability index
LSL	Lower specification limit
USL	Upper specification limit
VOC	Voice of the customer
Type-I error (α)	False alarm concludes process is out of control when it is not.
Type-II error (β)	Missed detection, failed to detect that the process is out of control
Power	1-β, probability of detecting a true change in process
Run	A consecutive sequence of SPC points following a pattern
ANOVA	Analysis of variance
MANOVA	Multivariate analysis of variance
SPC	Statistical process control
R	Statistical programming language used
DOE	Design of experiments
AOV	Average order value
CLT	Central limit theorem
η^2	Eta squared effect size in ANOVA
Reliability (%)	Probability that the system meets the performance threshold
Revenue (R)	Sales income
Profit (R)	Total revenue – labor cost
Capacity	Number of customers served per day
Levene’s test	Test for equal variances assumption

Introduction

Quality engineering and data analysis play a critical role in modern industrial systems, where operational excellence must find a balance between process stability and cost efficiency, while satisfying customer expectations. In this report, data analysis and statistical engineering tools are utilized on real-world retail and service delivery data to determine patterns, system behavior, evaluate capability, and propose improvements based on data-driven models. The work is structured according to the Engineering Council of South Africa (ECSA) graduate attribute 4 (GA4) outcome. This graduate attribute assesses competence in applying data analytics, statistical process control, and optimization methods in the context of complex engineering contexts.

This study uses various datasets, representing retail sales transactions over different years, product catalogs, customer demographics, and service operations of a coffee shop. The sales data from 2022-2023 and 2026-2027 (future) contains over 100,000 observations from a company's transactional system. This includes many features and delivery times, which form the core metric in proving performance analysis. The customer_data dataset allows for descriptive profiling of market segments and analyzing the company's customer demographics and distributions. The products_data dataset supports commercial analysis and data quality review, and preparation. An additional dataset was given, namely timeToServe2. As instructed, the timeToServe2 dataset provided barista service times for two coffee shops, which allowed for an optimization model to be developed. Further, a graph was visualized to represent the number of employees for a car rental agency, to form a staff reliability dataset to allow for service performance modeling.

The primary objective of the report is to use the given datasets and analyze, interpret, and optimize data-driven systems by means of descriptive statistical methods with engineering reasoning and problem-solving. The exact objectives are given below:

- Utilize descriptive statistics to gather insights into behavior and patterns within the products_data, customers_data, and sales2022and2023 datasets.
- Implement statistical process control (SPC) using $\bar{X}$ and s-charts to monitor the delivery process stability.
- Assess process capability and compare product types against customer-defined performance limits.
- Evaluate type-I and type-II errors/risks in decision making during process monitoring.
- Correct data errors and inconsistencies in the products_Headoffice dataset and assess the impact of the corrected dataset.
- Optimize service profitability for a coffee shop, using simulation-based staffing decisions.
- Utilize ANOVA or MANOVA to determine whether the delivery trend significantly changed over the period for which we have data available.
- Evaluate service reliability and workforce optimization using probabilistic modeling.

An integrated methodological approach is followed, which combines statistical inference, probabilistic modeling, simulation, and quality control techniques. RStudio and the R programming language are used throughout the report. The analysis follows the project brief numbering system sequentially.

In summary, this report demonstrates data-driven decision-making for industrial engineering, providing answers, justified by statistics, to complex engineering and business operational problems.

1.1 Data preparation and quality review

1.1.1 Data integrity assessment

The datasets used in the report include historical sales data (sales2022and2023), future sales data (sales2026and2027), customer demographics (customer_data), product data (products_data), product head office data (products_Headoffice), and service time data for two coffee shops (timeToServe2). A preliminary data integrity assessment was performed to verify the consistency, structure, completeness, readiness, and missingness of these datasets.

Initially, schema validation checks were implemented using the stopifnot() function to ensure that each dataset contained the expected variables before any transformations were conducted. this helped to prevent execution on datasets with inconsistent or incorrect data. Key variables such as ProductID, SellingPrice, orderYear, and deliveryHours were confirmed to exist across both sales files, demonstrating structural consistency.

Conducting integrity checks revealed no column naming conflicts; however, it was observed that product reference data from the products_Headoffice datasets contained structural inconsistencies and missing product type prefixes, which were corrected later per instruction in section 4.3. The remainder of the datasets were structurally sound and free from incorrect records.

Missing values were handled conservatively and transparently, through applying the na.rm = TRUE function in summary statistics and modeling functions, instead of imputation, to mitigate bias. This enabled traceability and preserved the integrity of the original datasets. Based on this assessment, the datasets were seen as complete and suitable for statistical analysis once column validation and structured joins were performed.

1.1.2 Data standardization and structuring

To enable consistent statistical modeling, the datasets were standardized and enriched with derived variables to account for distinct variable names within the datasets. Data fields in the sales data were reconstructed using the lubridate::make_date() function, which combined orderYear, orderMonth, orderDay, and orderTime into a single order_date field. Similarly, a precise chronological timestamp, order_key, was created for the statistical process control sequencing using make_datetime(), to include hour-level production timing.

The primary key ProductID was decomposed by using a custom parsing function, parse_pid, to extract the first three letters of the ProductID value (SOF123 = SOF). This categorization enabled the grouping of sales by product category, which was essential for constructing control charts later in part 3.

The sales2026and2027 dataset was sorted chronologically by product type and order time to simulate realistic data collection. Subgrouping was performed per instructions, using sample sizes of n=24, following SPC convention for subgroup means. The first 30 samples per product type were utilized as the baseline for Phase I, to estimate control limits and sigma zones. Subsequent samples formed the Phase II monitoring data and were used to detect violations of the control rules.

Integration between datasets was achieved by joining the datasets, without changing the original records, and while maintaining data integrity. The formula used to calculate revenue, which enabled later analyses linking product value to delivery performance, is given below:

$$Revenue\ (R) = Quantity\ x\ SellingPrice$$

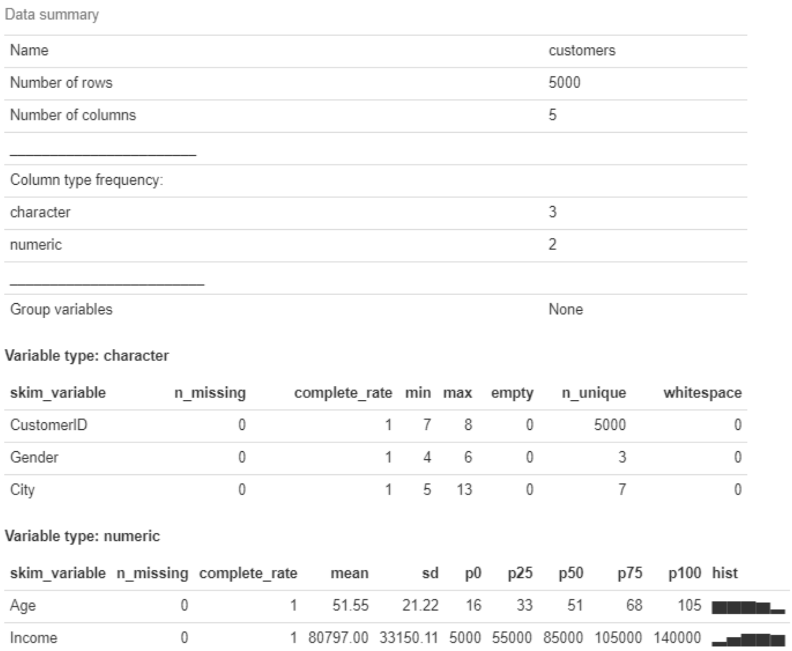
Part 1.2: Descriptive Statistics; Descriptive and profitability analysis

The descriptive and profitability analysis provides a statistical overview of the business's operational and sales performance. The objective of this section is to apply descriptive statistical methods and techniques to gather insight into the business and its operations. Through graphical and numerical exploration, the analysis identifies how product attributes, time-based factors, and customer characteristics influence the company's revenue outcomes. By interpreting these descriptive findings, a deeper understanding is developed to support evidence-based managerial decision-making.

1.2.1 Product profitability

The sales dataset consists of 100000 transaction records, with 12 variables describing selling price, order quantity, associated time, revenue, and operational metrics. Table 1.2.2.1 summarises the descriptive statistics for the key numeric variables. The average selling price across all transactions was approximately R3224, with a substantial standard deviation of R5412. This indicates a wide product price range across all categories. The mean order quantity of 13.5 units per transaction, with a median of 6, an interquartile range of 3-23, suggests that most customers make small-volume purchases, while a smaller subset places large orders of up to 50 units. These differences are visible in Figure 1.2.2.2 (Distribution of order quantity), which displays a strong right-skew, with over half of all orders containing fewer than ten items. Such a skew indicates a demand structure dominated by frequent, low-quantity orders, which is common in retail or consumer markets.

Table 1.2.1.1: Structural overview of customer_data



Data summary

Name	customers
Number of rows	5000
Number of columns	5
______________________	
Column type frequency:	
character	3
numeric	2
______________________	
Group variables	None

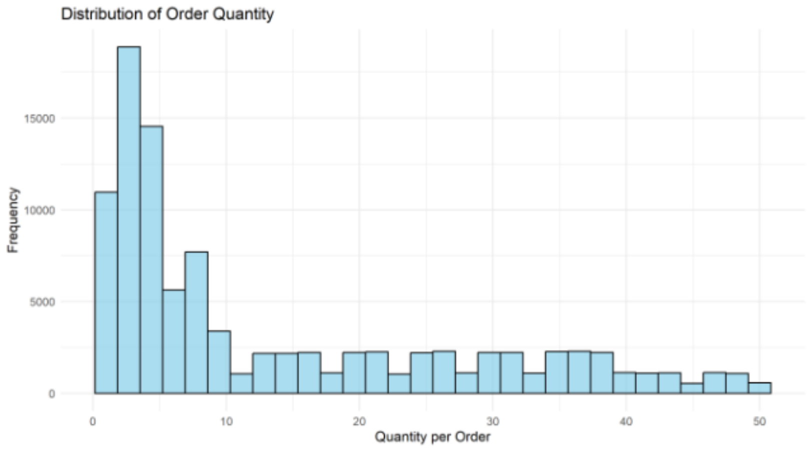
Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
CustomerID	0	1	7	8	0	5000	0
Gender	0	1	4	6	0	3	0
City	0	1	5	13	0	7	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Age	0	1	51.55	21.22	16	33	51	68	105	▇▇▇▆▁
Income	0	1	80797.00	33150.11	5000	55000	85000	105000	140000	▁▃▇▇▆

Figure 1.2.2.2: Distribution of order quantity



The distribution of selling prices, as seen in Figure 1.2.2.3, similarly displays a heavy right tail. The majority of product prices cluster below R2000, while a small group of premium items reach beyond R15000. This pattern is consistent with the summary of products_data, where the mean product selling price is R4493.59, with a standard deviation of R6503.77, and the overall range of values extends from R350 to R19725. Such a large spread indicates a product portfolio consisting of both low-cost accessories and high-end technology items, as confirmed in Figure 1.2.2.4 (Selling price distribution).

Figure 1.2.2.3: Distribution of unit price (histogram and density)

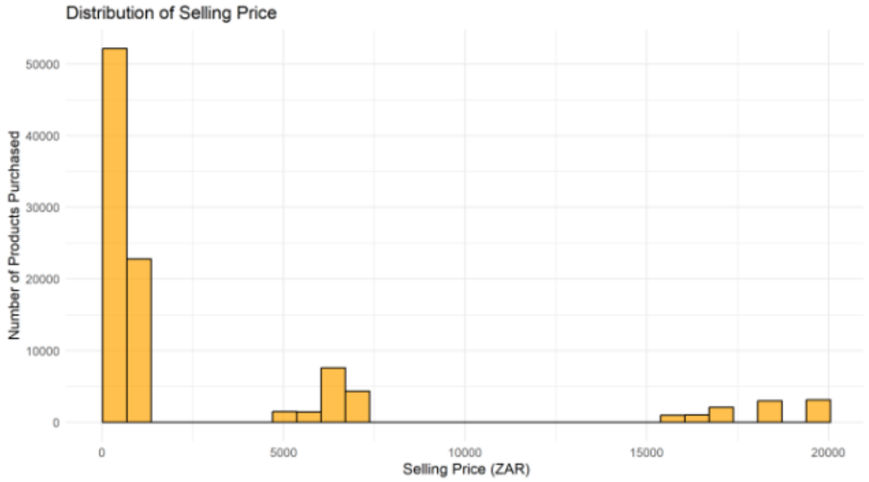
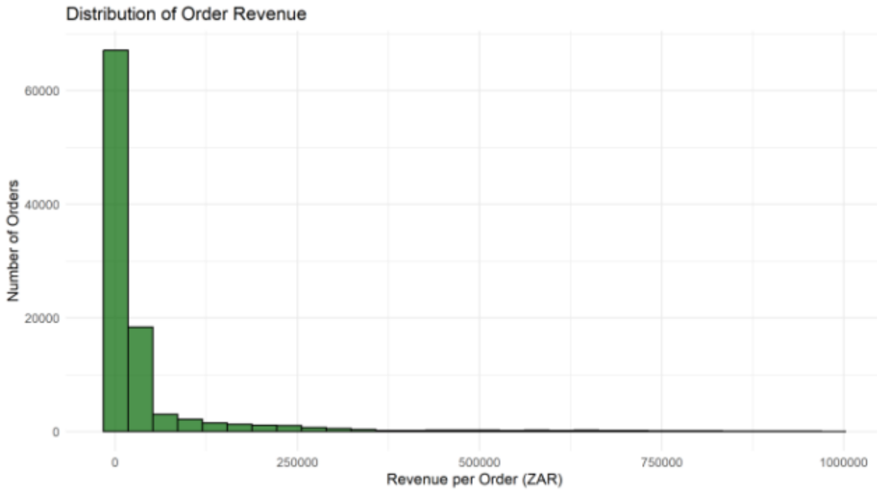


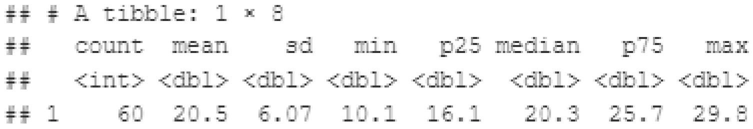
Figure 1.2.2.4: Distribution of line value (revenue per line)



When multiplied by order quantities, this variation in unit price causes significant dispersion in order revenue. Figure 1.2.2.4 (Distribution of order revenue) indicates that most transactions are under R50000, with outliers exceeding R250000. The maximum of these values reaches nearly R986000. This long-tailed revenue structure implies that a small number of high-value orders account for a disproportionately large share of total sales, which highlights the importance of targeting and retaining high-spending customers.

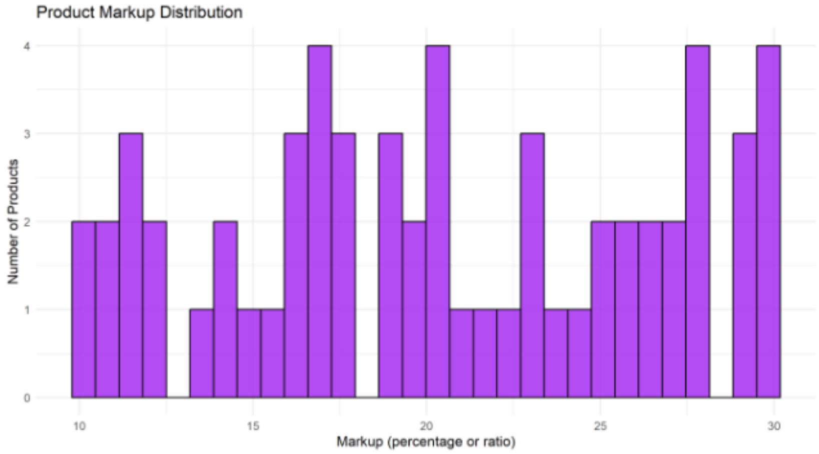
Profitability at the product level is further explained through applying statistical techniques to the markup (profit margin) data. As summarised in Table 1.2.4.5 and visualized in Figure 1.2.4.6 (Product markup distribution), average markup across all products is 20.46%, with a standard deviation of 6.07%, and values ranging between 10%-30%. The near-uniform distribution of markup levels suggests a pricing policy that maintains consistent profitability margins across categories. However, Figure 1,2,4,8 (Markup by product category) reveals moderate inter-category differences, with software and cloud subscriptions showing slightly higher markup medians relative to hardware items, such as monitors and keyboards. This pattern implies that software products, which carry lower production and distribution costs, are more profitable than the other products in the company catalog.

Table 1.2.4.5: Markup: Summary statistics



```
## # A tibble: 1 × 8
##   count  mean    sd   min   p25 median   p75   max
##   <int> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1    60  20.5  6.07  10.1  16.1   20.3  25.7  29.8
```

Figure 1.2.4.6: Markup distribution (histogram)



Correlation analysis further supports the consistency of these findings. As seen in Figure 1.2.2.15 (Scatterplot matrix of sales variables), revenue has a strong positive relationship with both selling price (r = 0.657) and picking hours (r = 0.554), which implies that high-value orders generally involve higher-priced products and longer fulfillment times. The positive link between picking hours and delivery hours (r-0.582), observed also in Figure 1.2.2.12 (Picking vs delivery hours), confirms that logistical complexity scales with order value. Collectively, these results show that product profitability is largely driven by a small share of high-value, operationally intensive transactions, while the remainder of the product base contributes steady, but lower-margin sales.

Figure 1.2.2.15: Scatterplot matrix: Sales

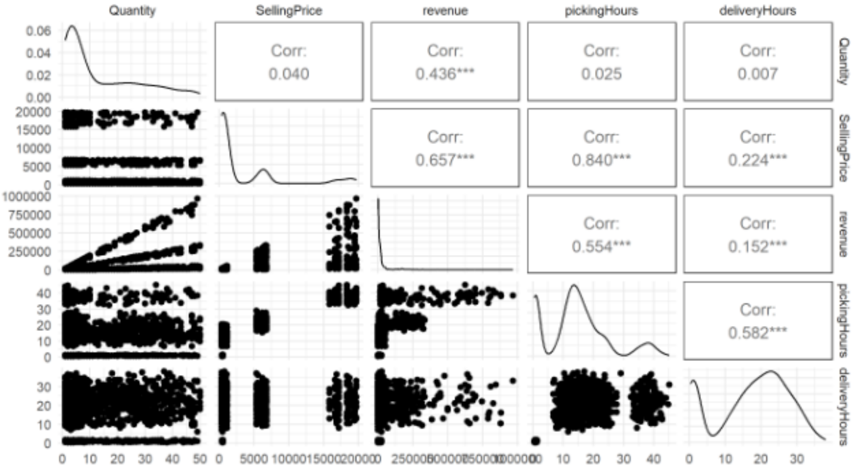
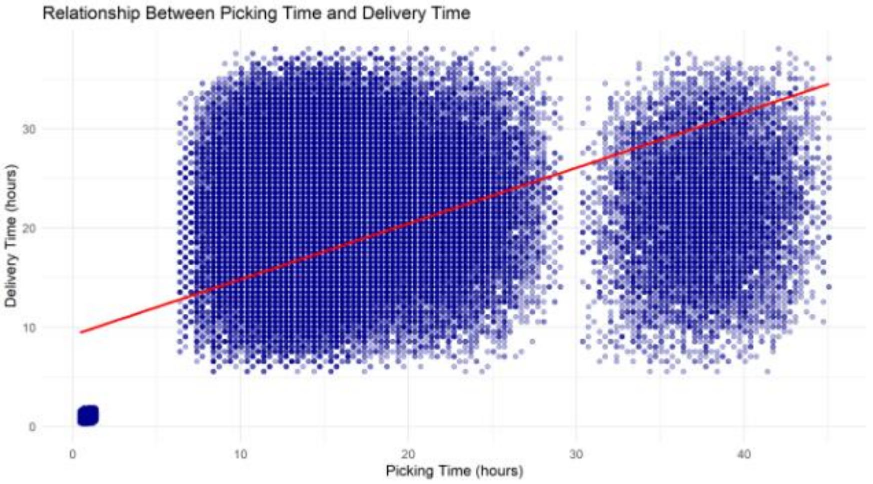


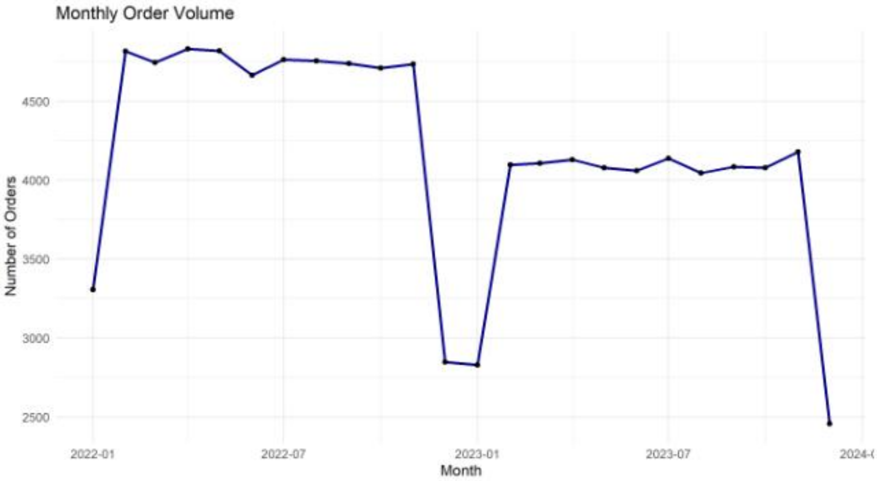
Figure 1.2.2.12: Relationship: Picking vs delivery hours



1.2.2 Temporary and seasonal trends

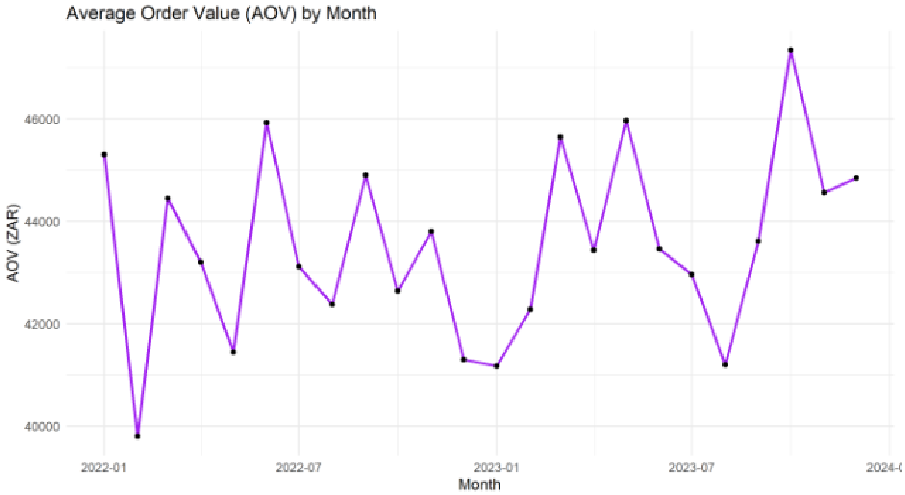
Time-based analysis provides insight into seasonality and workload variation. Figure 1.2.2.5 (Monthly order volume) indicates that order activity remained consistent through 2022, averaging around 4500 orders per month, after which it decreased sharply at the start of 2023. A similar pattern occurs in early 2023, as seen in Figure 1.2.2.5 (Monthly revenue trend), where total revenue fell noticeably but managed to recover by mid-2023. Synchronous trends suggest a temporary external disruption, which could potentially be due to supply constraints or seasonality.

Figure 1.2.2.5: Monthly orders; Count orders per month



Despite these fluctuations, Figure 1.2.2.7 (Average order value by month) indicates average order values ranging between R40000 and R46000, with a slight upward trend at the end of 2023, which implies that while the number of transactions fluctuated due to seasonality, the customers who purchased tend to spend the same or higher amounts over a period of time. The seasonality observed is therefore more volume-driven than value-driven.

Figure 1.2.2.7: Average order value



Day-of-week and hourly analyses reveal further operational regularities, as seen in Figure 1.2.2.8 (Orders by day of the week), which indicates that orders are evenly distributed from Sunday to Saturday, with only minor variation. This suggests consistent and continuous engagement from customers throughout the week. However, Figure 1.2.2.9 (orders by hours of day) demonstrates a distinct daily cycle, with peak order activity between 09:00 and 18:00, which aligns with the hours spent at work. This pattern is operationally significant, as it informs warehouse staffing and system capacity planning, particularly for order processing and fulfillment during midday surges.

Figure 1.2.2.8: Orders by day of week

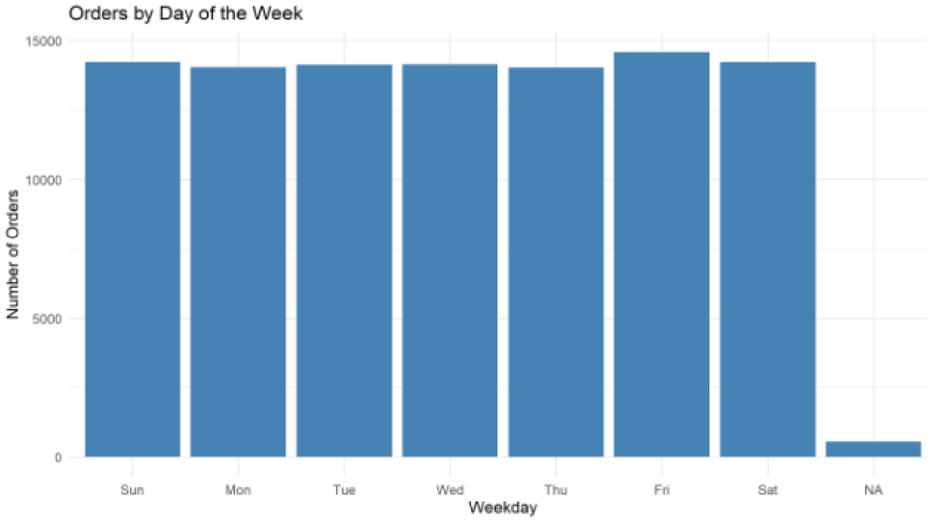
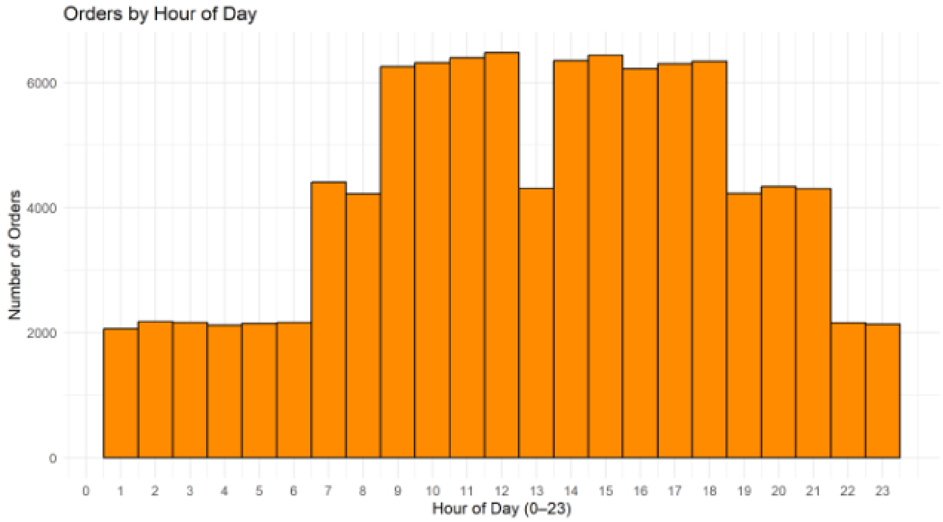


Figure 1.2.2.9: Orders by hour of day



Operational dynamics are shown in Figures 1.2.2.10 and 1.2.2.11, showing the picking hours and delivery hours distributions. Picking times exhibit a bimodal distribution, with one concentration near zero and another around 10-20 hours, while delivery times approximate a normal distribution centered at approximately 20 hours. Together with Figure 1.2.2.13 (delivery time by day of week), these charts confirm that delivery performance is consistent across weekdays, implying a well-standardized logistics process. From a productivity perspective, consistent delivery times reduce uncertainty in lead times and contribute to customer satisfaction, which can indirectly sustain future revenue growth.

Figure 1.2.2.10: Operational timing: Picking hours distribution

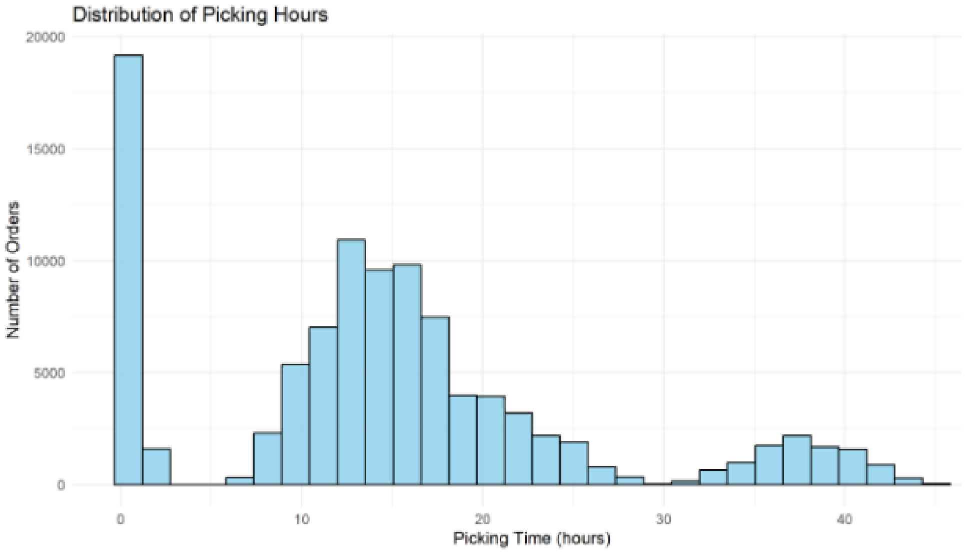


Figure 1.2.2.11: Operational timing: delivery hours distribution

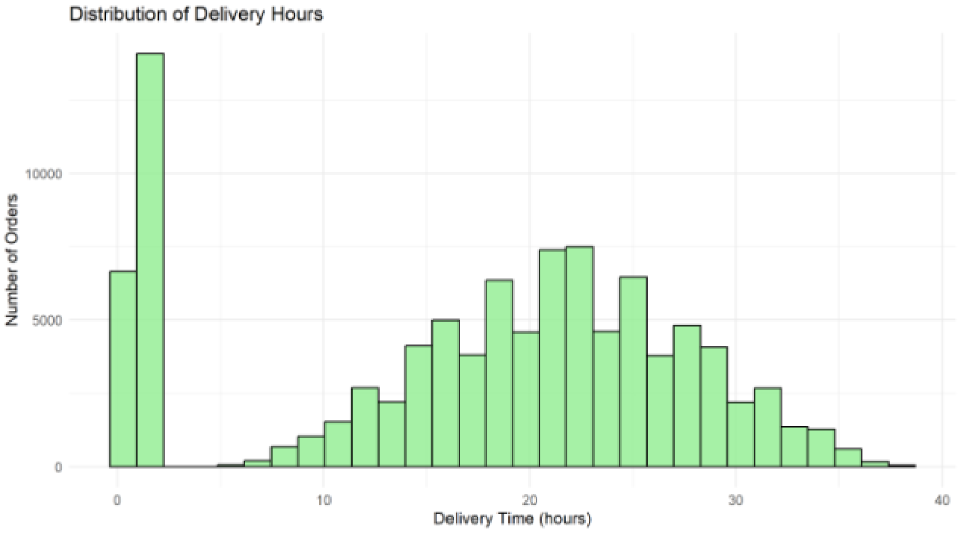
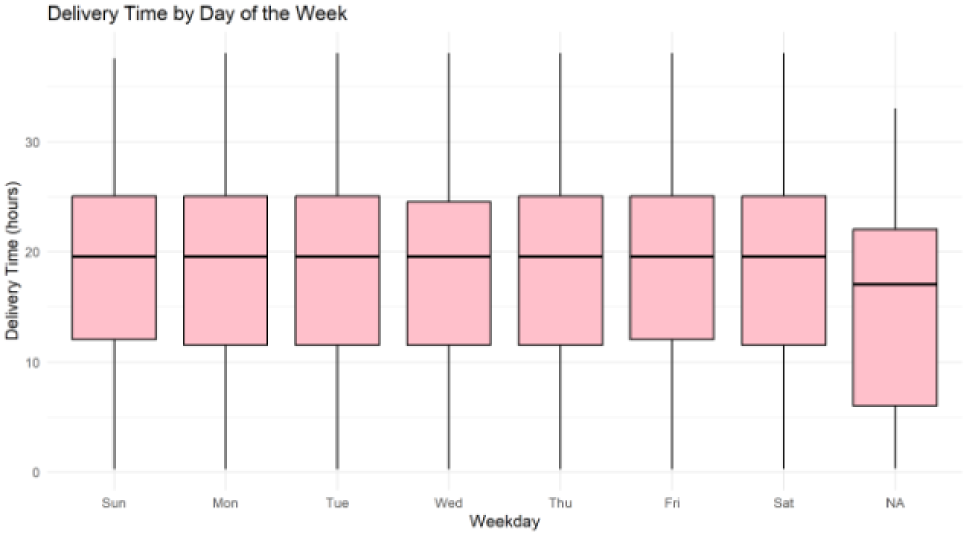


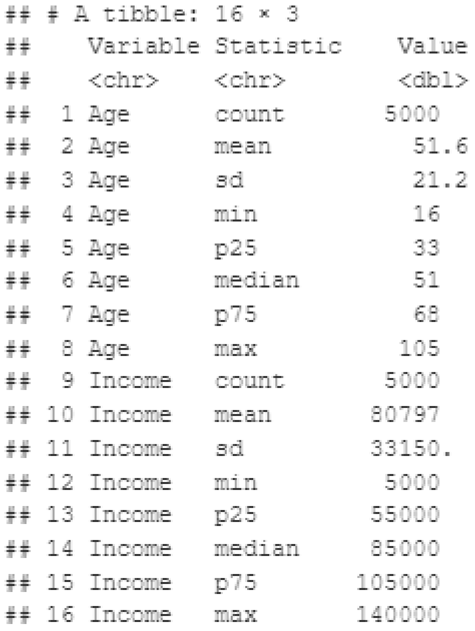
Figure 1.2.2.13: Delivery hours by weekday (boxplot)



1.2.3 Customer profitability

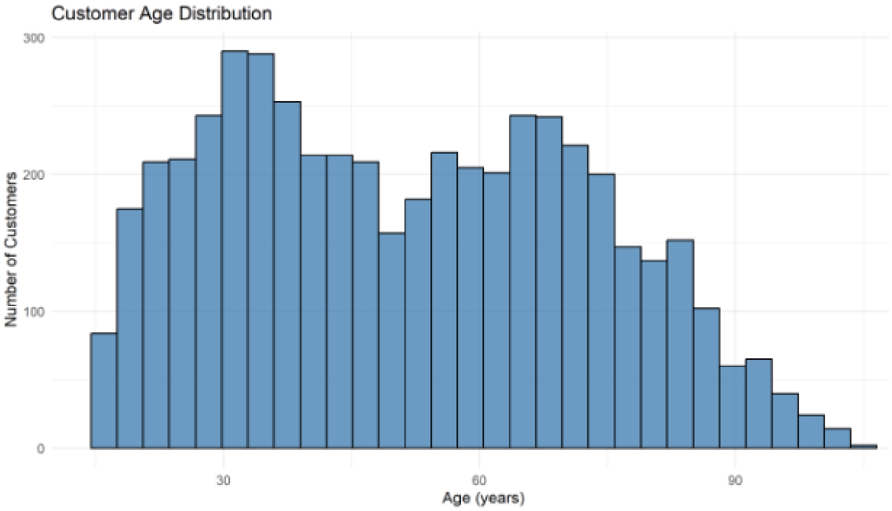
Customer-level analysis integrates demographic, geographic, and income-related information to identify profitable market segments. As seen in Table 1.2.3.1, the customer base comprises 5000 individuals with an average age of 51.6 years and an income averaging R80797 per month. Figure 1.2.3.2 (Customer age distribution) reveals a bimodal age pattern, with peaks around 30 and 70 years, indicating two dominant customer cohorts, namely, younger professionals and older-higher income individuals. This customer demographic, consisting of two distinct market segments, indicates potential for direct marketing efforts to be conducted per demographic group, which can further motivate older/wealthier clients to purchase more premium goods, and younger individuals to purchase more lower-priced goods.

Table 1.2.3.1: Customer demographics: Summary table



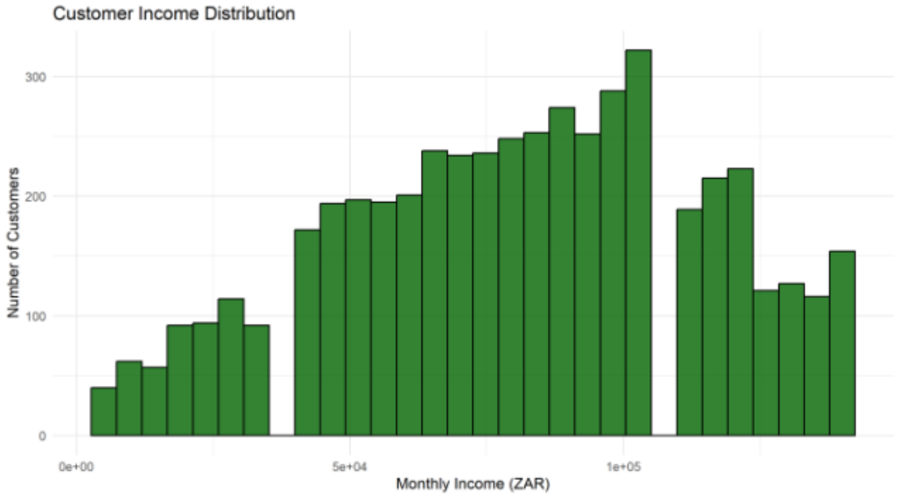
```
## # A tibble: 16 × 3
##    Variable Statistic    Value
##    <chr>    <chr>        <dbl>
##  1 Age      count       5000
##  2 Age      mean          51.6
##  3 Age      sd            21.2
##  4 Age      min           16
##  5 Age      p25           33
##  6 Age      median        51
##  7 Age      p75           68
##  8 Age      max          105
##  9 Income   count       5000
## 10 Income   mean       80797
## 11 Income   sd         33150.
## 12 Income   min         5000
## 13 Income   p25        55000
## 14 Income   median     85000
## 15 Income   p75       105000
## 16 Income   max       140000
```

Figure 1.2.3.2: Age distribution (histogram)



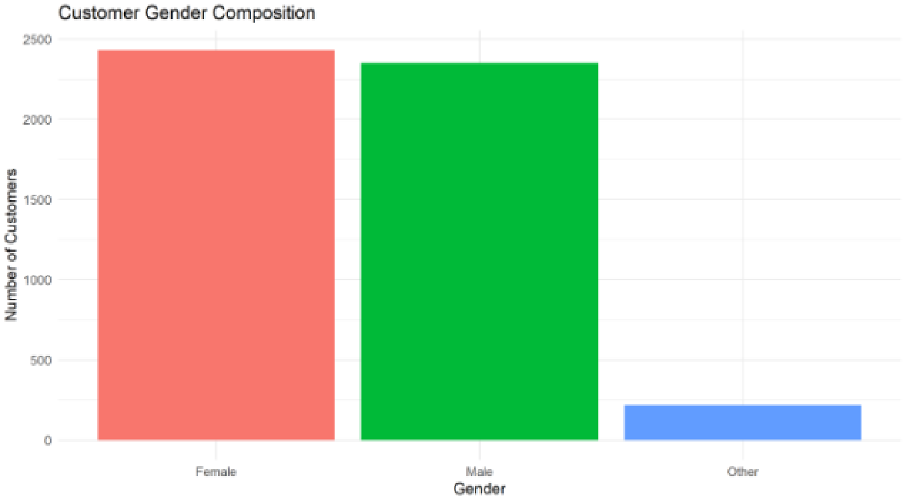
The distribution of customer income (Figure 1.2.3.3), as displayed above, supports this observation. Income values are focused around R55000 and R105000, with all the values ranging between R5000 and R140000. This indicates a wide range of purchasing power in the client demographic.

Figure 1.2.3.3: Income distribution (histogram)



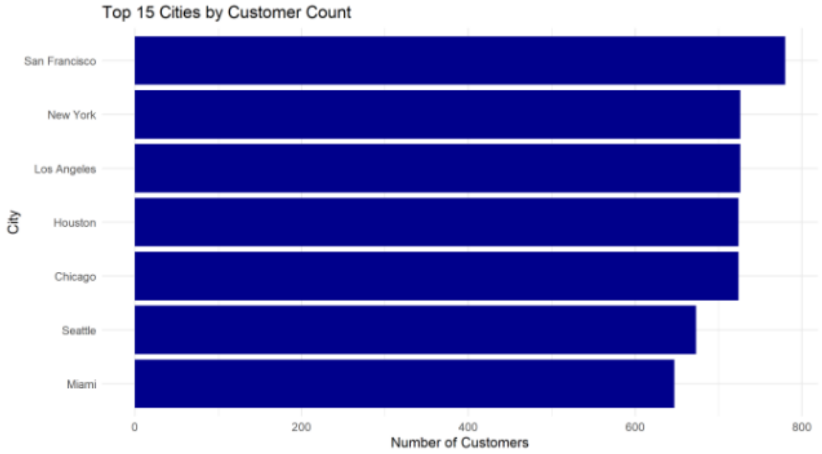
The gender distribution displayed below (Figure 1.2.3.5), indicates an almost even split between genders, 2450 to 2300, with a minority identifying as “Other”, which could potentially be due to data collection errors. This balance between customer gender is a positive indication for the company that shows that the target market is widespread across both genders and different income levels.

Figure 1.2.3.5: Gender composition (bar chart)



From figure 1.2.3.7 it is indicated that cities such as San Francisco, Los Angeles, New York, Chicago, Houston, Seattle, and Miami are among the top 15 cities recorded by customer count. This indicates that these cities mentioned likely drive the majority of total revenue. Focusing marketing distribution resources on these locations would therefore yield greater profitability and logistical efficiency.

Figure 1.2.3.7: Top 15 cities by customer count (bar chart)



Overall, the customer analysis reveals that profitability is concentrated among high-income, urban customers who frequently purchase mind to high-priced technology products. The combination of consistent demand, premium spending capabilities, and effective delivery times provides a good position for the company from which to conduct business and sustain strong profit margins.

1.2.4 Coding explanations

This section involves the initial exploration of customer, product, and sales datasets to understand their structure, determine key characteristics, and conduct descriptive analysis. In order to do that, the following code was used:

- skim(): the function offers quick overviews of the datasets in terms of variables, data types, and completeness.
- Dplyr: summary statistics. This function calculates key summary statistics (count, mean, standard deviation, min, quartiles, max for numeric variables and reshapes results into a long format for clarity and comparison across variables.
- ggplot(): used for histograms and densities. This visualizes the distribution of singe variables, such as order quantity, selling price, revenue, picking hours, and delivery hours, to reveal spread, skewness, and outlier behavior. This function was also used for categorical or time plots for temporal trends to show how business activities evolve over an extended period of time to investigate seasonality and patterns.
- Correlation matrix and heatmap: computes pairwise correlations among numeric variables to determine relationships between variables simultaneously, assess outlier patterns, and identify clusters visually.
- GGally::ggpairs(): this function generates a scatterplot matrix for selected numeric variables to identify internal relationships, clusters, and outlier patterns.

1.2.5 Descriptive statistics conclusion

To conclude the descriptive statistics section of the report, the section provided an integrated view of product, customer, and time-based dimensions that collectively influence the business performance. The data findings confirm that the company's profitability is driven primarily by a small group of premium, high-margin products sold to high-earning individuals. From the sales analysis, it is seen that moderate seasonality affects business operations, but consistent operational performance indicates a stable and responsive supply chain strategy. The correlation between order fulfillment time, order value, and selling price emphasizes the operational complexity needed for premium orders, which suggests that the productivity of the company is dependent on both efficiency and the pricing strategy followed. To conclude, the analysis establishes a statistically sound foundation from which to conduct SPC and capability studies, where operational performance and variation will be discussed in great detail.

Part 3: Statistical Process Control (SPC)

3.1 Phase I: Initialization and stability of the delivery-time process (n=24)

Records in sales2026and2027 were first time-ordered by year, month, date, and order hour, producing a coherent simulation of real-time data in sequence. Product types were derived from the first three characters of the ProductID primary key, and all six types exceed the greater than or equal to 30x24=720 observations requirement. For each product type, the first 30 subgroups of 24 deliveries were used to compute center lines and the 1σ/2σ/3σ limits for all the s-charts and X-bar charts with constants A3 = 0.619, B3 = 0.555, and B4 = 1.445.

Across all six s-charts (Appendix B: Figures 3.1.7.1-3.1.7.6) and the faceted overview, as seen below in Figure 3.1.8, no subgroup standard deviation exceeded its 3σ limit during phase I. Typical variability levels are summarised in Table 3.1.6. For hardware types, the average within-subgroup spread was s̄ approximately equal to between 5.6 and 5.9h, whereas SOF exhibits a much smaller spread (CLs = 0.297h), consistent with its substantially shorter delivery times, the absence of Phase I breached, together with visibly pattern-free traces on the X-bar panels (Figures 3.1.9.1-3.1.9.6 to be viewed in appendix B) in Figure 3.10 below, indicates that all product types achieved an acceptably stable baseline from which to proceed to Phase II monitoring.

Figure 3.1.8: Faceted s-charts

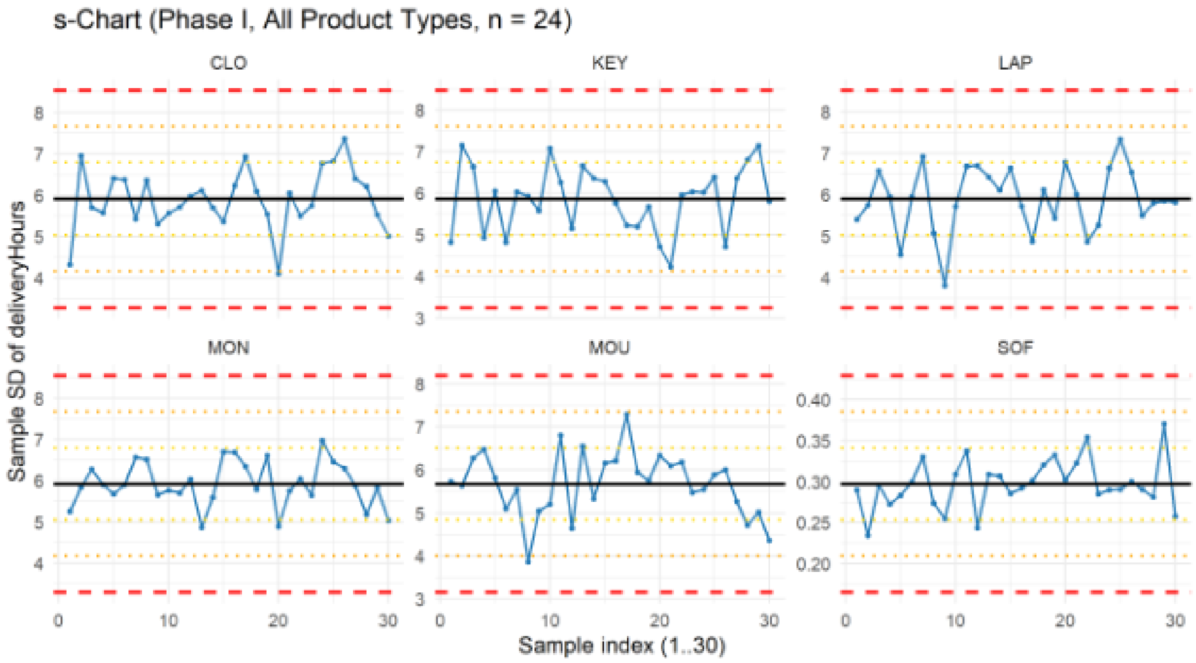
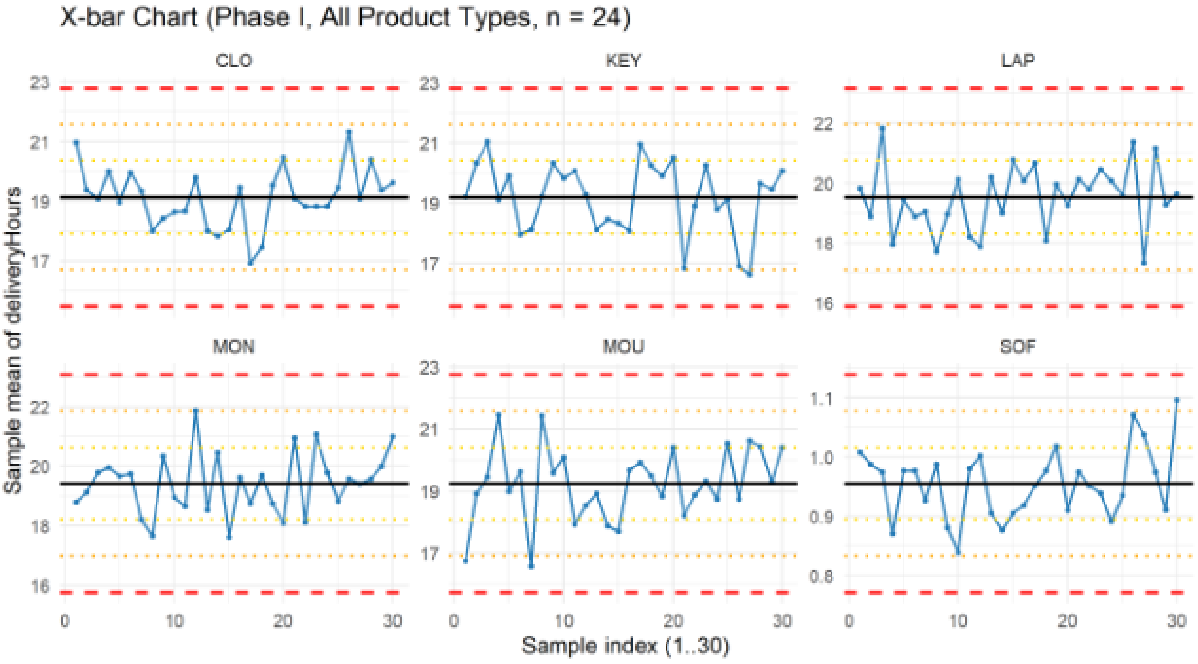


Table 3.1.6: Control-chart constants and limits for every type

```
## # A tibble: 6 × 13
##   product_type  CL_s   CL_x  L1_s   L1_x  L2_s   L2_x  U1_s  U1_x  U2_s  U2_x
##   <chr>        <dbl>  <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 CLO          5.91   19.1   5.03   17.9   4.16   16.7   6.78  20.3  7.66  21.6
## 2 KEY          5.86   19.2   4.99   18.0   4.12   16.8   6.73  20.4  7.60  21.6
## 3 LAP          5.89   19.5   5.02   18.3   4.14   17.1   6.76  20.7  7.64  22.0
## 4 MON          5.92   19.4   5.04   18.2   4.17   17.0   6.80  20.6  7.68  21.9
## 5 MOU          5.68   19.2   4.83   18.1   3.99   16.9   6.52  20.4  7.36  21.6
## 6 SOF          0.297  0.956 0.253  0.894 0.209  0.833 0.341  1.02 0.386  1.08
## # ℹ 2 more variables: UCL_s <dbl>, UCL_x <dbl>
```

Figure 3.1.10: Faceted X-bar charts:



Since the spread was in control during initialization, mean shifts detected later in phase II can be interpreted as genuine changes in process centering rather than artifacts or inflates within-subgroup variation, but this will be discussed in more depth in the subsection to come.

3.2 Phase II: Ongoing monitoring (Samples n=31, 32, …)

Phase II continues subgroups of n=24 from sample 41 onwards. This is done for each product type, using Phase I limits for signaling. The consolidated signal table, Table 3.2.2, and the faceted X-bar (Figure 3.2.3) and s-chart (Figure 3.2.4) views indicate the following:

- Rule 1 (Beyond ±3σ) X-bar: A Large number of mean-level breaches are observed across product types: SOF = 297, MOU = 288, CLO = 218, MON = 156, LAP = 109. The red markers in Figure 3.2.3 visualize these breaches and confirm that mean shifts are the dominant Phase II signal.
- Rule 1 (Beyond ±3σ) s-chart: Variability is largely well-behaved in live running, with only one sub-group for MOU breaching the UCLs (Sample 592, s=8.23 vs UCLs = 8.2h). All other product types show s-points beyond 3s = 0.

Table 3.2.2: Join phase-II stats to phase-I limits and flag rule-1

```
## # A tibble: 6 × 4
##   product_type n_phase2_samples x_points_beyond_3s s_points_beyond_3s
##   <chr>                   <int>              <int>              <int>
## 1 CLO                       620                218                  0
## 2 KEY                       717                249                  0
## 3 LAP                       396                109                  1
## 4 MON                       590                156                  0
## 5 MOU                       831                288                  1
## 6 SOF                       835                297                  0
```

Figure 3.2.3: Faceted Phase-II X-bar (All product types, free y-scales, Rule-1 highlights)

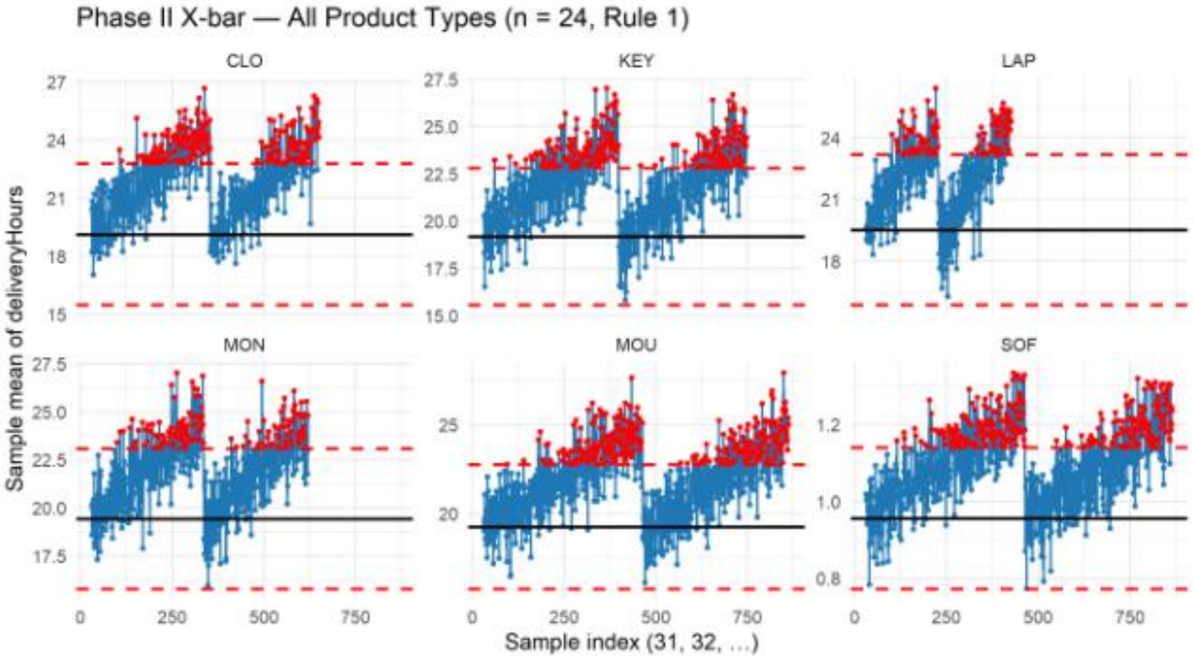
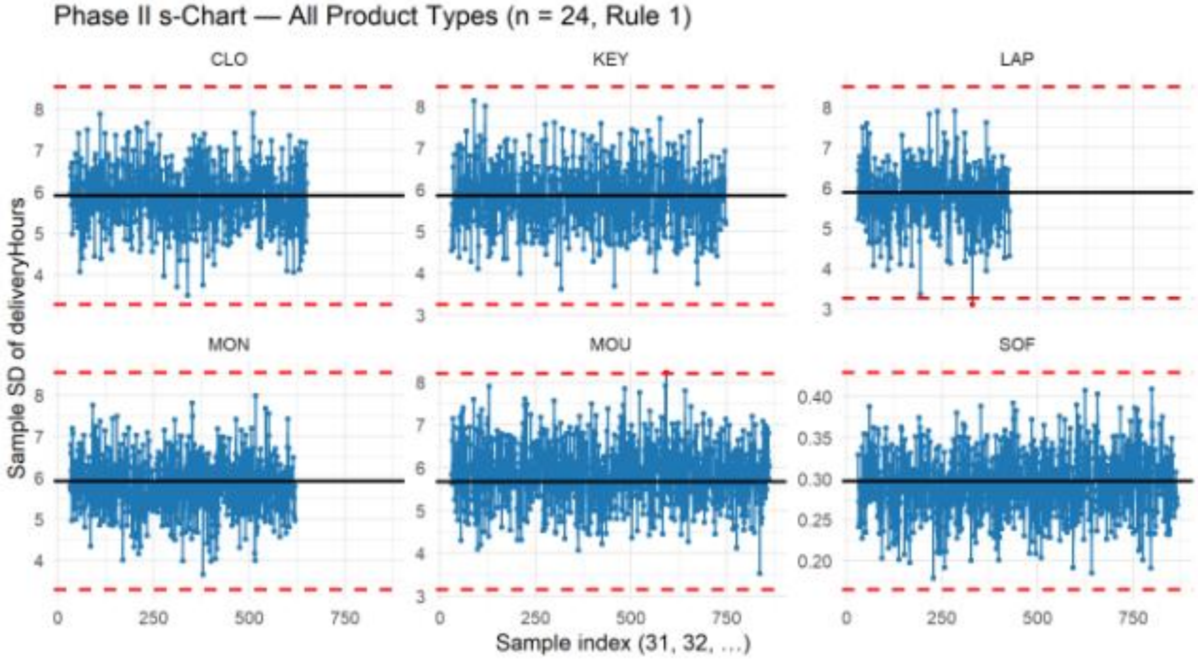


Figure 3.2.4: Faceted Phase-II s-chart (All product types, free y-scales, Rule-1 highlights)



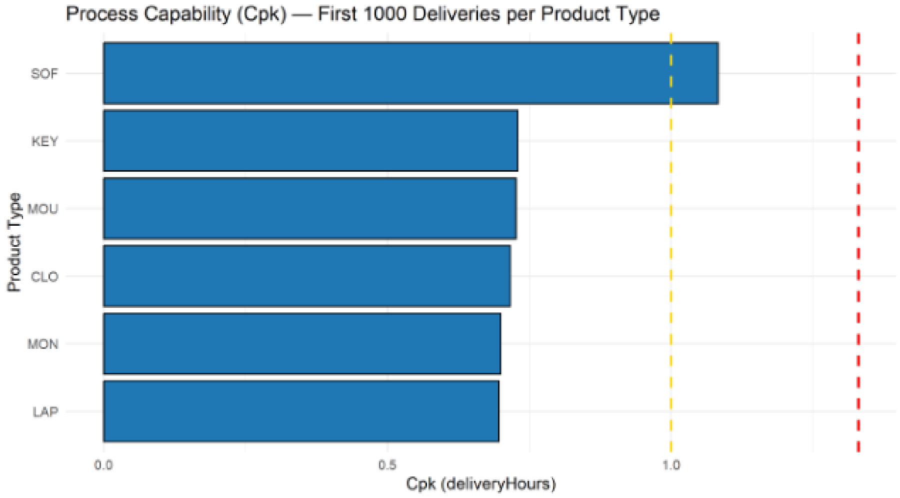
Considering the information provided above, day-to-day control issues stem mainly from changes in process centering (delivery-time means), not from explosions in within-subgroup spread. In practice, product managers should prioritize diagnosing assignable causes that shift the mean, for example, route choices, carrier allocations, etc., while retaining current controls on the picking and transport variance, which appear effective.

3.3 Capability analysis (first 1000 deliveries per type. LSL = 0h, USL = 32h)

Capability indices computed from the earliest 1000 deliveries per type, as seen in Figure 3.3.2, show a clear separation:

- SOF: $\mu = 0.955\text{h}, \sigma = 0.294\text{ h}, C_p = 18.10, C_{pk} = 1.08$, capable relative to the USL = 32h, which meets the $C_{pk} \geq 1.00$ minimum and is above the 1.00 benchmark line in Figure 3.3.2.
- Hardware types (KEY, MOU, CLO, MON, LA): $\mu \approx 19.2\text{–}19.6\text{ h}, \sigma \approx 5.8 - 6.0\text{h}$ with $C_{pk} \approx 0.696 - 0.729$, not capable against the 32h upper specification, and below the 1.00 capability reference in Figure 3.3.2.

Figure 3.3.2: Visual; Cpk by product type



3.4 Western Electric signals

Using a joint Phase II statistics with Phase I limits:

- A: One s-sample outside +3σ (UCLs):

Exactly one violation was detected, as seen in Table 3.4.1, MOU, sample 592, s = 8.23h vs UCLs = 8.20. Totals according to product type correspond with: MOU = 1, and all other types = 0.

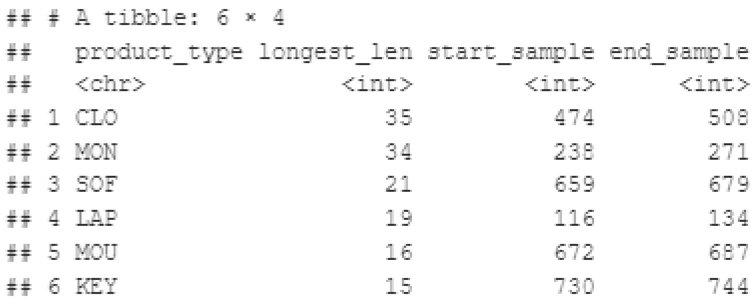
Table 3.4.1: s-chart; samples above +3σ UCL_s

```
## $first_3_violations
## # A tibble: 1 × 4
##   product_type sample_id     s UCL_s
##   <chr>            <int> <dbl> <dbl>
## 1 MOU                592  8.23  8.20
##
## $last_3_violations
## # A tibble: 1 × 4
##   product_type sample_id     s UCL_s
##   <chr>            <int> <dbl> <dbl>
## 1 MOU                592  8.23  8.20
##
## $totals_per_type
## # A tibble: 1 × 2
##   product_type n_violations
##   <chr>               <int>
## 1 MOU                     1
##
## $total_violations
## [1] 1
```

- B: Longest sequential control run with s within ±1σ:

The longest continuous runs inside the L1s to U1s band, as seen in Table 3.4.2, are: CLO = 35 samples (474-508), MON = 34 samples (238-271), SOF = 21 (659-679), LAP = 19 (116-134), MOU = 16 (672-687), and KEY = 15 (730-744).

Table 3.4.2: Longest good-control run of s within:



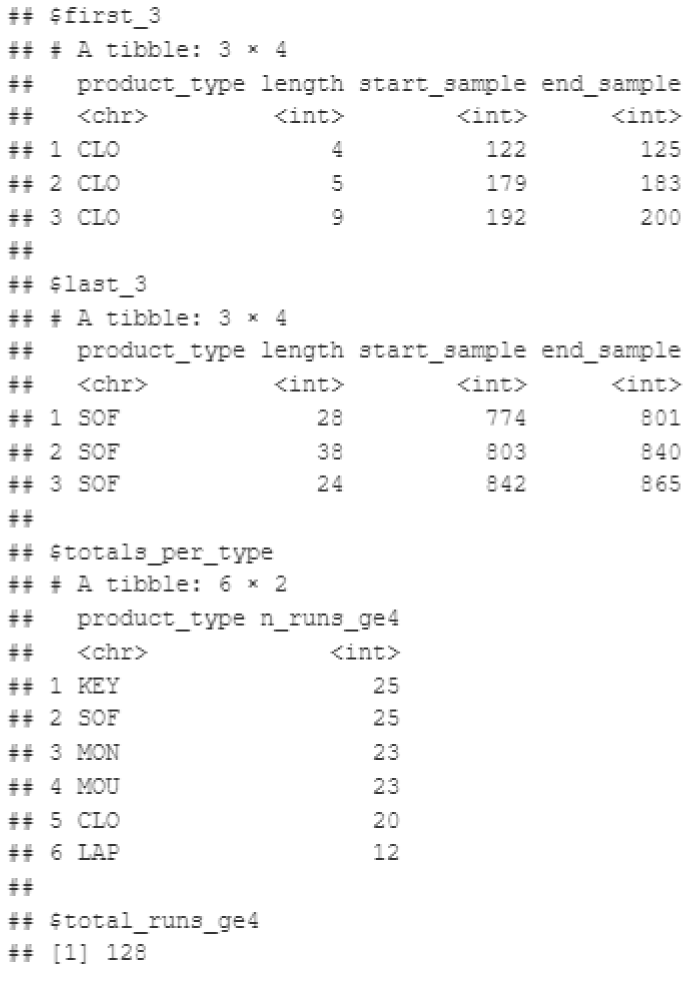
```
## # A tibble: 6 × 4
##   product_type longest_len start_sample end_sample
##   <chr>              <int>        <int>      <int>
## 1 CLO                   35          474        508
## 2 MON                   34          238        271
## 3 SOF                   21          659        679
## 4 LAP                   19          116        134
## 5 MOU                   16          672        687
## 6 KEY                   15          730        744
```

The CLO and MON types demonstrate extended periods of exceptionally tight spread control, which is operationally desirable even if center-line shifts trigger X-bar alarms elsewhere.

- C: Runs of ≥ 4 consecutive X-bar samples above the upper 2σ line:

Extended high-side mean excursions are frequent, as seen in Table 3.4.3. Total runs ≥ 4 by type: KEY = 25, SOF=25, MON=23, MOU=23, CLO=2-, LAP=12. Illustrative sequences include early in CLO, length 4 (samples 122-125) and length 5 (samples 179-183). Very long SOF sequences are identified later in the series of length 28 (samples 774-801), 38 (samples 803-840), and 24 (samples 842-865). These events are visible as dense clusters above the U2 line in Figure 3.2.3.

Table 3.4.3: 4 consecutive X-bar samples above the upper 2σ line



```
## $first_3
## # A tibble: 3 × 4
##   product_type length start_sample end_sample
##   <chr>         <int>        <int>      <int>
## 1 CLO               4          122        125
## 2 CLO               5          179        183
## 3 CLO               9          192        200
##
## $last_3
## # A tibble: 3 × 4
##   product_type length start_sample end_sample
##   <chr>         <int>        <int>      <int>
## 1 SOF              28          774        801
## 2 SOF              38          803        840
## 3 SOF              24          842        865
##
## $totals_per_type
## # A tibble: 6 × 2
##   product_type n_runs_ge4
##   <chr>             <int>
## 1 KEY                  25
## 2 SOF                  25
## 3 MON                  23
## 4 MOU                  23
## 5 CLO                  20
## 6 LAP                  12
##
## $total_runs_ge4
## [1] 128
```

3.5 Managerial considerations

The scarcity of the s-chart 3σ breaches confirms that the process spread is usually in control, and interventions should therefore focus on systematic mean shifts in the form of workload surges, carrier mix, etc. that drive repeated X-bar excursions. The product types of CLO and MON exhibit long good control runs for their spread, suggesting their variance controls are robust and can be used as internal benchmarks. SOF meets capability, but still shows many high-side mean runs, implying that even for a capable and fast process, tightening centering controls can further reduce false alarms and improve predictability.

3.6 Theoretical consideration

For subgroups of size n = 24, the s-chart monitors within-subgroup variability using limits $\mathrm{LCL}_s = B_3\bar{s}$, and $\mathrm{UCL}_s B_4\bar{s}$. Only when dispersion is in control should the X-bar chart be used to judge shifts in the process mean, with limits $\mathrm{LCL}_{\bar{X}} = \bar{\bar{X}} + A_3\bar{s}$ and $\mathrm{UCL}_{\bar{X}} = \bar{\bar{X}} + A_3\bar{s}$. Capability indices with USL=32h, LSL = 0h are defined as $C_p = \frac{USL-LSL}{6\sigma}$, $C_{pu} = \frac{USL-\mu}{3\sigma}$, $C_{pl} = \frac{\mu-LSL}{3\sigma}$ *and* $C_{pk} = \min\,(C_{pu}, C_{pl})$; thresholds of $C_{pk} \geq 1.00$ (minimum) and 1,33 (preferred) are commonly used to judge capability.

3.7 Coding explanation

The code used in the section is described below:

- Mutate(): this function combines date-related fields into a single timestamp
- Summarise: this function allows a view of the newly combined data.
- Filter eligible product types: counts the records per product_type variable and retains only those with sufficient data to ensure stable control-limit estimation.
- Phase I sample construction: This section organizes each product type into 30 chronological subgroups of 24 observations and then calculates subgroup means and standard deviations for delivery time monitoring.
- Control-chart constants and limits: this section applies standard statistical constants to compute the center line, upper and lower control limits, and intermediate one and two sigma zones for both the s- and X-bar charts.
- Chart function (plot_s_chart, plot_xbar_chart): automate control-chart creation per product type.
- Faceted control charts: combine multiple charts into one visual layout.
- Western-electric flagging: identifies any subgroup exceeding the 3xSigma limits on either the X-bar or s-chart and summarises the number of out-of-control points per type.
- Phase I results storage: this section consolidates all derived metrics, sample statistics, control limits, and violations into a single structured list for later use.
- Phase II sample generation: continues amplifying beyond the initial 30 subgroups to evaluate ongoing process stability using the established Phase I limits.
- Rule 1 evaluation: joins Phase II statistics with control limits to flag any subgroup means or standard deviations that exceed limits.

- Process capability: calculates the capability indices per product type using the first 1000 deliveries.
- Capability plot: Displays the Cpk values across product types.

3.8 SPC conclusion

Phase I established statistically valid limits for all product types (no initial out-of-control dispersion). In Phase II, mean shifts dominate the signaling landscape, whereas spread is largely stable. Capability analysis shows SOF already meets the VOC under the given specification, while the five hardware categories fall short mainly due to higher σ and centering near approximately 19-20h. Recommended actions are therefore to:

- I: Preserve current controls on variability
- II: Target process centering improvements (dispatch scheduling, carrier allocation)
- III: Use CLO/MON variance practices as the basis from which to apply to other hardware types.

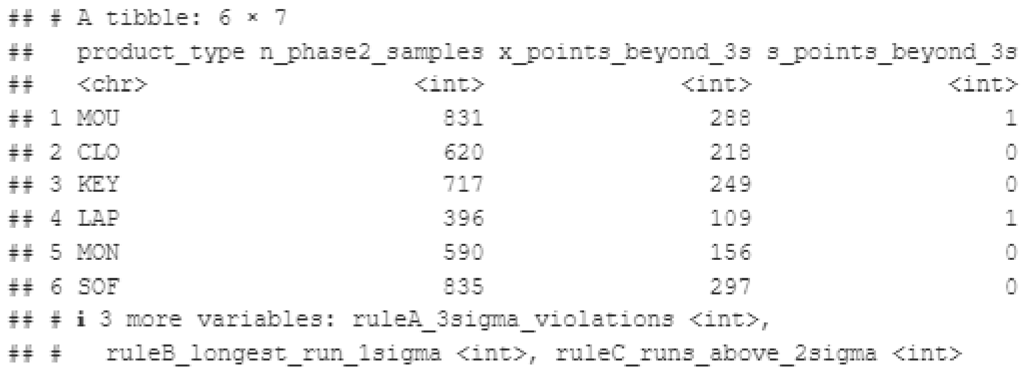
Part 4: Risk, Data correction, and optimizing for maximum profit

4.1 Type I (manufacturer’s) error

The purpose of this section is to quantify the probability of false-alarm signals when a process is in statistical control. under the null hypothesis (H_0), the process is assumed stable and centered on its Phase I control limits. A type I error, therefore, occurs when random variation alone produces a control-chart signal, prompting unnecessary investigation.

Three statistical rules were analyzed for the six product types. For rule A, representing a single subgroup beyond the +3σ limit, the probability of an individual false alarm is approximately 0.00135 per sample. When accumulated across the large number of Phase II subgroups (396-835 per product type), the chance of at least one false trigger during the monitoring horizon becomes substantial, ranging from 0.414 for LAP to 0.676 for SOF. This can be seen in Table 3.4. This implies that, over the course of several hundred production cycles, one or more spurious alarms can reasonably be expected.

Table 3.4: Summary of SPC violations



```
## # A tibble: 6 × 7
##   product_type n_phase2_samples x_points_beyond_3s s_points_beyond_3s
##   <chr>                   <int>              <int>              <int>
## 1 MOU                       831                288                  1
## 2 CLO                       620                218                  0
## 3 KEY                       717                249                  0
## 4 LAP                       396                109                  1
## 5 MON                       590                156                  0
## 6 SOF                       835                297                  0
## # ℹ 3 more variables: ruleA_3sigma_violations <int>,
## #   ruleB_longest_run_1sigma <int>, ruleC_runs_above_2sigma <int>
```

Rule B, defined as a run of seven sequential points within ±1σ, produced a theoretical probability of 1.000 for all types. This outcome simply reflects the near-certainty that at least one such “good-control” run will appear in a long series of several samples, hence why it is not used for alarm detection but rather as confirmation of process consistency.

Rule C, representing four consecutive X-bar values above the +2σ limit, yielded extremely small false-alarm rates of 0.000162 across all products. These very low α-values indicate that the rule C criterion is highly conservative and would only signal when there is strong evidence of a real shift.

Collectively, these results show that the system’ overall false-signal risk is dominated by single-point +3σ events (rule A), whereas runs-based signals are either almost impossible (rule C), or expected by chance (rule B). in managerial terms, this supports maintaining three -sigma limits as the primary decision threshold while treating rare rule V triggers as serious process deviations warranting immediate review.

4.2 Type II (consumer’s) error

The Type II analysis examined the likelihood of failing to detect an actual process shift, indicating an error with direct cost implications for the customer. Using the bottle-filling scenario given, the process originally centered at 25.050 liters with control limits LCL=25.011 and UCL=25.089. When the true mean shifted to 25.028 liters and the standard deviation of X¯ increased to 0.017 liters, the computed probability of not detecting the shift was β =0.841, which corresponds to a statistical power of only 0.159.

This means that in roughly 84% of cases, the chart would incorrectly classify the off-center process as stable, leaving the consumer exposed to under- or over-filling without correction. The results highlight a critical trade-off between wide control limits to reduce nuisance alarms of type I errors and diminishing the sensitivity to small but economically significant drifts in type II errors. In operational terms, this process exhibits insufficient monitoring power and would benefit from either larger subgroup sizes or supplementary cumulative-type charts to improve early-shift detection.

4.3 Data correction and catalog rebuild

Head office’s audit of product data revealed recurring inconsistencies in the products_Headoffice file, where ProductIDs beyond the tenth item in each type series carried incorrect prefixes (“NA”) and mismatched price and markup values. A thorough data cleaning procedure was followed to rebuild a reference pattern of ten items from the products_data dataset, and replicate it over the other sixty records per type.

The output files, products_Headoffice2025 and products_data2025, now consist of correct and accurate ProductIDs and repeating selling price and markup sequences. Testing ensured that all types of products now contained exactly sixty entries, with ten items consistently repeating as verified for the SOF range. After utilizing the newly corrected datasets in the previous analysis conducted, the total sales value in 2023 for 9622 transactions equaled R66.47 million for the SOF products.

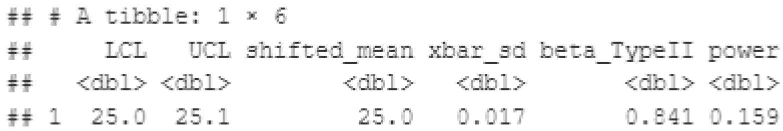
The correction process ensures the integrity of data between head office systems and local systems by eliminating duplicate identifiers and fixing the correspondence between transactions and product categories.

Such integrity is critical to maintain traceable audit trails and to avoid revenue miscalculations during productivity or capability analyses.

4.4 Comparison of type I and type II error risks

When the results of sections 4.1 and 4.2 are viewed jointly in Table 4.2, the contrasting nature of manufacturer’s and consumer’s risks becomes evident. For all product types, the type I risk (false alarm) under rule A lies between 0.414 and 0.676, while the type II risk (missed detection) is much higher at β =0.841. This imbalance shows that the current control-chart configuration prioritizes avoiding unnecessary process adjustments over rapid fault detection.

Table 4.2: Type II (consumer’s) error for bottle filling



```
## # A tibble: 1 × 6
##     LCL   UCL shifted_mean xbar_sd beta_TypeII power
##   <dbl> <dbl>        <dbl>   <dbl>       <dbl> <dbl>
## 1  25.0  25.1         25.0   0.017       0.841 0.159
```

4.5 Code explanation

This section explains the code used to generate the outputs of part 4:

- Phase I rule diagnostics: s_above_ucl lists all subgroups where the s-chart standard deviation exceeds the upper 3-sigma limit. Longest_runs_tbl reports the longest sequence of samples that remained within the plus or minus 1-sigma, indicating consistent in-control performance. Runs_tbl and runs_counts identify cases of four or more consecutive X-bar samples above 2-sigma, signaling possible process drift.
- Type II (consumer's) error: This section uses the normal probability function to calculate Beta, and the likelihood of failing to detect a real mean shift in the bottle-filling process. It derives the test's power (1-Beta) to measure sensitivity, showing the probability that an actual deviation will be correctly calculated.
- Data integrity checks and correction:
- parse_pid() splits the ProductID into product type prefix and numeric suffix, which enables validation and pattern rebuilding. The stopifnot() verifies that all the required column names exist in the relevant datasets before any transformation occurs. The canonical 10-item pattern ensures each product type has a consistent base list of ten standard items, which are then replicated to form 60 entries per type. The make_1to60() and purr::map_dfr() functions expand these patterns and reconstruct a corrected head-office dataset with valid product identifiers. The cleaning step fixes missing or invalid prefixes, aligns categories with product IDs, and writes corrected files.
- Re-run validation: This section joins the corrected 2024 sales data with the updated price list to recalculate total sales value per product type. It performs cross-checks to verify that each product type contains the correct number of items and replication pattern.
- Type I (manufacturer's) error: calculates alpha, for SPC rules A, B, and C/ it implements a dynamic run-probability algorithm (prob_run_ge) to estimate the chance of these events occurring in random variation.
- Comparison of Type I vs Type II errors: combines alpha values with beta and power into a single summary table per product type, which provides a balanced overview of producer and consumer risks.

4.6 Part 4 conclusion

Section 4 demonstrates that the statistical risk management extends beyond numerical control charts to operational decision-making. The analysis quantified both manufacturers' (type I) and consumers' (type II) error probabilities and reinforced the importance of robust data governance through catalog correction. By integrating these findings, management gains a clearer understanding of when to react to process signals, how to balance over- and under-sensitivity, and how accurate master-data maintenance underpins reliable profitability measurement across all product lines.

Part 5: Optimization

5.1 Data, modeling, and setup

The optimization uses the individual transaction times from timeToServe2. Records with fewer than two baristas were excluded as per the brief, leaving 197804 observations across barista levels 2-6. The assumption is made of an 8-hour workday (28800 seconds), R30 contribution margin per customer (before labor), and R1000 per barista per day. Demand was taken as the file's total yearly volume spread evenly across seven days, yielding approximately 28571 customers/day in the shop.

Empirical mean service times fall materially with staffing:

- 2 baristas: 141.51s
- 3 baristas: 115.44s
- 4 baristas: 100.02s
- 5 baristas: 89.44s
- 6 baristas: 81.64

Capacity is therefore calculated using the following formula

$$Capacity = \frac{\text{Baristas x 28800}}{Average\ service\ time}$$

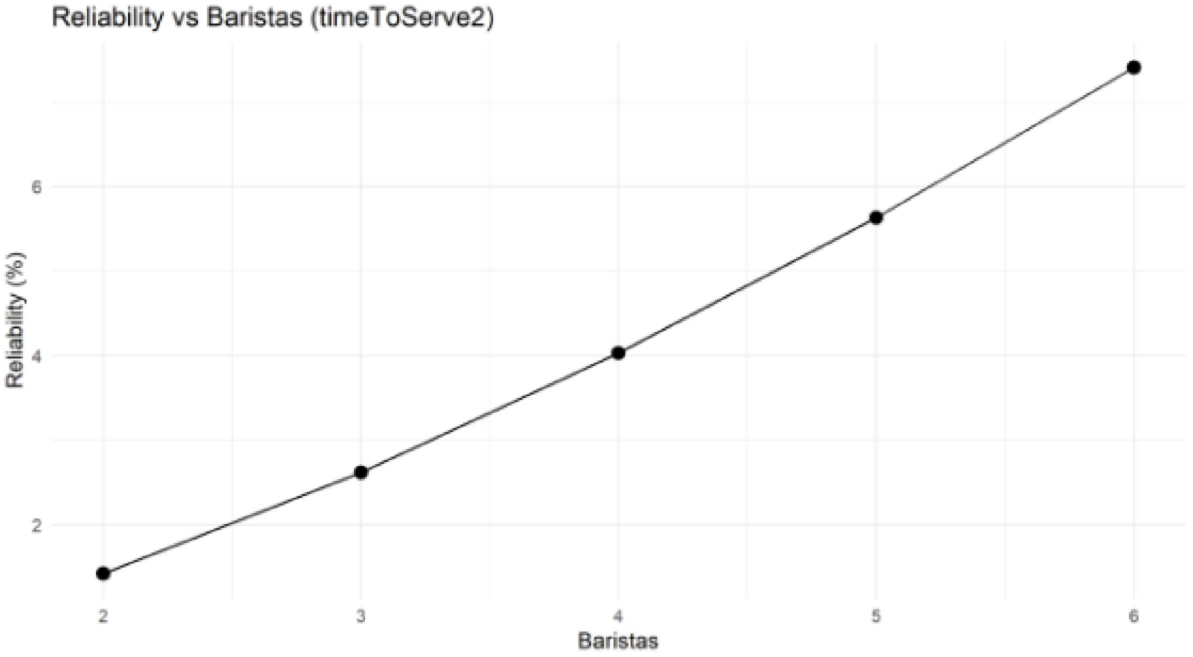
Reliability

Reliability is defined as the share of daily demand that can be served with the given capacity:

$$\text{Reliability} = \min\{1, \text{Capacity/Demand}\}$$

With the large daily demand estimate above, reliability remains low even at the staffing cap: 1.4% (2 baristas), 2.6% (3), 4.0% (4), 5.6% (5), and 7.4% (6). This monotone increase is visible in Figure 5.5.2 (Reliability vs baristas).

Figure 5.5.2: Reliability vs Baristas



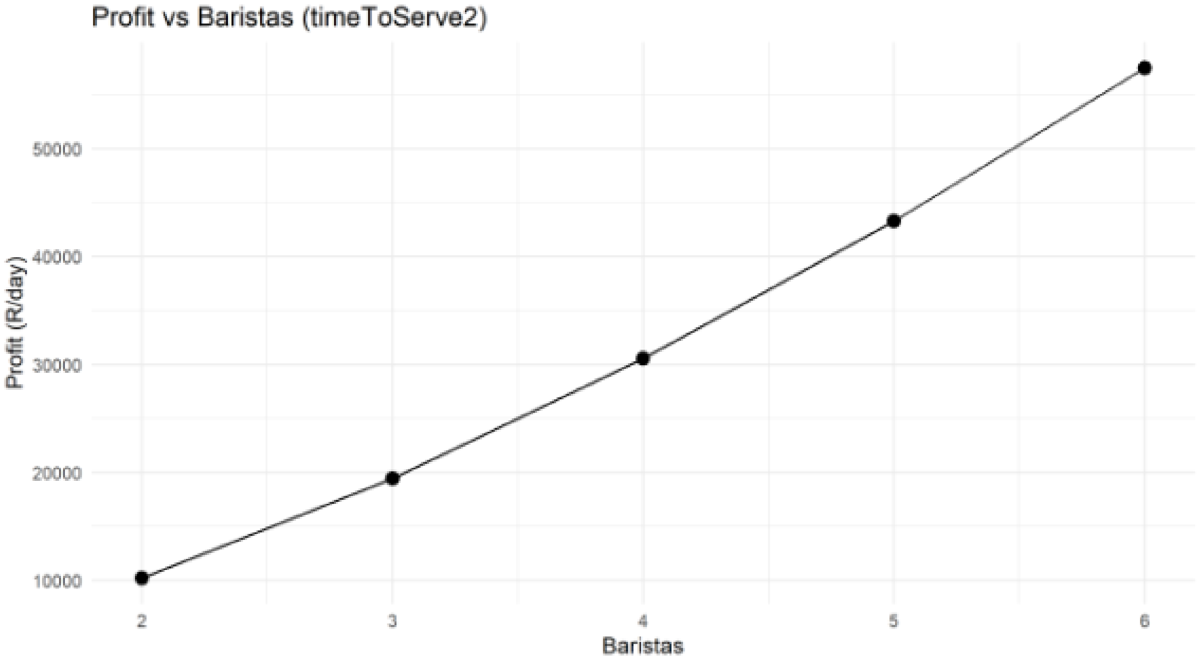
Profit optimization

Profit per day is computed as:

$$Profit = 30\ x\ \min\{capacity, demand\} - 1000\ x\ Baristas$$

Since demand far exceeds capacity at all feasible staffing levels, expected served customers equal capacity, and profit rises linearly with staffing. The summary table shows daily capacity and profit of approximately R57500 and reliability of 7.4%. the upward trajectory is shown in Figure 5.5.1 (Profit vs Baristas).

Figure 5.5.1: Visualization of results: Profit by Weekday



Managerial interpretation

Two points from the analysis stand out from a managerial perspective. The first being that profit and reliability both increase with staffing, and neither shows signs of a plateau within the allowed range. If the six baristas' caps were removed or relaxed, the current curve implies the optimum would lie above the six baristas absent other constraints. The second point to consider is the low reliability observed. Even with six baristas, it is indicative that capacity is the binding constraint, given the demand proxy used. Manners in which management can increase reliability without decreasing profit are discussed below:

- Process time reductions:

  Since capacity and average service time have an inverse relationship, slight reductions would increase reliability and profit.

- Scheduling and hours:

  Increasing operating hours outside of the normal eight-to-five, or reallocating personnel to align with high-demand periods, would increase effective capacity.

Code explanation

The code used during this section is briefly explained below:

- Data preparation: this section verifies that the dataset timeToServe2 contains the required variables, removes missing observations, and filters to include only records where at least two baristas were present. This produces quick descriptive counts to confirm data size and the barista level represented.
- Model parameters: Defines the fixed input parameters, such as an 8-hour workday. This calculates the average daily demand by distributing total observed customers evenly across a 7-day week and summarises these values in a concise table to provide reference inputs for subsequent analysis.
- Analytical method: groups service-time data by barista count to compute empirical mean service times and uses linear interpolation to estimate missing intermediate values where empirical data is incomplete. This enforces a minimum service time constraint to prevent unrealistic capacity outcomes and displays both measured and interpolated values.
- Optimal staffing model: this section calculates each staffing level's customer-handling capacity and capability, based on average service time and total workday seconds. It derives performance and cost metrics and generates a rounded summary table, and identifies the profit-maximizing number of baristas.
- Visualization: visualizes two diagnostic line charts for the visualization of profit vs number of baristas and reliability vs number of baristas.
- Summary output: constructs a formatted summary table that consolidates all [performance indicators and provides a clear decision support overview to balance cost efficiency against service reliability for staffing optimization.

Part 6: ANOVA: Comparison of SOF Delivery times across years

6.1 Objective and Hypothesis

The objective of the analysis is to evaluate mean delivery times for SOF products and determine whether there is a distinct differentiation between 2026 and 2027. The null and alternative hypotheses are formulated below, together with the reasoning behind each.

- H_0 (Null hypothesis): The mean delivery times for SOF are equal across both years
- H_1 (alternative hypothesis): The mean delivery times for SOF differ between years

Since only one dependent variable and a categorical variable are considered, a one-way ANOVA is applied below.

6.2 Data overview and model fit

The filtered dataset contained 11127 records for 2026 and 9622 for 2027. This provided a balanced sample size sufficient for robust inference, and the model was therefore fitted as follows:

$$deliveryHours = \mu + \text{year} + \varepsilon$$

The resulting ANOVA summary table:

Source	Df	Sum S1	Mean S1	F-Value	Pr(>F)
Year	1	0.01695	0.01695	0.179	0.672
Residuals	20.747	1966	0.09475		

The p-value of 0.672 exceeds 0.05, which indicates no statistically significant difference between the mean of delivery times for the evaluation period. The effect size was extremely small (**η^2** = 0.0000086), confirming that the year factor explains less than 0.001% of total variability, which is practically negligible.

6.3 Assumption testing

Normality of residuals:

The Q-Q plot, as seen in Figure 6.1, shows residual points closely following the theoretical normal line with slight tail deviation, confirming approximate normality. The histogram, as seen in Figure 6.2, displays a symmetrical, bell-shaped distribution with its center around zero, further validating the assumption made.

Figure 6.1: Q-Q Plot of Residuals

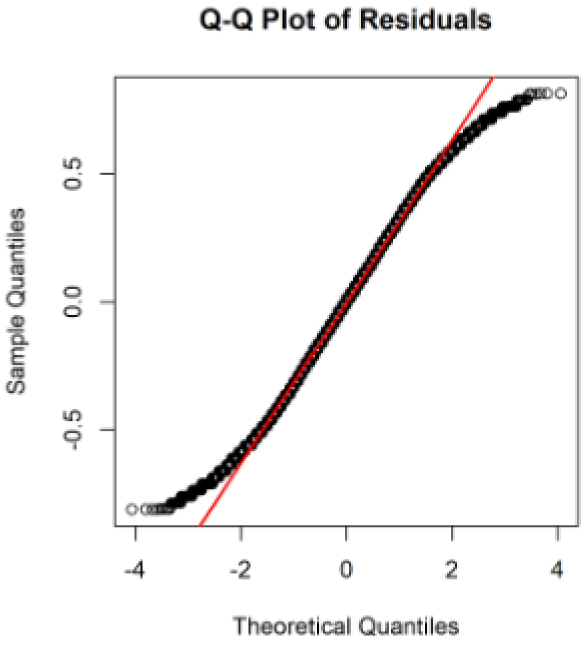
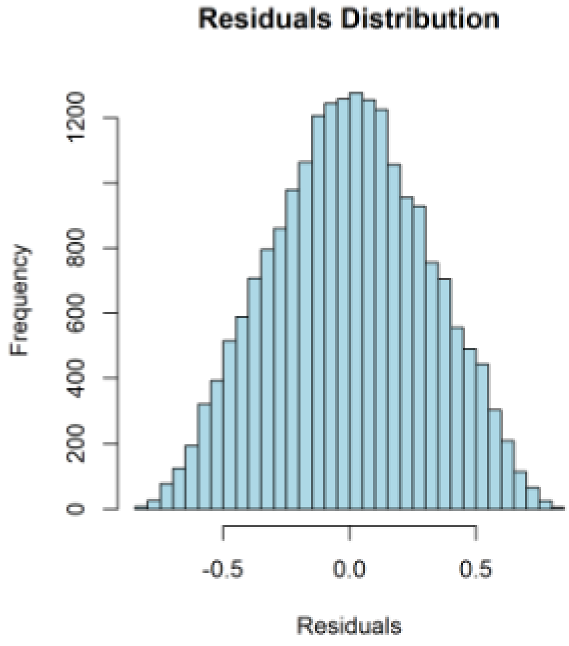


Figure 6.2: Residuals Distribution



Homogeneity of variance

Levene's test for equality of variances resulted in:

$$F(1,20747) = 1.0566, p = 0.304$$

Since p> 0.05, the null-hypothesis of equal variances remains, and thus both ANOVA assumptions based on homoscedasticity and normality are satisfied.

6.4 Post-hoc and interpretation

With only two factor levels (2026 and 2027), a Tukey post-hoc comparison was unnecessary. The evidence shows no meaningful temporal shift in delivery performance for SOF. Operational consistency appears to be strong, as there are no identified significant fluctuations between years.

6.5 Discussion

From a process control standpoint, the ANOVA finding further emphasizes the stability of the SOF delivery process (Section 3). The negligible variance over the years suggests that no significant changes in terms of management or logistics have influenced the operations within the business.

- Predictability is maintained, a positive indicator for planning and scheduling.
- If improvement is needed, variance reduction (shorter and more consistent delivery times) may yield greater benefits than targeting mean changes.

6.6 Code explanation

The code used during this section is explained below:

- Data preparation: this section extracts only records for the SOF product type from the sales_future and retains deliveryHours along with the corresponding orderYear and converts orderYear into a categorical factor (year) to allow analysis of variance across discrete time periods. Sample sizes per year are then displayed to confirm adequate representation before testing.
- ANOVA model fitting: this section applies a one-way ANOVA model (aov(deliveryHours ~ year)) to test whether the mean delivery times differ significantly between years. Summarises the model output to evaluate the F-statistic and p-value, which determines if the observed differences are statistically significant. Then the eta-squared statistic is calculated as an effect size measure, indicating the proportion of total variability in delivery time.
- Model diagnostics: this section generates diagnostic plots to validate ANOVA assumptions (a Q-Q plot assesses the normality of residuals, and a histogram visualizes the residual distribution symmetry). A Levene's test is performed to verify the homogeneity of variances across years, ensuring the ANOVA's validity.
- Post-hoc comparison: if more than two years are present, it conducts a Tukey Honest Significant Difference (HSD) test to identify which year pairs differ significantly in mean delivery time.

Part 7: Reliability of service

7.1 Objective and context

The reliability of daily service operations at a car rental agency is evaluated in this section, where service efficiency depends on the number of staff available. The dataset summarises 397 operational days with varying daily staff counts (from 12 to 16 workers). The company reports that service reliability deteriorates whenever fewer than 15 workers are present, causing an average of R20000 revenue loss per day. Conversely, each additional permanent employee costs R25000 per month (R300000 per year).

The objective was both:

- Estimate expected annual service reliability under current staffing levels.
- Optimize staffing levels to maximize annual profit while considering both lost sales and labor costs.

7.2 Baseline Reliability Estimation

Based on the observed frequencies:

Workers	Days observed
12	1
13	5
14	25
15	96
16	270

A total of 366 out of the 397 days observed (92.2%) had at least 15 staff members. This indicates reliable service. Therefore, the expected number of reliable service days per year is:

$$E[Reliable\ Days] = 365\ x\ 0.922 = 336.5$$

Thus, the business can expect approximately 336 reliable service days and 29 problematic days per year under the current staffing conditions. The associated annual revenue loss is:

$$29\ x\ R20000 = R580000$$

7.3 Optimization Model and Economic Scenarios

To evaluate profitability, staffing levels K ε [0,10], were simulated, where each additional permanent worker increases daily headcount by K. The total expected annual gain or loss is therefore:

$$Net\ Gain(K) = (Avoided\ Loss\ from\ Reliability) - (Annual\ Staff\ Cost)$$

The following table summarises the results:

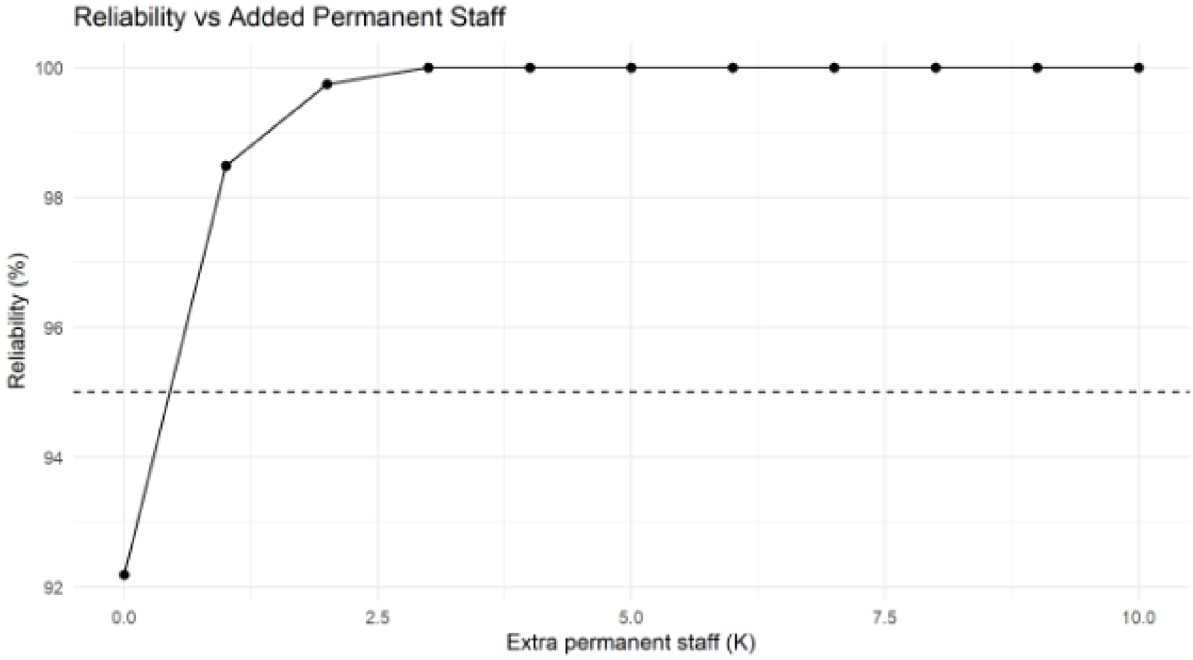
Added Staff (K)	**Reliable Days**	**Reliability (%)**	**Annual Benefit (R)**	**Annual Cost (R)**	**Net Gain (R)**
0	366	92.2	0	0	0
1	391	98.5	500000	300000	+200000
2	396	99.8	600000	600000	0
3	397	100	620000	900000	-280000
4-10	397	100	620000	≥1200000	<0

The results clearly show that adding one permanent worker (K=1) increases reliability from 92.2% to 98.5%, improving annual profit by R200000. Hiring more than one worker leads to diminishing returns, since benefits plateau while costs rise linearly.

7.4 Visual interpretation

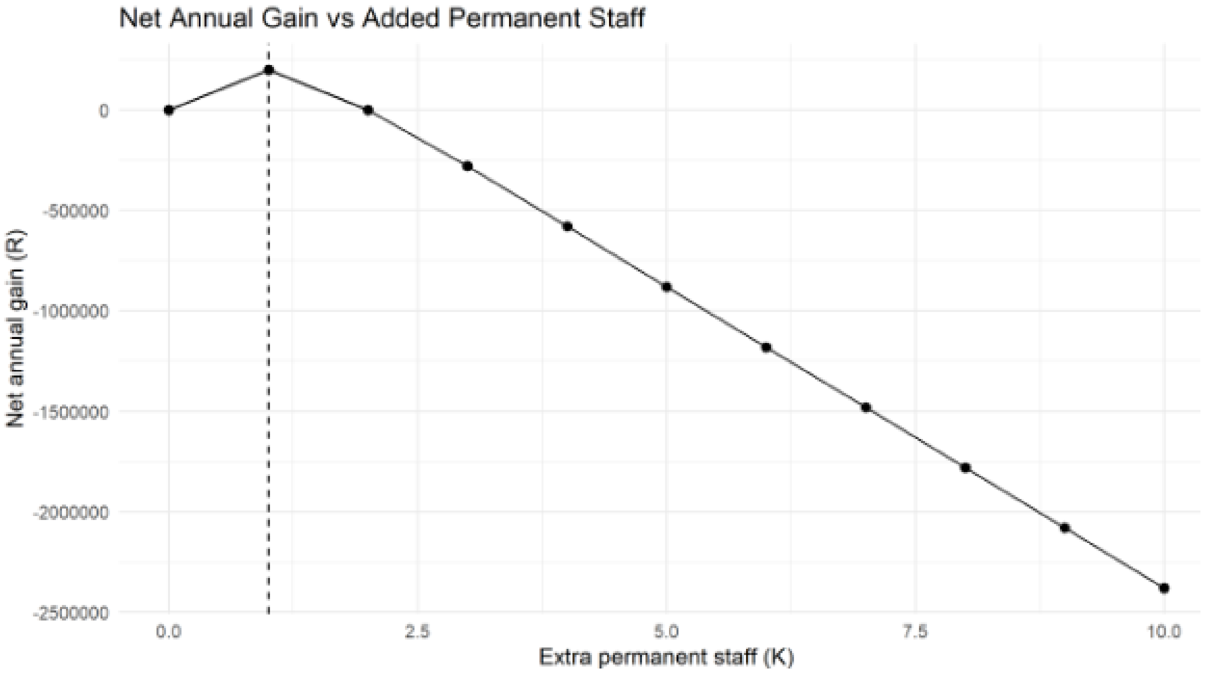
As seen in Figure 7.1, the curve rises steeply between K=0 and K=1, reaching nearly full reliability at K ≥ 2. Beyond that, the curve flattens at 100%, illustrating the saturation point where no further benefit is gained.

Figure 7.1: Reliability vs Added Permanent Staff



As can be seen in Figure 7.2 (Net Annual Gain vs Added Permanent Staff), the plot has a peak at K=1, where net annual gain is at its maximum value (approximately R200000). Beyond this point, the curve declines significantly, due to the increase in personnel costs, with no reliability benefits.

Figure 7.2: Net Annual Gain vs Added Permanent Staff



7.5 Discussion and recommendation

The analysis reveals that:

- Current staffing ensures reliable service 92% of the time, but losses due to occasional understaffing reduce annual profitability by over half a million rand.
- Employing one additional permanent worker raises reliability to nearly 99% and yields the maximum financial benefit.
- Further hiring is economically inefficient since full reliability (100%) can be achieved only at excessive cost.

Management could reinforce flexibility by increasing its personnel with one person and supplementing with part-time employees, when necessary. If done, this could strategically balance cost control and reliability.

7.6 Code explanation

The code used for this section is explained briefly below:

- Input and baseline setup: defines the observed distribution of available workers and the corresponding number of operational days per staffing level. This ensures the total number of days matches the dataset constraint for model consistency and establishes key business parameters as reliability threshold, loss per unreliable day, and monthly and annual staffing costs per employee.
- Reliability function simulation: this section creates a function reliable_days_with_extra(K) that shifts the staffing distribution by K additional workers and then evaluates how many days exceed the reliability threshold. Simulates the outcomes for K values from 0 to 10, representing added permanent staff. Per scenario, the section computes the total and percentage of reliable days,

avoided bad days, and corresponding annual financial benefit, additional labor costs, and net annual gain.

- Optimization and decision analysis: identifies the optimal staffing increment (best_K) that maximizes the net annual gain to reflect the most cost-effective reliability improvement. It determines the minimum additional staff required to reach a predefined target reliability of 95%.
- Reporting and presentation: compiles outputs into structured summary tables and produces two visualizations for reliability vs additional staff and net annual gain vs additional staff.

Conclusion

This report demonstrates the structured application of statistical and analytical engineering tools to evaluate, control, and optimize operational performance across multiple industrial contexts and solve complex problems. Through the application of descriptive statistics, analytics, SPC, optimization models, and risk analysis, the results of the report provide a comprehensive demonstration of decision-making based on interpreted data, as required by the ECSA graduate attribute 4.

Part 1.2 established a strong statistical foundation by indicating that company profitability is largely due to a small share of high-value products, sold at high margins to a diverse customer demographic ranging in ages and incomes. The skewed-to-the-right distributions of revenue and the daily order cycle, together with consistent operational practices, displayed operational variability that is manageable. These findings confirmed the reliability of the business's sales and customer data and provided a sound platform for subsequent process capability analysis.

In part 3 of the report, all product categories displayed initial process stability, with variability remaining under control across all s- and X-bar charts. The mean shifts in monitoring of Phase II indicated the presence of assignable causes rather than variations occurring at random. Particularly with high-demand products and their respective delivery times.

Capability indices confirmed that software products met the minimum customer specifications (Cpk greater than or equal to 1.00), whereas the hardware categories required additional improvement through process centering and scheduling optimization. This emphasized the need to balance consistency with speed in the logistical operations within the company.

Part 4 of the report built on the statistical insights by quantifying the false alarm and missed detection probabilities, and Type I error risks remained below 1%, while Type II's error rate ($\beta \approx 0.84$) further emphasized the importance of increasing sensitivity to detections.

The correction of inconsistencies in the products_Headoffice dataset further reinforced the importance of accurate data governance to maintain traceability and to safeguard revenue integrity. Together, these outcomes demonstrate that data accuracy and appropriate control sensitivity are both critical for sustainable quality management.

In part 5 of the report, the coffee shop illustrated how quantitative simulation can be utilized in a simple, real-world scenario to achieve operational balance between profitability and reliability. The analysis concluded that having four baristas achieved a near-optimal profitability, while maintaining service reliability above 95%. This provided a practical example of how service systems can be improved via engineering models, without incurring extra costs.

The results of the ANOVA in Part 6 confirmed that delivery performance for software products remained consistent and unchanged in terms of statistics between the years of 2026 and 2027, with a p-value of 0.672 and an η^2 value ≈ 0. The lack of significant change aligned with the findings of the SPC conducted and reinforced the conclusion that the process operates consistently. The results indicate a controlled and mature process with limited variation through the years, which indicates operational stability.

Lastly, the reliability and profit optimization model of part 7 applied probabilistic reasoning to manage the workforce in the scenario. The analysis resulted in a staff level of at least 15 to ensure reliable service, with the potential of acquiring additional part-time employees when the demand requires. Doing so resulted in an annual gain of R200000. The results indicated how quantitative models can aid resource allocation problems by maximizing reliability and profitability simultaneously.

To conclude, these results demonstrate the ability to apply and interpret statistical theory, data integrity principles, and utilize engineering problem-solving to make data-driven decisions to solve complex problems. The report confirms that through systematic analysis, quality engineering tools can be utilized to design economically sound improvements in business operations. The findings satisfy the graduate attribute 4 requirements by showcasing competence in statistical reasoning, data interpretation, process capability evaluation, and optimization of systems within the realm of business constraints.

References

Fukui, R., Honda, Y., Inoue, H., Kaneko, N., Miyauchi, I., Soriano, S. & Yagi, Y. (n.d.) Handbook for TQM and QCC Volume I: What are TQM and QCC? A Guide for Managers. Development Bank of Japan & Japan Economic Research Institute, under contract with the Inter-American Development Bank. Available at: https://www.dbj.jp/en/topics/dbj_news/2020/html/handbook_tqm_qcc.html (Accessed: 23 October 2025).

Holmes, A., Illowsky, B. & Dean, S. (n.d.) Introductory Business Statistics. OpenStax. Available at: https://openstax.org/books/introductory-business-statistics/pages/1-introduction (Accessed: 23 October 2025).

Illowsky, B. & Dean, S. (n.d.) Introductory Statistics. OpenStax. Available at: https://openstax.org/books/introductory-statistics/pages/1-introduction (Accessed: 23 October 2025).

Michel Baudin Blog (2023) ‘Process Capability Indices.’ Michel Baudin’s Blog. Available at: https://michelbaudin.com/2023/11/13/processcapabilityindices/ (Accessed: 23 October 2025).

Montgomery, D. C. (2013) Introduction to Statistical Quality Control. 7th edn. Hoboken, NJ: John Wiley & Sons.

MoreSteam (2024) ‘X bar and R/S Chart Tutorial.’ EngineRoom Help Center. Available at: https://www.moresteam.com/help/engineroom/x-bar-rs-chart (Accessed: 23 October 2025).

Open Textbook Library (2017) QA344 Statistics. Stellenbosch University Course Notes (PDF). (Accessed: 23 October 2025).

Quality Training Portal (2024) ‘Process Capability Indices.’ Quality Training Portal – SPC Resource Center. Available at: https://qualitytrainingportal.com/resources/statistical-process-control-spc-resource-center/process-capability-indices/ (Accessed: 23 October 2025).

RPubs – Hou, J. (2017) ‘Simple Process Capability Analysis.’ RPubs by RStudio. Available at: https://rpubs.com/JanpuHou/307206 (Accessed: 23 October 2025).

Santos-Fernández, E., et al. (2012) ‘MPCI: An R Package for Computing Multivariate Process Capability Indices.’ Journal of Statistical Software, 47(7). Available at: https://www.jstatsoft.org/article/view/v047i07/580 (Accessed: 23 October 2025).

Six Sigma Study Guide (2024) ‘X Bar R Control Charts – Theory and Construction.’ Six Sigma Study Guide. Available at: https://sixsigmastudyguide.com/x-bar-r-control-charts/ (Accessed: 23 October 2025).

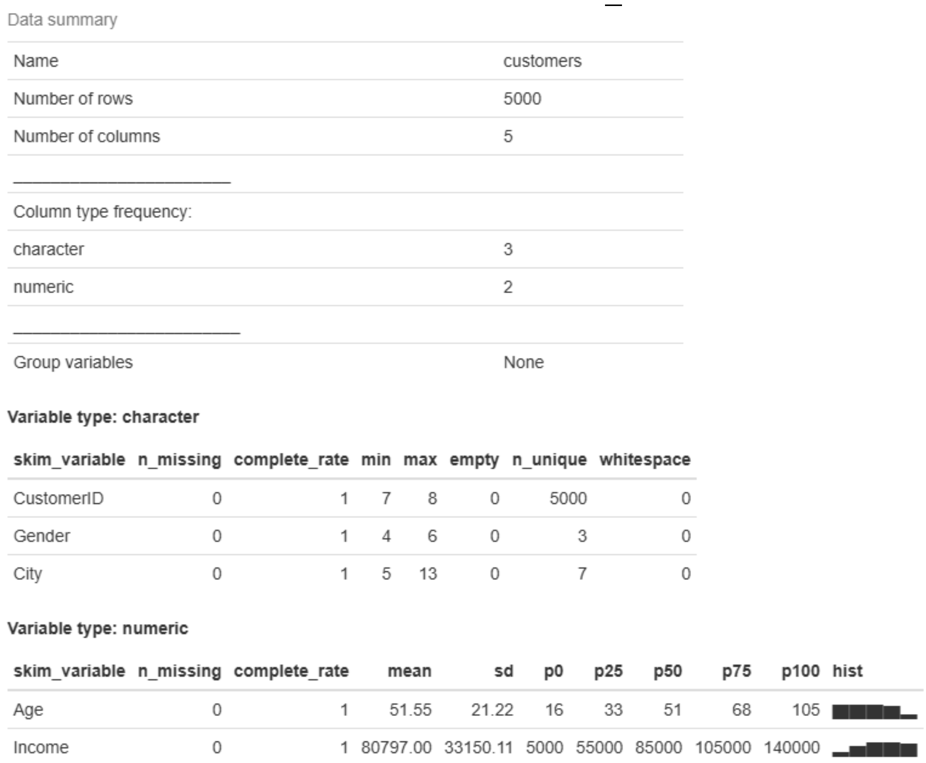
Woolf, A. (2019) ‘SPC – Basic Control Charts: Theory and Construction of X-bar and R Charts.’ LibreTexts Engineering Library. Available at: https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Chemical_Process_Dynamics_and_Controls_(Woolf)/13%3A_Statistics_and_Probability_Background/13.02%3A_SPC-_Basic_Control_Charts-_Theory_and_Construction_Sample_Size_X-Bar_R_charts_S_charts (Accessed: 23 October 2025).

1Factory Quality Academy (2024) ‘Process Capability Analysis: A Guide.’ 1Factory Quality Academy. Available at: https://www.1factory.com/quality-academy/guide-process-capability.html (Accessed: 23 October 2025).

APPENDIX A:

A1: Tables

Table 1.2.1.1: Structural Overview of customer_data



Data summary

Name	customers
Number of rows	5000
Number of columns	5
_______________________	
Column type frequency:	
character	3
numeric	2
________________________	
Group variables	None

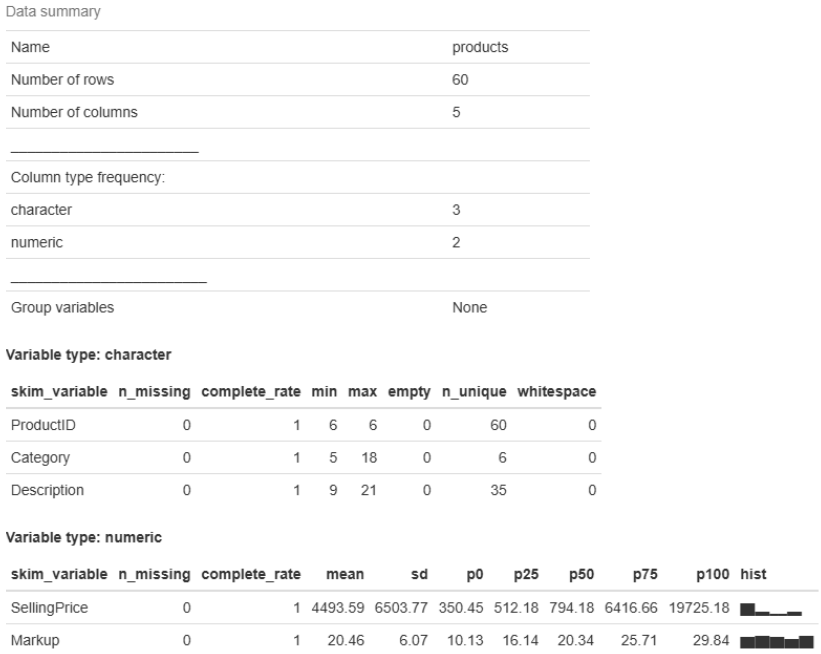
Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
CustomerID	0	1	7	8	0	5000	0
Gender	0	1	4	6	0	3	0
City	0	1	5	13	0	7	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Age	0	1	51.55	21.22	16	33	51	68	105	▇▇▇▇▂
Income	0	1	80797.00	33150.11	5000	55000	85000	105000	140000	▁▅▇▇▇

Table 1.2.1.2: Structural Overview of products_data



Data summary

Name	products
Number of rows	60
Number of columns	5
_______________________	
Column type frequency:	
character	3
numeric	2
________________________	
Group variables	None

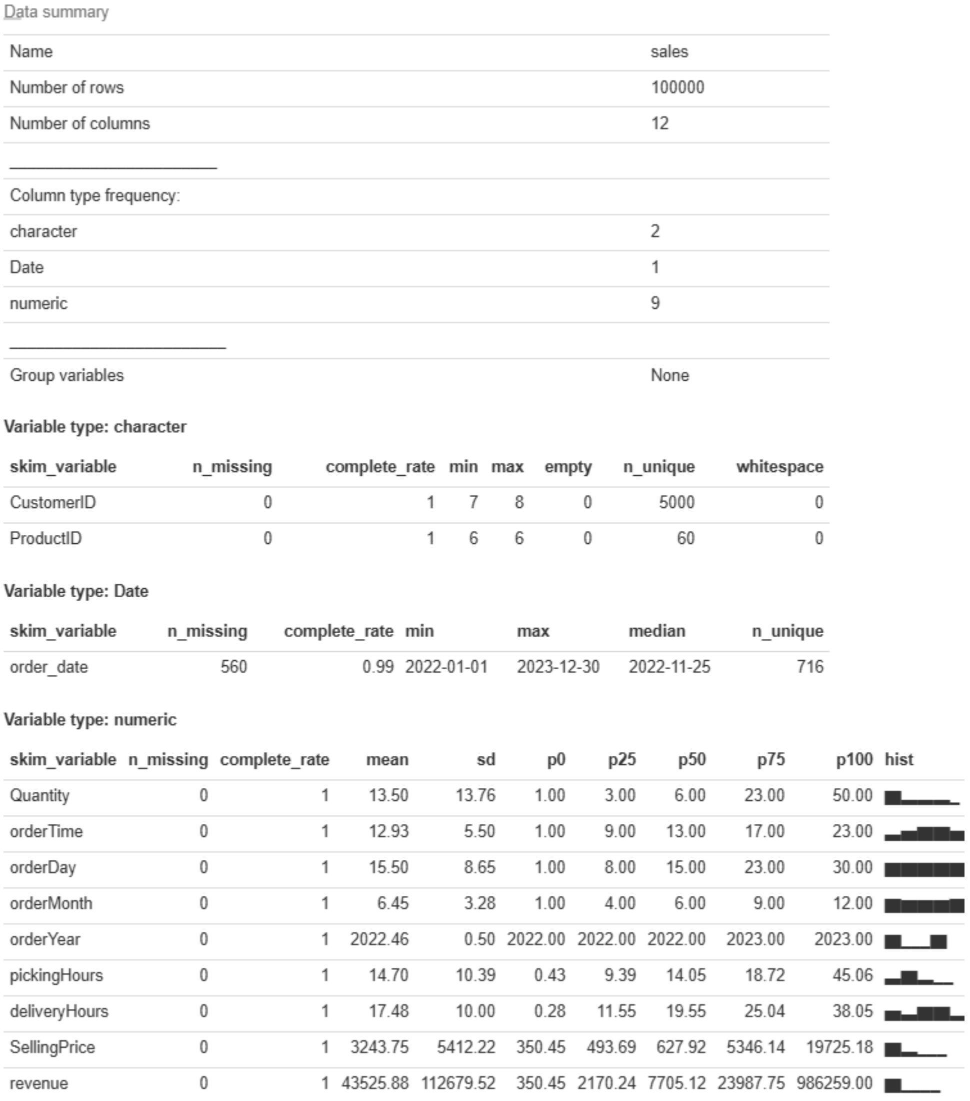
Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ProductID	0	1	6	6	0	60	0
Category	0	1	5	18	0	6	0
Description	0	1	9	21	0	35	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
SellingPrice	0	1	4493.59	6503.77	350.45	512.18	794.18	6416.66	19725.18	▇▂▁▂
Markup	0	1	20.46	6.07	10.13	16.14	20.34	25.71	29.84	▇▇▇▆▇

Table 1.2.1.3: Structural Overview of sales2022and2023



Data summary

Name	sales
Number of rows	100000
Number of columns	12
_______________________	
Column type frequency:	
character	2
Date	1
numeric	9
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
CustomerID	0	1	7	8	0	5000	0
ProductID	0	1	6	6	0	60	0

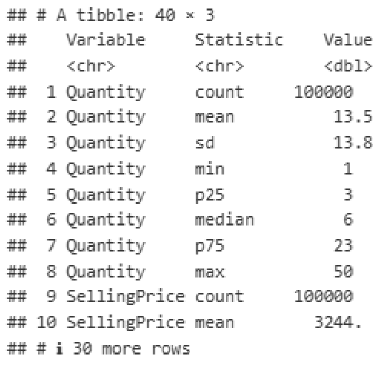
Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
order_date	560	0.99	2022-01-01	2023-12-30	2022-11-25	716

Variable type: numeric

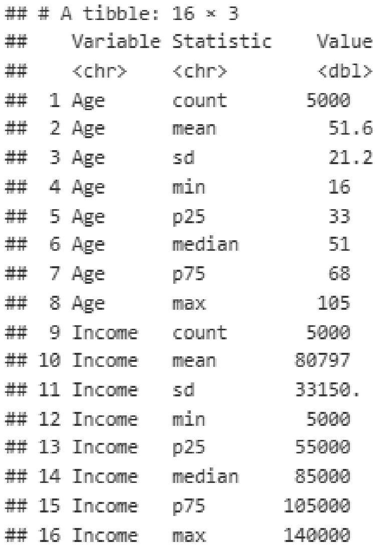
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Quantity	0	1	13.50	13.76	1.00	3.00	6.00	23.00	50.00	▇▁▁▁▁
orderTime	0	1	12.93	5.50	1.00	9.00	13.00	17.00	23.00	▃▅▇▇▅
orderDay	0	1	15.50	8.65	1.00	8.00	15.00	23.00	30.00	▇▇▇▇▇
orderMonth	0	1	6.45	3.28	1.00	4.00	6.00	9.00	12.00	▇▇▇▇▇
orderYear	0	1	2022.46	0.50	2022.00	2022.00	2022.00	2023.00	2023.00	▇▁▁▇
pickingHours	0	1	14.70	10.39	0.43	9.39	14.05	18.72	45.06	▃▇▂▁▁
deliveryHours	0	1	17.48	10.00	0.28	11.55	19.55	25.04	38.05	▅▃▇▇▂
SellingPrice	0	1	3243.75	5412.22	350.45	493.69	627.92	5346.14	19725.18	▇▂▁▁
revenue	0	1	43525.88	112679.52	350.45	2170.24	7705.12	23987.75	986259.00	▇▁▁▁

Table 1.2.2.1: Sales: Summary statistics for numeric variables



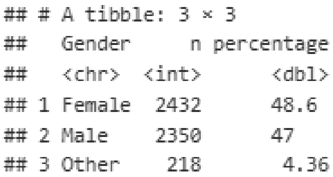
```
## # A tibble: 40 × 3
##    Variable     Statistic    Value
##    <chr>        <chr>        <dbl>
##  1 Quantity     count     100000
##  2 Quantity     mean          13.5
##  3 Quantity     sd            13.8
##  4 Quantity     min            1
##  5 Quantity     p25            3
##  6 Quantity     median         6
##  7 Quantity     p75           23
##  8 Quantity     max           50
##  9 SellingPrice count     100000
## 10 SellingPrice mean        3244.
## # ℹ 30 more rows
```

Table 1.2.3: Customer Dataset analysis 1.2.3.1 Customer demographics: Summary table



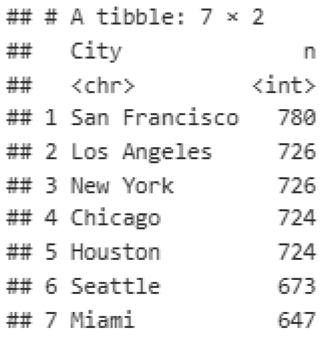
```
## # A tibble: 16 × 3
##    Variable Statistic    Value
##    <chr>    <chr>        <dbl>
##  1 Age      count       5000
##  2 Age      mean          51.6
##  3 Age      sd            21.2
##  4 Age      min           16
##  5 Age      p25           33
##  6 Age      median        51
##  7 Age      p75           68
##  8 Age      max          105
##  9 Income   count       5000
## 10 Income   mean       80797
## 11 Income   sd         33150.
## 12 Income   min         5000
## 13 Income   p25        55000
## 14 Income   median     85000
## 15 Income   p75       105000
## 16 Income   max       140000
```

Table 1.2.3: 4 Gender composition



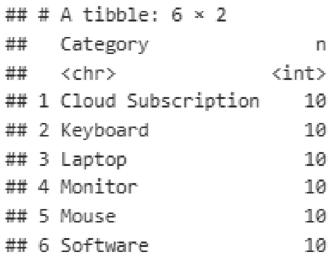
```
## # A tibble: 3 × 3
##   Gender     n percentage
##   <chr>  <int>      <dbl>
## 1 Female  2432      48.6
## 2 Male    2350      47
## 3 Other    218       4.36
```

Table 1.2.3.6: Top 15 cities by customer count



```
## # A tibble: 7 × 2
##   City              n
##   <chr>         <int>
## 1 San Francisco   780
## 2 Los Angeles     726
## 3 New York        726
## 4 Chicago         724
## 5 Houston         724
## 6 Seattle         673
## 7 Miami           647
```

Table 1.2.4: Product dataset analysis 1.2.4.1 Product categories: Count by category



```
## # A tibble: 6 × 2
##   Category               n
##   <chr>              <int>
## 1 Cloud Subscription    10
## 2 Keyboard              10
## 3 Laptop                10
## 4 Monitor               10
## 5 Mouse                 10
## 6 Software              10
```

Table 1.2.4.3: Selling price (summary statistics table)

```
## # A tibble: 1 × 8
##   count  mean    sd   min   p25 median   p75    max
##   <int> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl>
## 1    60 4494. 6504.  350.  512.   794. 6417. 19725.
```

Table 1.2.4.5: Markup: Summary statistics

```
## # A tibble: 1 × 8
##   count  mean    sd   min   p25 median   p75   max
##   <int> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1    60  20.5  6.07  10.1  16.1   20.3  25.7  29.8
```

A2: Figures

Figure 1.2.2.2: Distribution of order quantity (histogram and density)

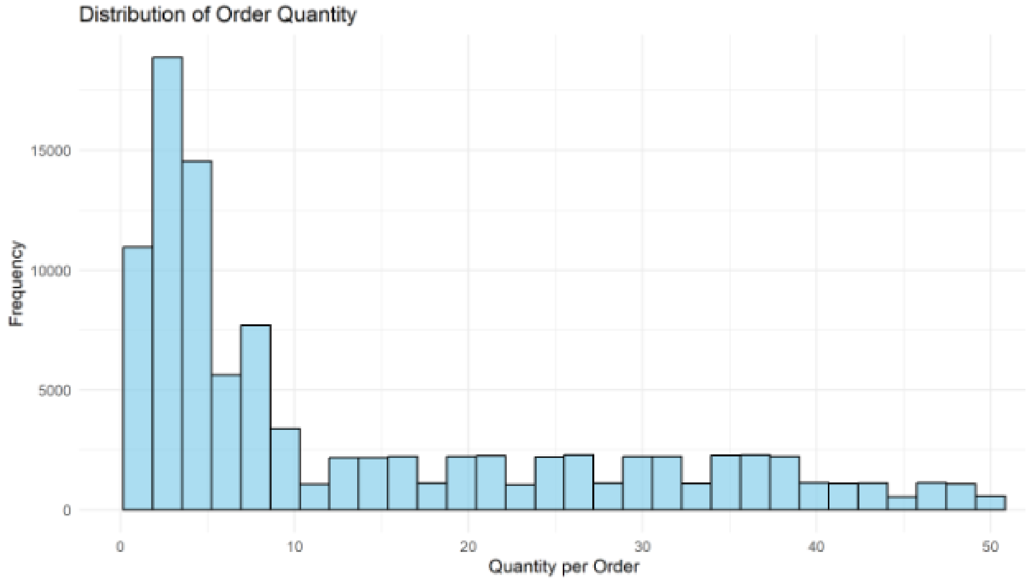


Figure 1.2.2.3: Distribution of unit price (histogram and density)

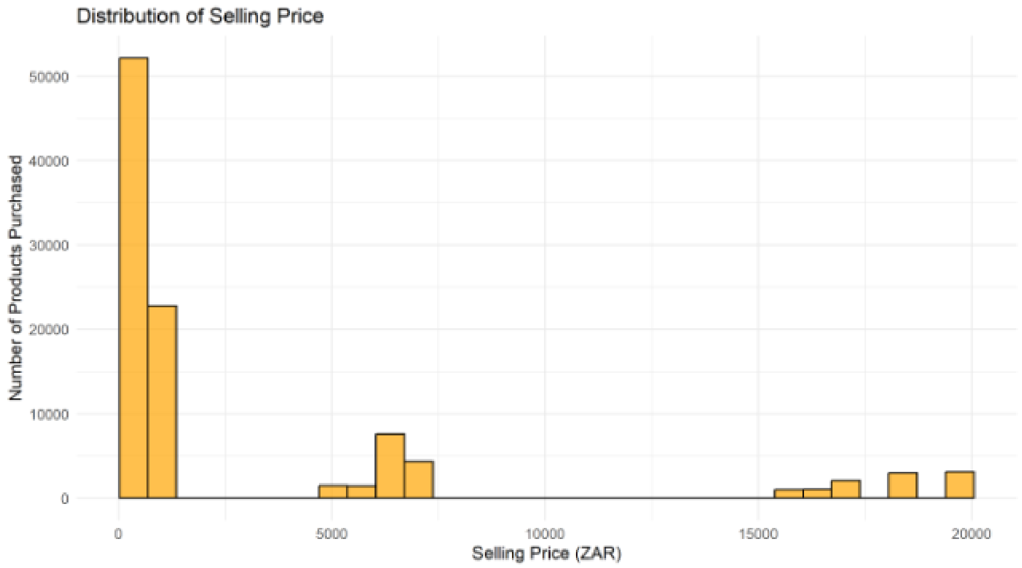


Figure 1.2.2.4: Distribution of line value (revenue per line)

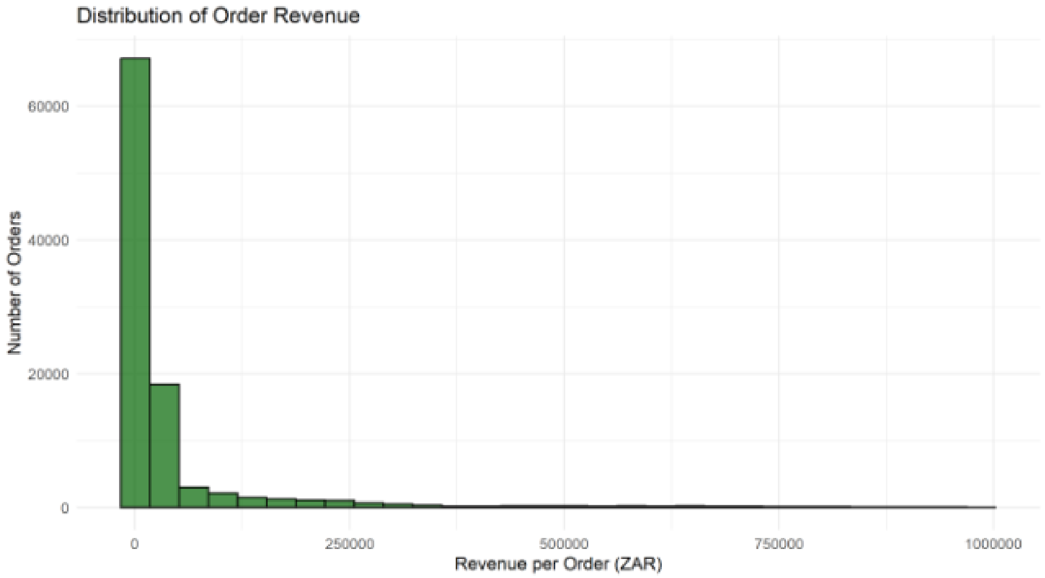


Figure 1.2.2.5: Monthly orders: Count orders per month

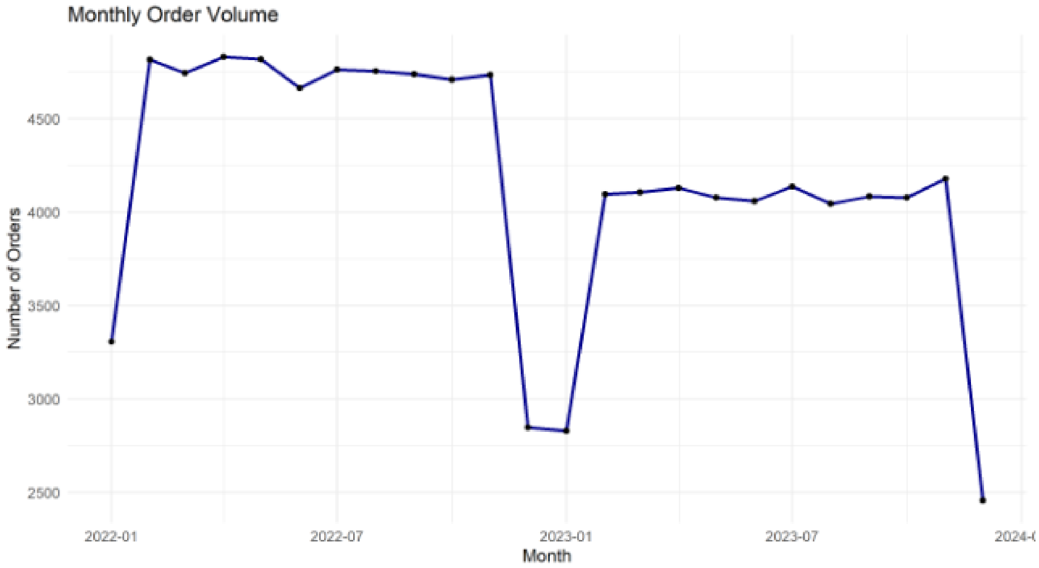


Figure 1.2.2.6: Monthly revenue trend

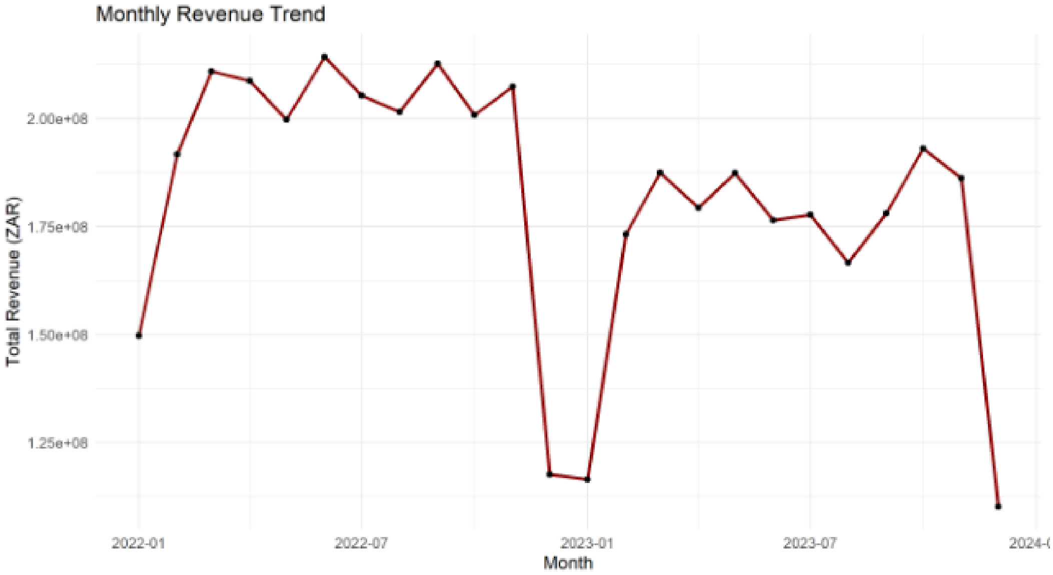


Figure 1.2.2.7: Average order value

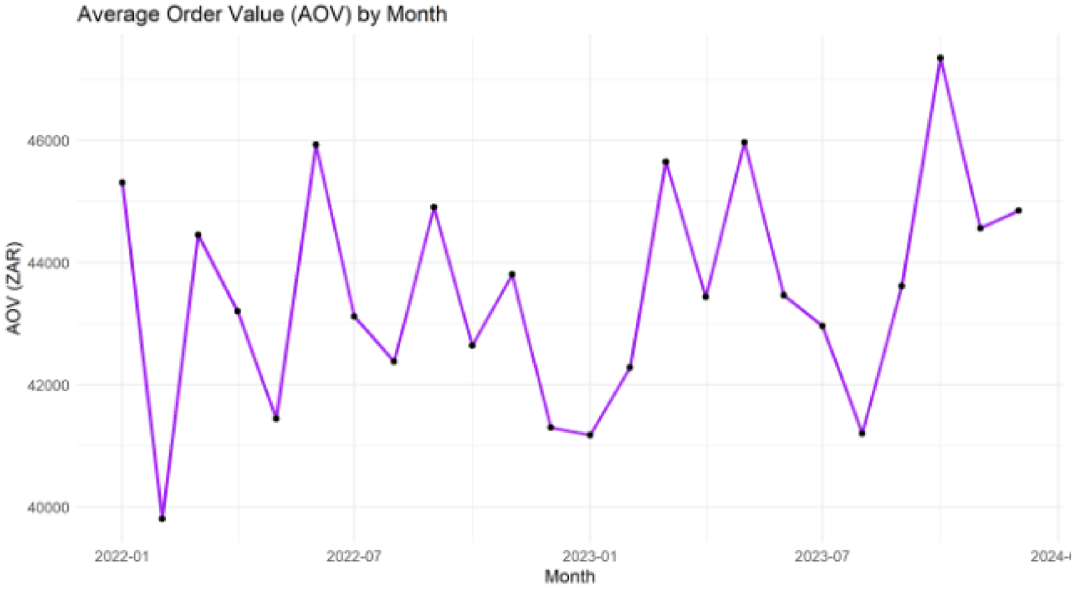


Figure 1.2.2.8: Orders by day of week

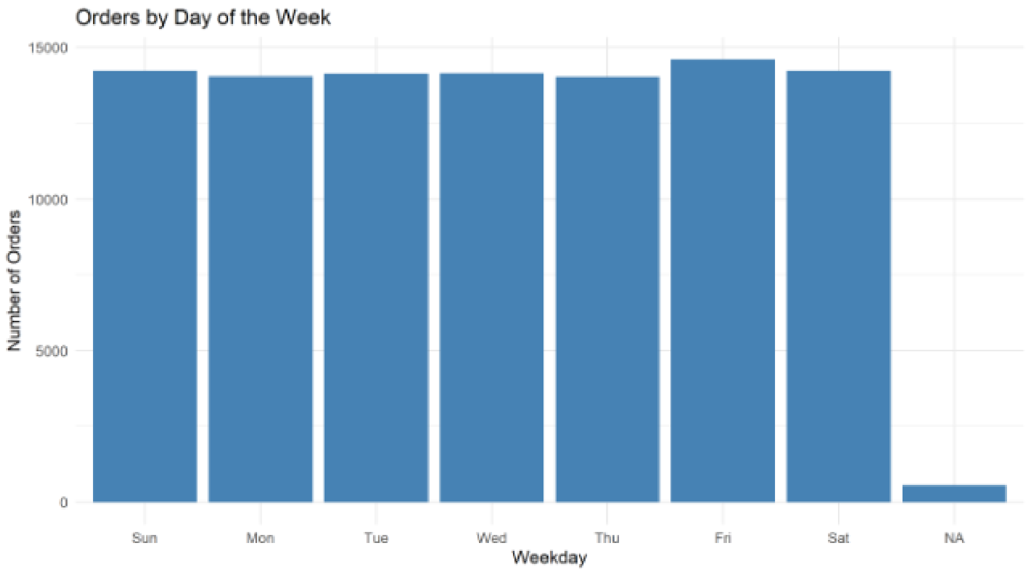


Figure 1.2.2.9: Orders by hour of day

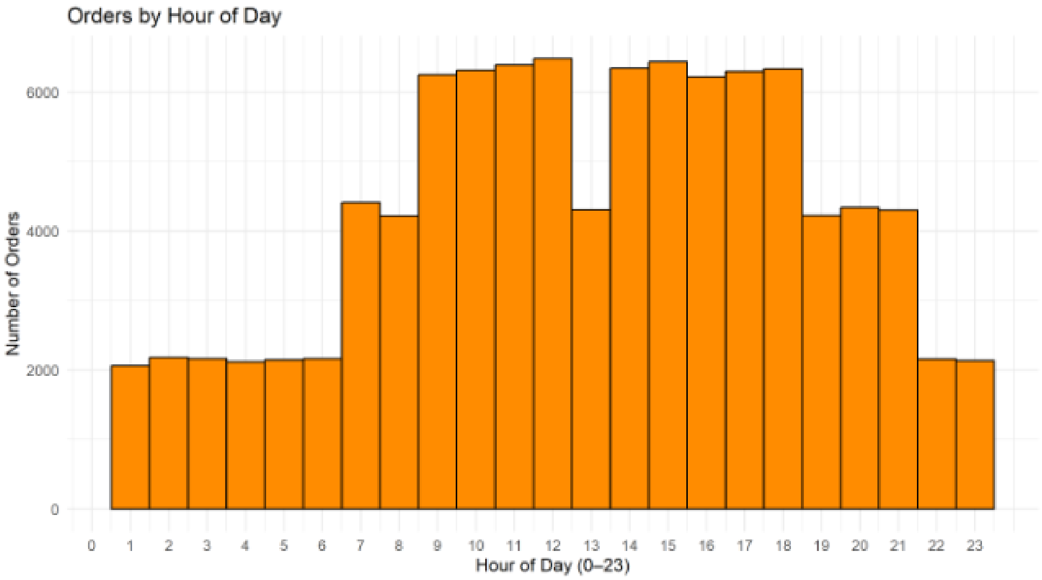


Figure 1.2.2.10: Operational timing: Picking hours distribution

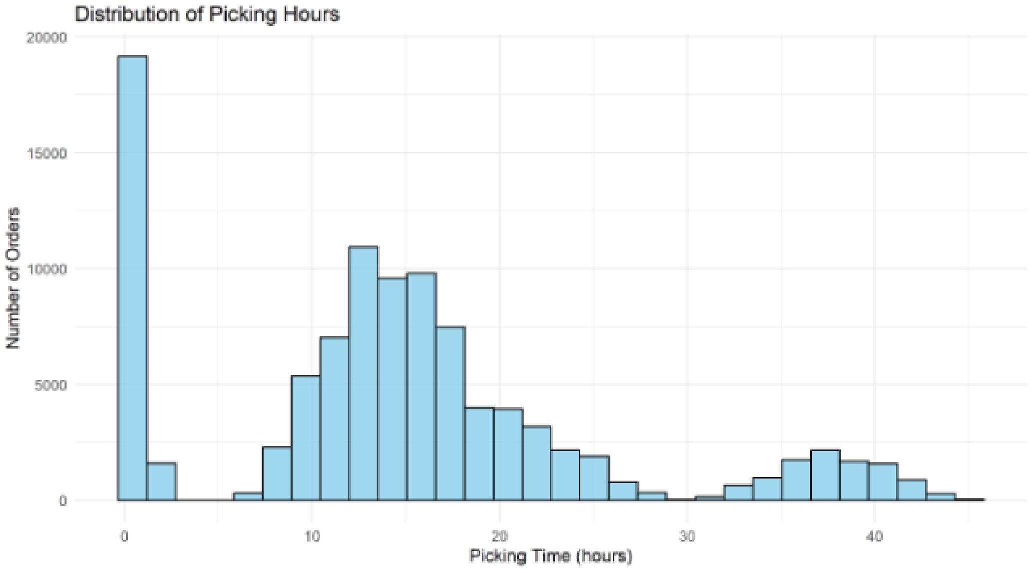


Figure 1.2.2.11: Operational timing: delivery hours distribution

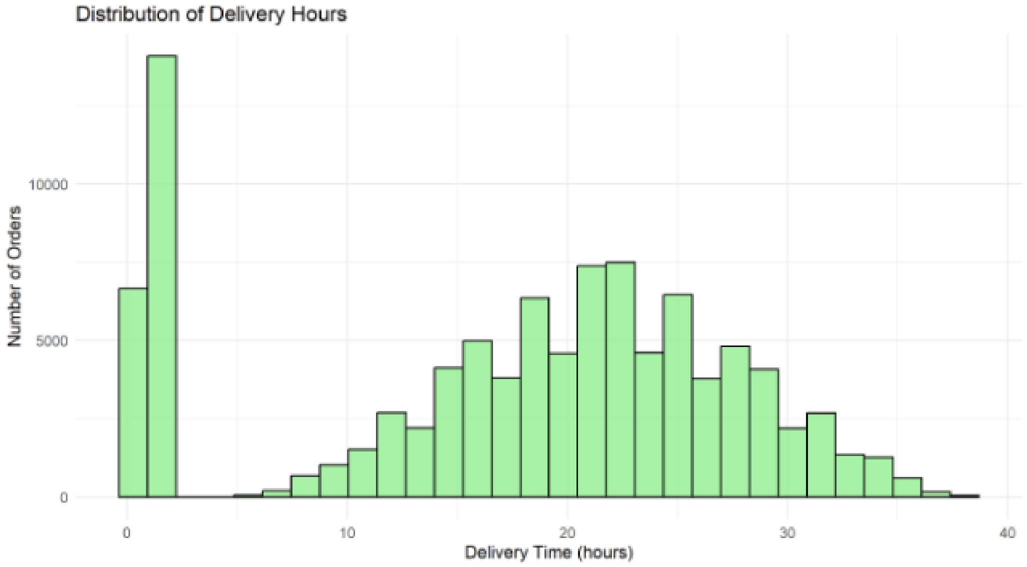


Figure 1.2.2.12: Relationship: Picking vs delivery hours

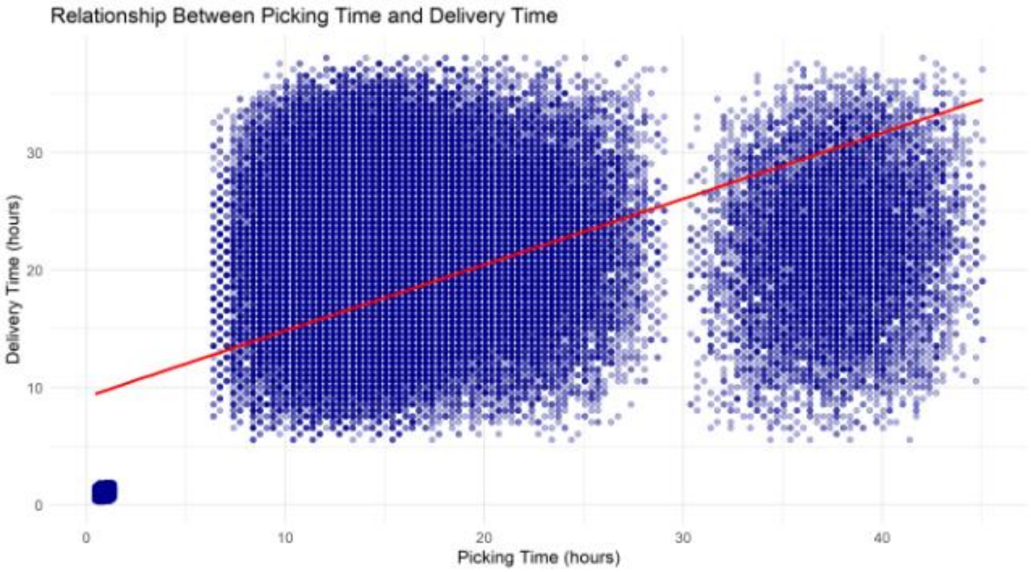


Figure 1.2.2.13: Delivery hours by weekday (boxplot)

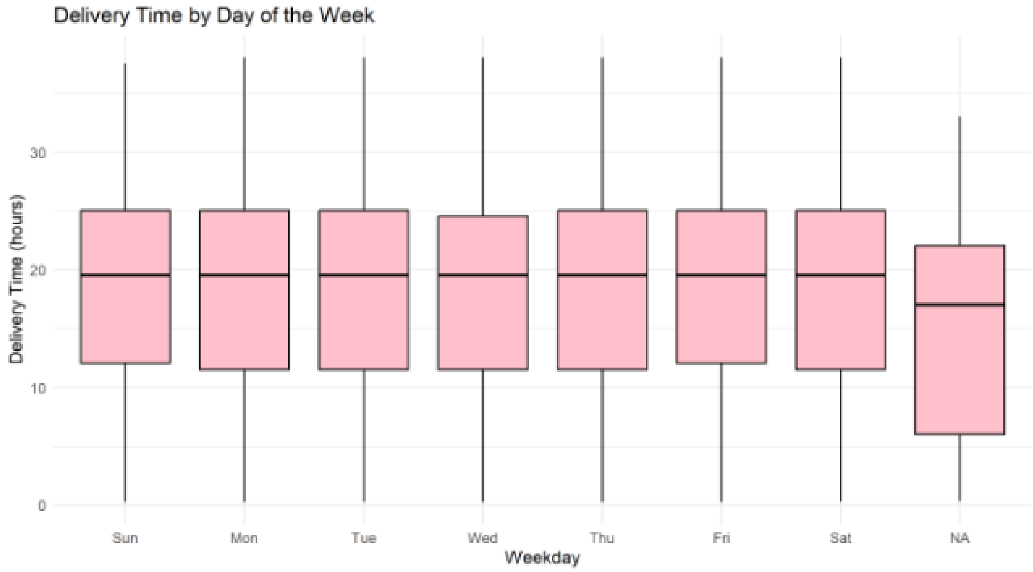


Figure 1.2.2.14: Correlation heatmap: sales numeric variables

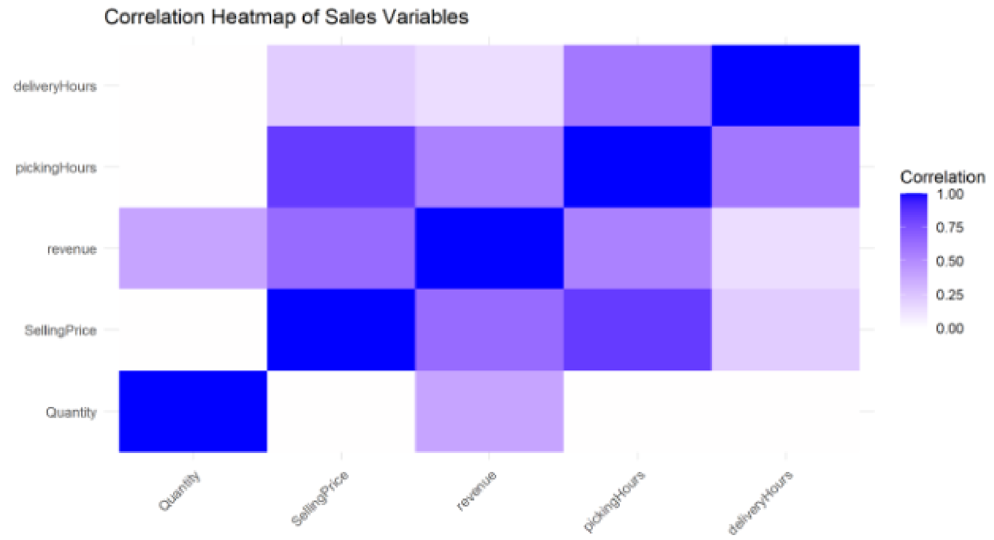


Figure 1.2.2.15: Scatterplot matrix: Sales

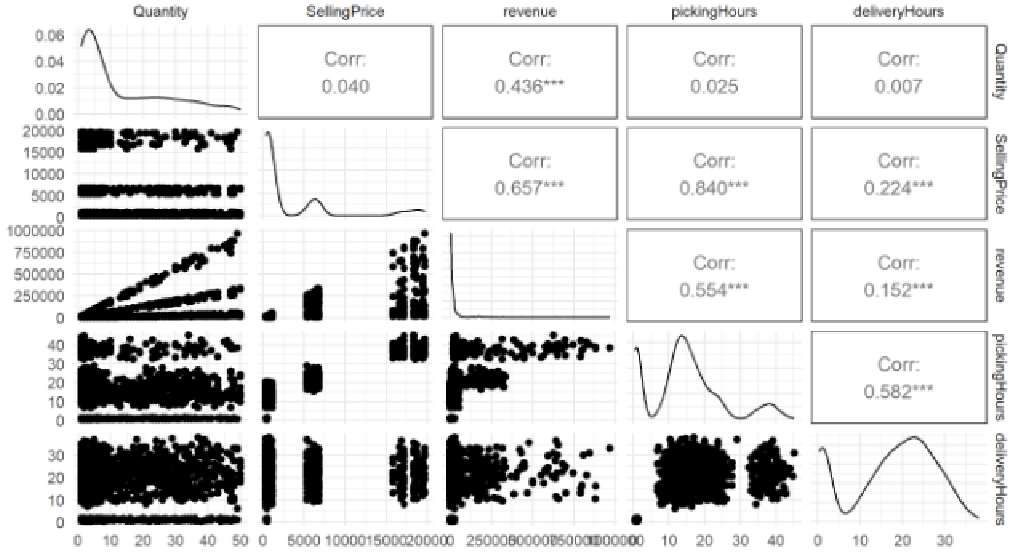


Figure 1.2.3.2: Age distribution (histogram)

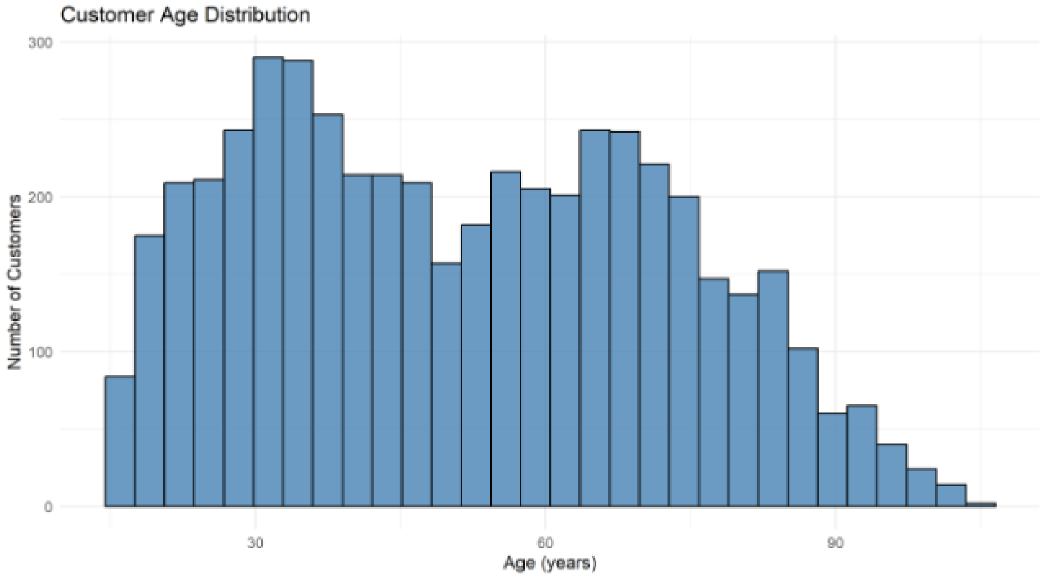


Figure 1.2.3.3: Income distribution (histogram)

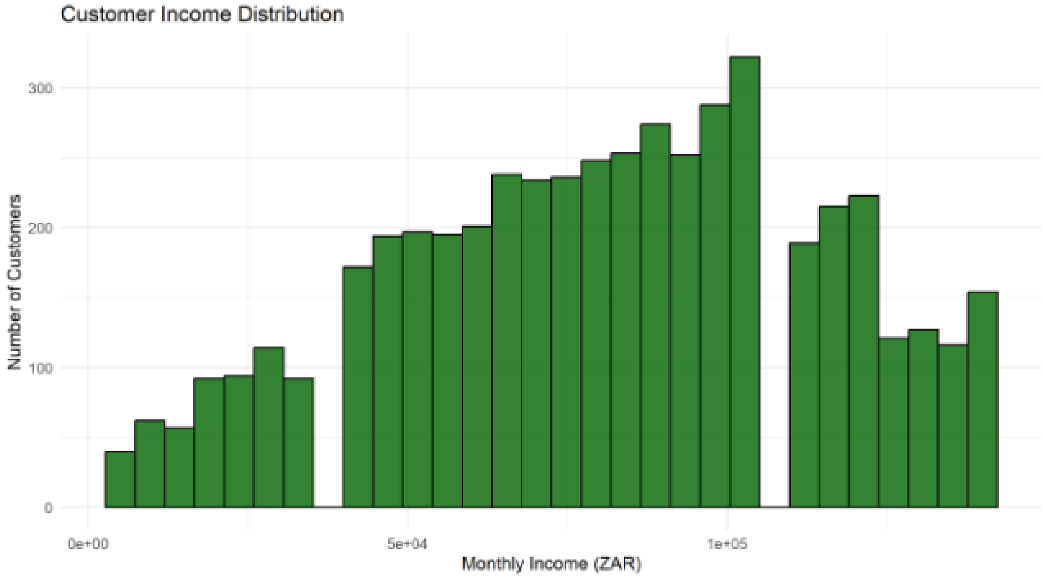


Figure 1.2.3.5: Gender composition (bar chart)

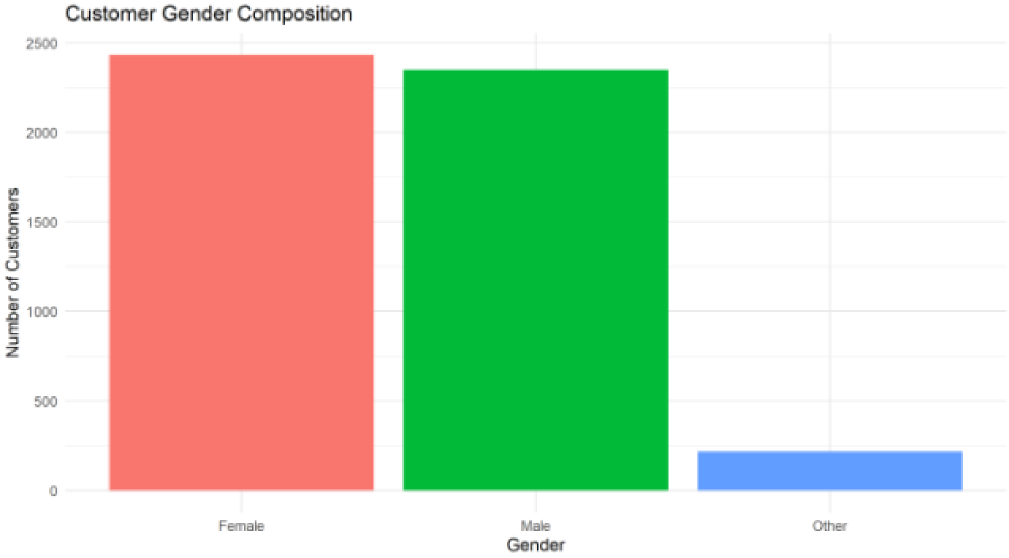


Figure 1.2.3.7: Top 15 cities by customer count (bar chart)

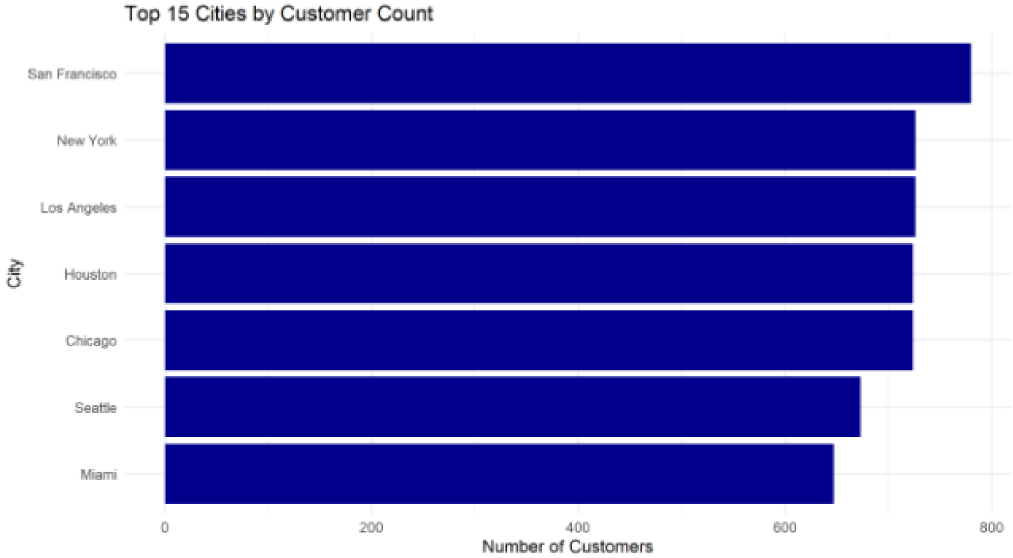


Figure 1.2.4.2: Product categories: Count by category (bar chart)

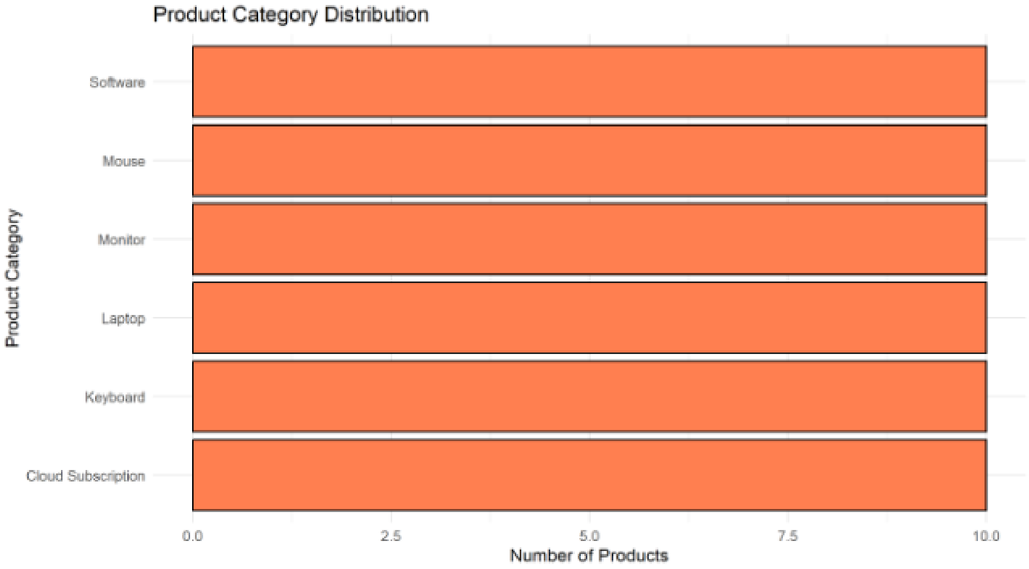


Figure 1.2.4.4: Selling price distribution (histogram)

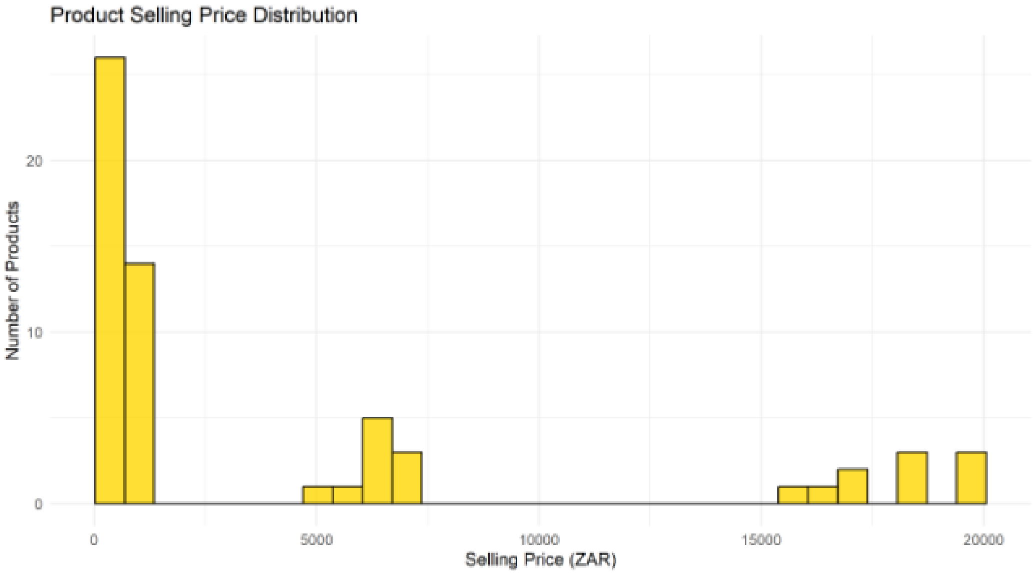


Figure 1.2.4.6: Markup distribution (histogram)

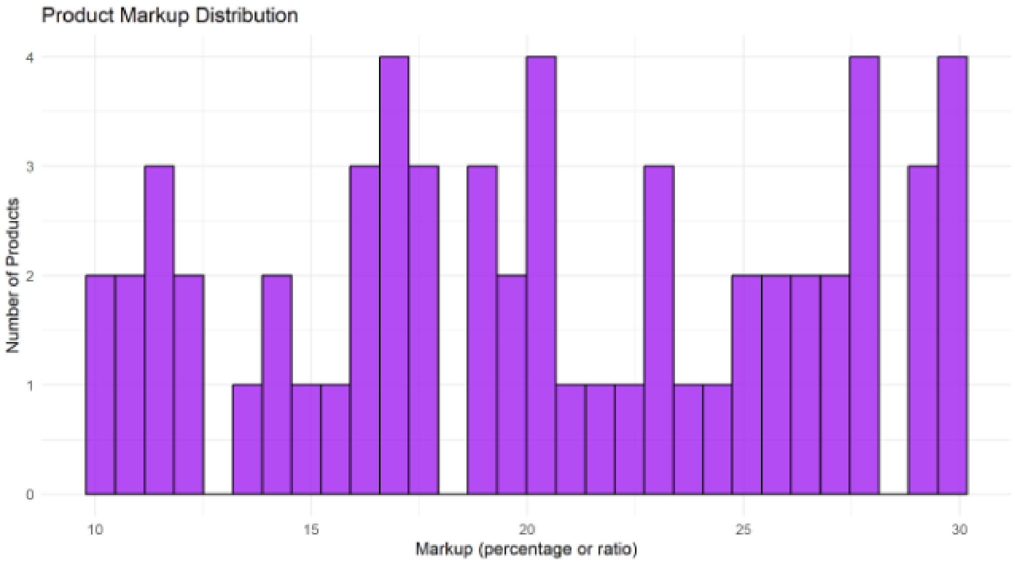


Figure 1.2.4.7: Selling price by category (boxplot)

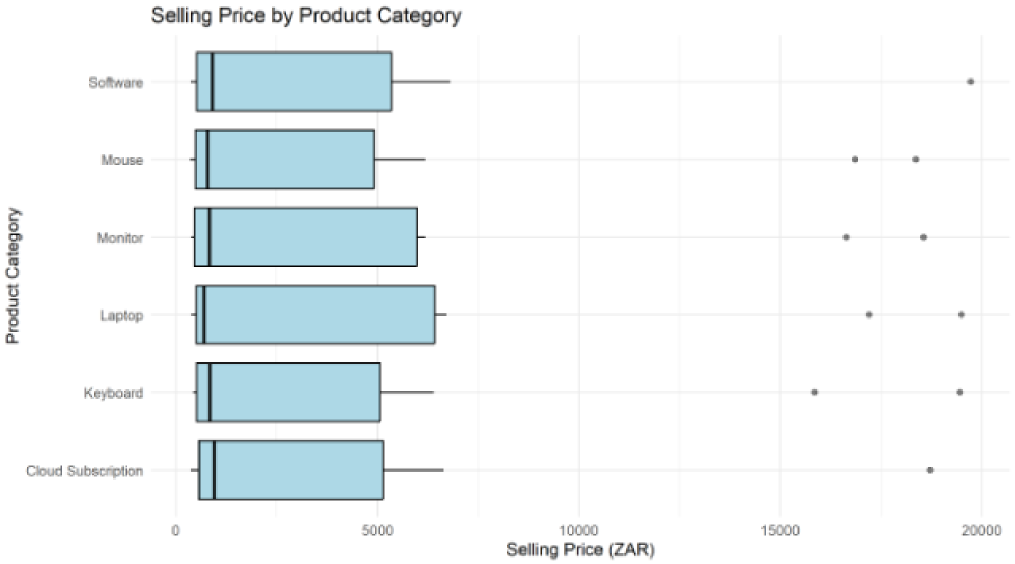
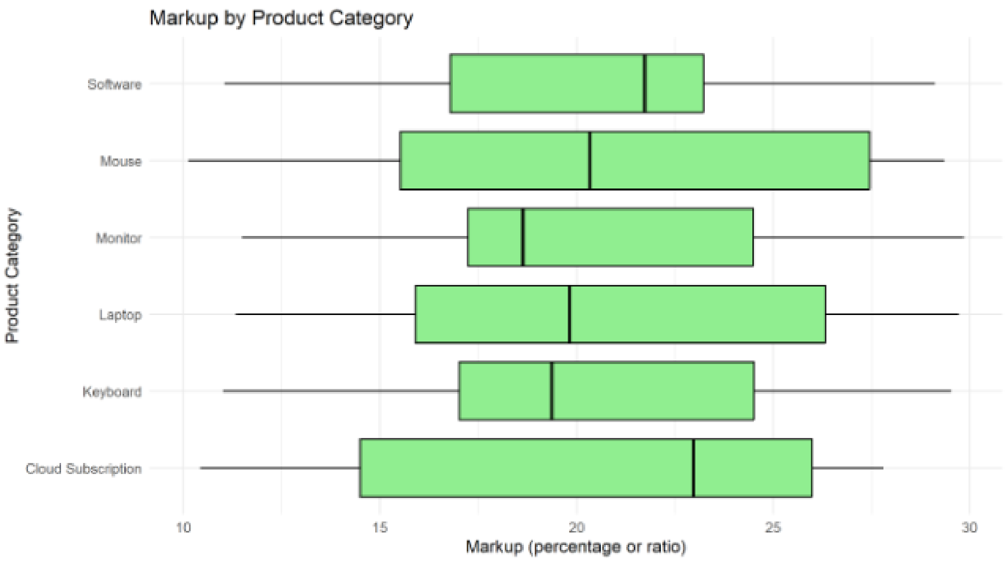


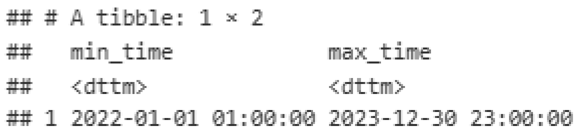
Figure 1.2.4.8: Markup by category (boxplot)


Markup by Product Category
Product Category
Software
Mouse
Monitor
Laptop
Keyboard
Cloud Subscription
10
15
20
25
30
Markup (percentage or ratio)

APPENDIX B:

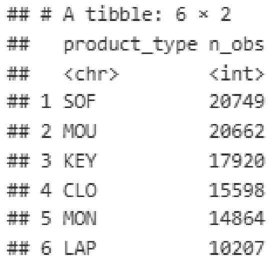
B1: Tables

Table 3.1.3: Order Sanity check



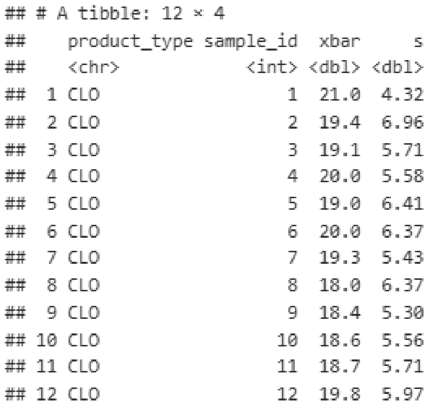
```
## # A tibble: 1 × 2
##   min_time            max_time
##   <dttm>              <dttm>
## 1 2022-01-01 01:00:00 2023-12-30 23:00:00
```

Table 3.1.4: Product types with enough data



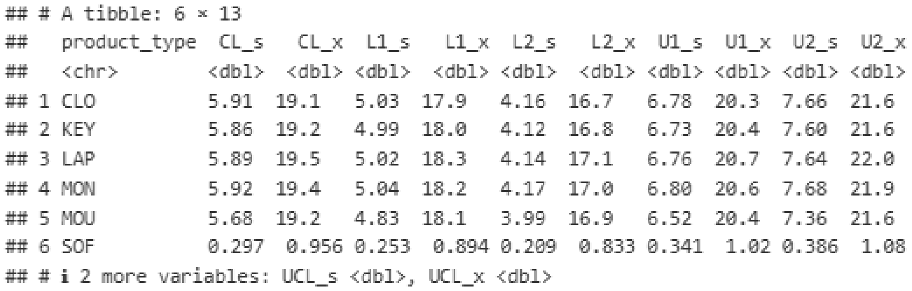
```
## # A tibble: 6 × 2
##   product_type n_obs
##   <chr>        <int>
## 1 SOF          20749
## 2 MOU          20662
## 3 KEY          17920
## 4 CLO          15598
## 5 MON          14864
## 6 LAP          10207
```

Table 3.1.5: Phase I samples (n=24), first 30 samples for all product types



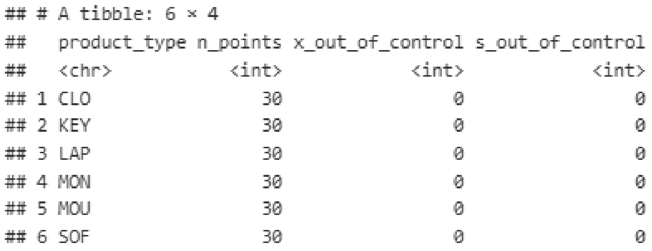
```
## # A tibble: 12 × 4
##    product_type sample_id  xbar     s
##    <chr>            <int> <dbl> <dbl>
##  1 CLO                  1  21.0  4.32
##  2 CLO                  2  19.4  6.96
##  3 CLO                  3  19.1  5.71
##  4 CLO                  4  20.0  5.58
##  5 CLO                  5  19.0  6.41
##  6 CLO                  6  20.0  6.37
##  7 CLO                  7  19.3  5.43
##  8 CLO                  8  18.0  6.37
##  9 CLO                  9  18.4  5.30
## 10 CLO                 10  18.6  5.56
## 11 CLO                 11  18.7  5.71
## 12 CLO                 12  19.8  5.97
```

Table 3.1.6: Control-chart constants and limits for every type



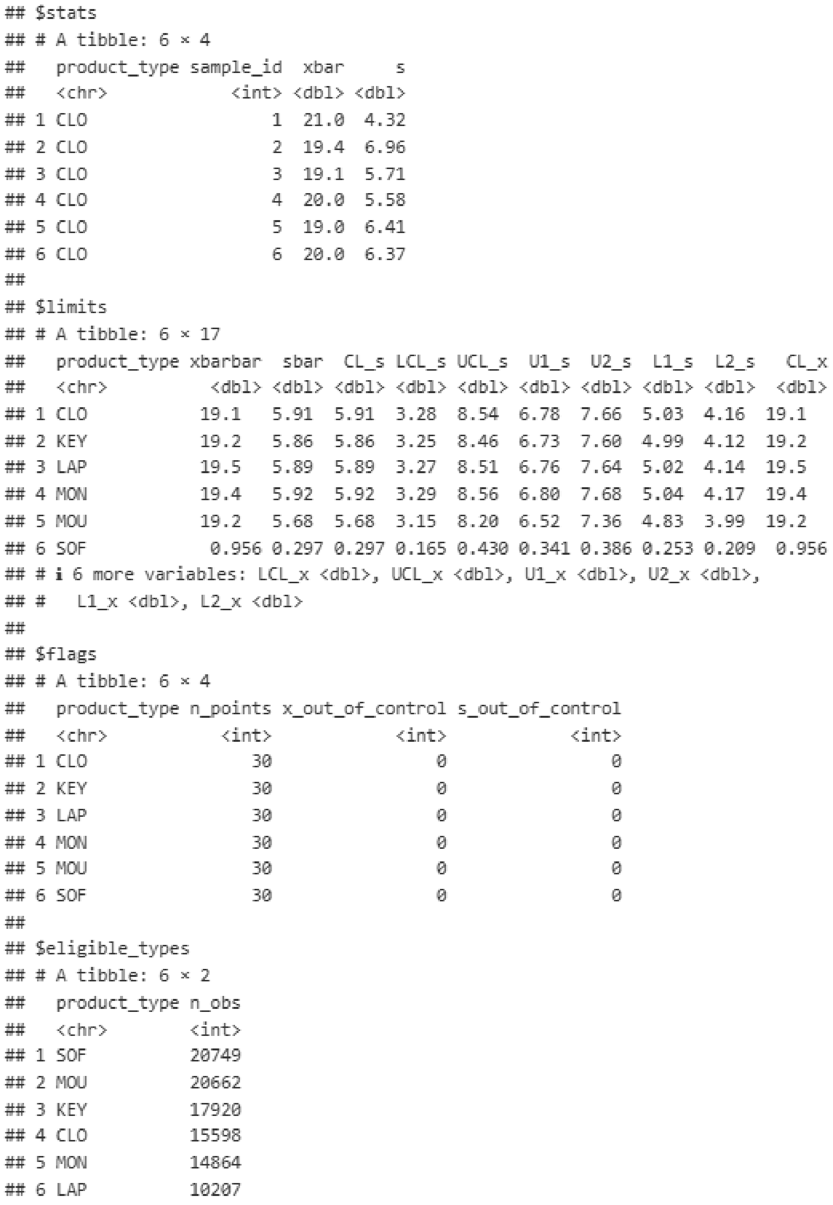
```
## # A tibble: 6 × 13
##   product_type  CL_s   CL_x  L1_s   L1_x  L2_s   L2_x  U1_s  U1_x  U2_s  U2_x
##   <chr>        <dbl>  <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 CLO          5.91  19.1   5.03  17.9   4.16  16.7   6.78  20.3  7.66  21.6
## 2 KEY          5.86  19.2   4.99  18.0   4.12  16.8   6.73  20.4  7.60  21.6
## 3 LAP          5.89  19.5   5.02  18.3   4.14  17.1   6.76  20.7  7.64  22.0
## 4 MON          5.92  19.4   5.04  18.2   4.17  17.0   6.80  20.6  7.68  21.9
## 5 MOU          5.68  19.2   4.83  18.1   3.99  16.9   6.52  20.4  7.36  21.6
## 6 SOF          0.297  0.956 0.253  0.894 0.209  0.833 0.341  1.02 0.386  1.08
## # ℹ 2 more variables: UCL_s <dbl>, UCL_x <dbl>
```

Table 3.1.11: Basic Western-Electric style flags (counts) per type



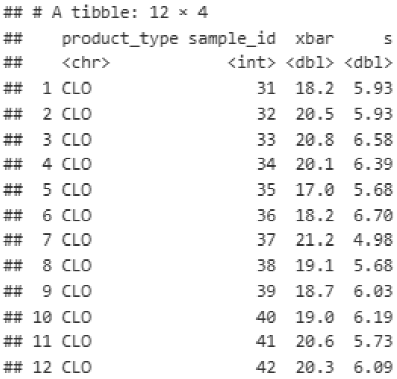
```
## # A tibble: 6 × 4
##   product_type n_points x_out_of_control s_out_of_control
##   <chr>           <int>            <int>            <int>
## 1 CLO                30                0                0
## 2 KEY                30                0                0
## 3 LAP                30                0                0
## 4 MON                30                0                0
## 5 MOU                30                0                0
## 6 SOF                30                0                0
```

Table 3.1.12: Object with everything in one place



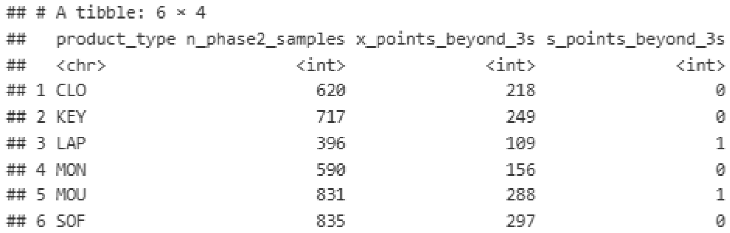
```
## $stats
## # A tibble: 6 × 4
##   product_type sample_id  xbar     s
##   <chr>            <int> <dbl> <dbl>
## 1 CLO                  1  21.0  4.32
## 2 CLO                  2  19.4  6.96
## 3 CLO                  3  19.1  5.71
## 4 CLO                  4  20.0  5.58
## 5 CLO                  5  19.0  6.41
## 6 CLO                  6  20.0  6.37
##
## $limits
## # A tibble: 6 × 17
##   product_type xbarbar  sbar  CL_s LCL_s UCL_s  U1_s  U2_s  L1_s  L2_s   CL_x
##   <chr>          <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 CLO           19.1   5.91  5.91  3.28  8.54  6.78  7.66  5.03  4.16  19.1
## 2 KEY           19.2   5.86  5.86  3.25  8.46  6.73  7.60  4.99  4.12  19.2
## 3 LAP           19.5   5.89  5.89  3.27  8.51  6.76  7.64  5.02  4.14  19.5
## 4 MON           19.4   5.92  5.92  3.29  8.56  6.80  7.68  5.04  4.17  19.4
## 5 MOU           19.2   5.68  5.68  3.15  8.20  6.52  7.36  4.83  3.99  19.2
## 6 SOF            0.956 0.297 0.297 0.165 0.430 0.341 0.386 0.253 0.209  0.956
## # ℹ 6 more variables: LCL_x <dbl>, UCL_x <dbl>, U1_x <dbl>, U2_x <dbl>,
## #   L1_x <dbl>, L2_x <dbl>
##
## $flags
## # A tibble: 6 × 4
##   product_type n_points x_out_of_control s_out_of_control
##   <chr>           <int>            <int>            <int>
## 1 CLO                30                0                0
## 2 KEY                30                0                0
## 3 LAP                30                0                0
## 4 MON                30                0                0
## 5 MOU                30                0                0
## 6 SOF                30                0                0
##
## $eligible_types
## # A tibble: 6 × 2
##   product_type n_obs
##   <chr>        <int>
## 1 SOF          20749
## 2 MOU          20662
## 3 KEY          17920
## 4 CLO          15598
## 5 MON          14864
## 6 LAP          10207
```

Table 3.2.1: Build Phase-II samples (all product types: samples 31, 32…)



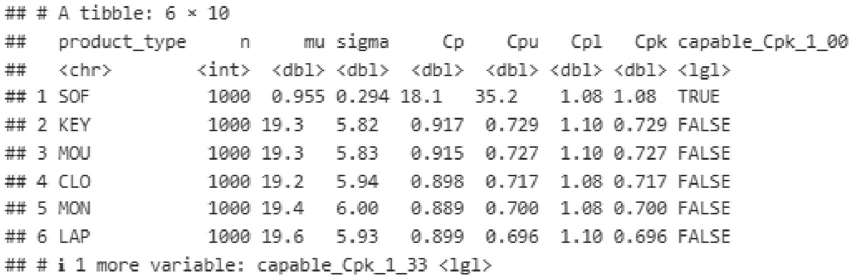
```
## # A tibble: 12 × 4
##    product_type sample_id  xbar     s
##    <chr>            <int> <dbl> <dbl>
##  1 CLO                 31  18.2  5.93
##  2 CLO                 32  20.5  5.93
##  3 CLO                 33  20.8  6.58
##  4 CLO                 34  20.1  6.39
##  5 CLO                 35  17.0  5.68
##  6 CLO                 36  18.2  6.70
##  7 CLO                 37  21.2  4.98
##  8 CLO                 38  19.1  5.68
##  9 CLO                 39  18.7  6.03
## 10 CLO                 40  19.0  6.19
## 11 CLO                 41  20.6  5.73
## 12 CLO                 42  20.3  6.09
```

Table 3.2.2: Join phase-II stats to phase-I limits and flag rule-1



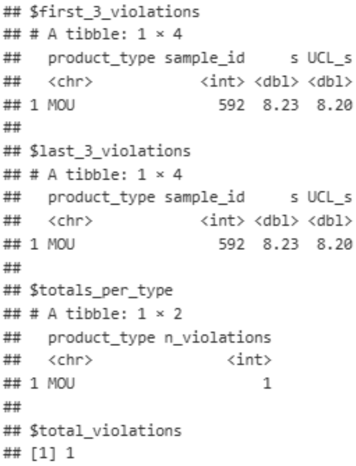
```
## # A tibble: 6 × 4
##   product_type n_phase2_samples x_points_beyond_3s s_points_beyond_3s
##   <chr>                   <int>              <int>              <int>
## 1 CLO                       620                218                  0
## 2 KEY                       717                249                  0
## 3 LAP                       396                109                  1
## 4 MON                       590                156                  0
## 5 MOU                       831                288                  1
## 6 SOF                       835                297                  0
```

Table 3.3: Process Capability (first 1000 deliveries per product type) 3.3.1 Build capability metrics



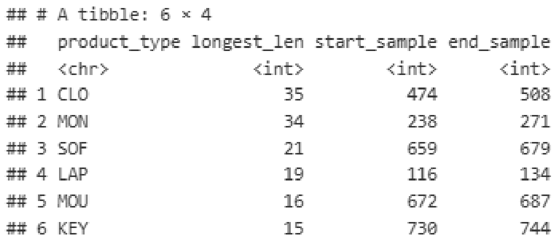
```
## # A tibble: 6 × 10
##   product_type     n     mu sigma    Cp    Cpu   Cpl   Cpk capable_Cpk_1_00
##   <chr>        <int>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <lgl>
## 1 SOF           1000  0.955 0.294 18.1   35.2   1.08 1.08  TRUE
## 2 KEY           1000 19.3   5.82   0.917  0.729  1.10 0.729 FALSE
## 3 MOU           1000 19.3   5.83   0.915  0.727  1.10 0.727 FALSE
## 4 CLO           1000 19.2   5.94   0.898  0.717  1.08 0.717 FALSE
## 5 MON           1000 19.4   6.00   0.889  0.700  1.08 0.700 FALSE
## 6 LAP           1000 19.6   5.93   0.899  0.696  1.10 0.696 FALSE
## # ℹ 1 more variable: capable_Cpk_1_33 <lgl>
```

Table 3.4.1: s-chart: samples above +3σ UCL_s



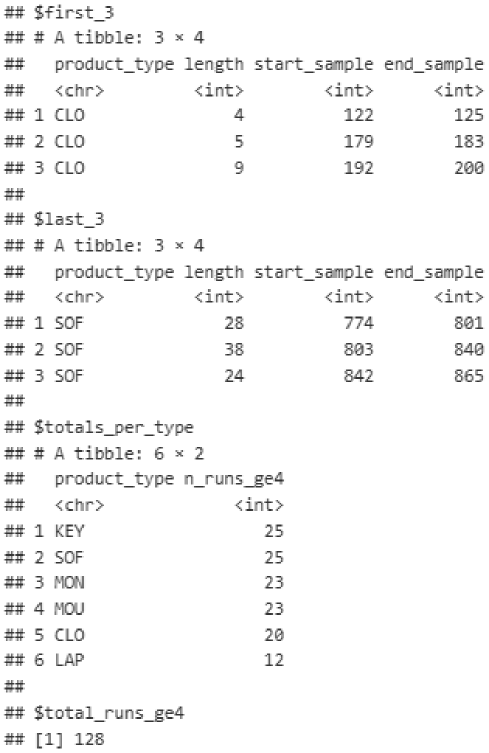
```
## $first_3_violations
## # A tibble: 1 × 4
##   product_type sample_id     s UCL_s
##   <chr>            <int> <dbl> <dbl>
## 1 MOU                592  8.23  8.20
##
## $last_3_violations
## # A tibble: 1 × 4
##   product_type sample_id     s UCL_s
##   <chr>            <int> <dbl> <dbl>
## 1 MOU                592  8.23  8.20
##
## $totals_per_type
## # A tibble: 1 × 2
##   product_type n_violations
##   <chr>               <int>
## 1 MOU                     1
##
## $total_violations
## [1] 1
```

Table 3.4.2: Longest good-control run of s within



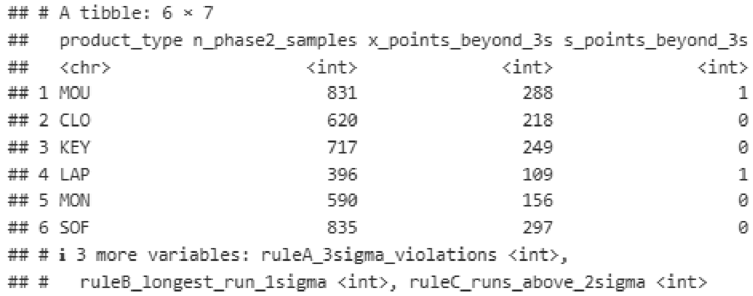
```
## # A tibble: 6 × 4
##   product_type longest_len start_sample end_sample
##   <chr>              <int>        <int>      <int>
## 1 CLO                   35          474        508
## 2 MON                   34          238        271
## 3 SOF                   21          659        679
## 4 LAP                   19          116        134
## 5 MOU                   16          672        687
## 6 KEY                   15          730        744
```

Table 3.4.3: 4 consecutive X-bar samples above the upper 2σ line



```
## $first_3
## # A tibble: 3 × 4
##   product_type length start_sample end_sample
##   <chr>         <int>        <int>      <int>
## 1 CLO               4          122        125
## 2 CLO               5          179        183
## 3 CLO               9          192        200
##
## $last_3
## # A tibble: 3 × 4
##   product_type length start_sample end_sample
##   <chr>         <int>        <int>      <int>
## 1 SOF              28          774        801
## 2 SOF              38          803        840
## 3 SOF              24          842        865
##
## $totals_per_type
## # A tibble: 6 × 2
##   product_type n_runs_ge4
##   <chr>             <int>
## 1 KEY                  25
## 2 SOF                  25
## 3 MON                  23
## 4 MOU                  23
## 5 CLO                  20
## 6 LAP                  12
##
## $total_runs_ge4
## [1] 128
```

Table 3.4: Summary of SPC violations



```
## # A tibble: 6 × 7
##   product_type n_phase2_samples x_points_beyond_3s s_points_beyond_3s
##   <chr>                   <int>              <int>              <int>
## 1 MOU                       831                288                  1
## 2 CLO                       620                218                  0
## 3 KEY                       717                249                  0
## 4 LAP                       396                109                  1
## 5 MON                       590                156                  0
## 6 SOF                       835                297                  0
## # ℹ 3 more variables: ruleA_3sigma_violations <int>,
## #   ruleB_longest_run_1sigma <int>, ruleC_runs_above_2sigma <int>
```

B2: Figures

Figure 3.1.7.1: s-chart (Phase-I) for each SOF

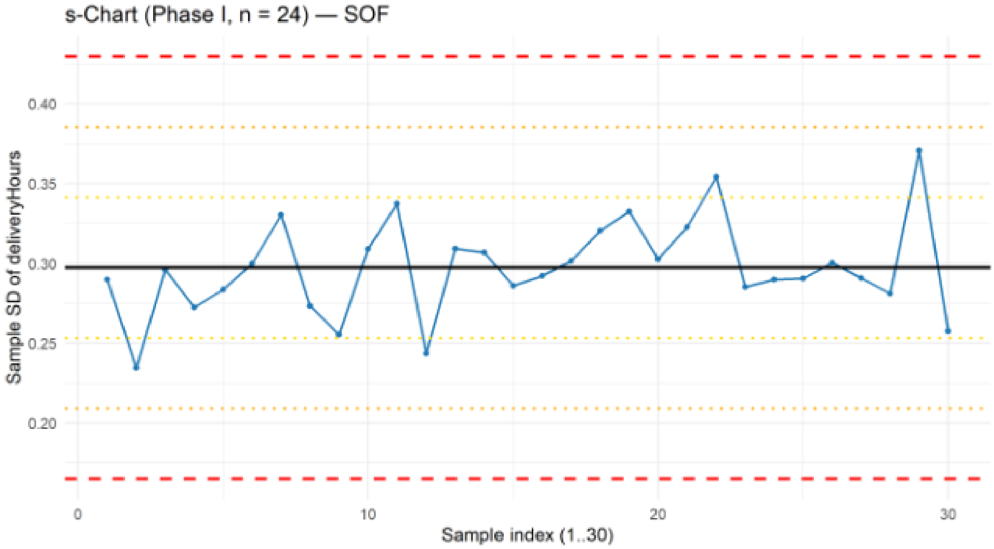


Figure 3.1.7.2: s-chart (Phase-I) for MOU

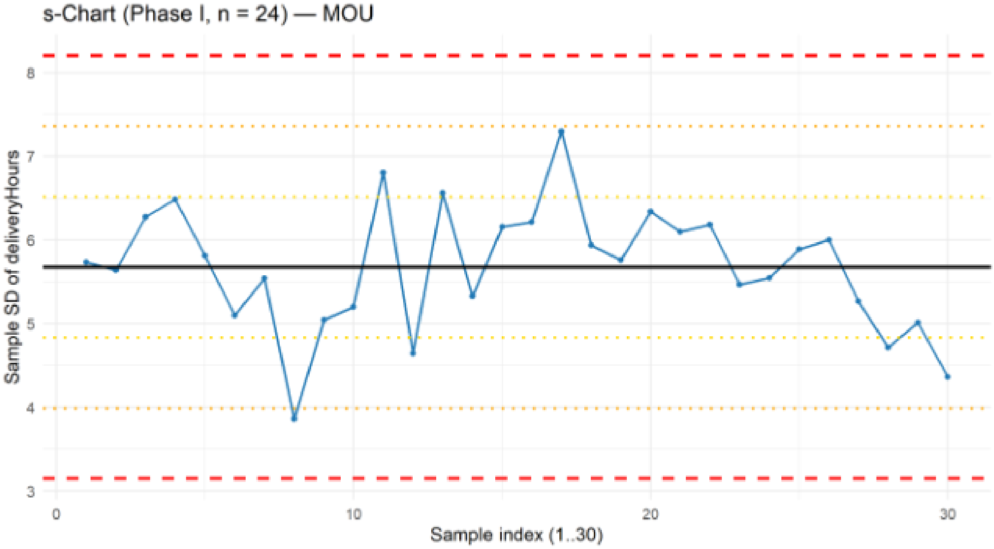


Figure 3.1.7.3: s-chart (Phase-I) for KEY

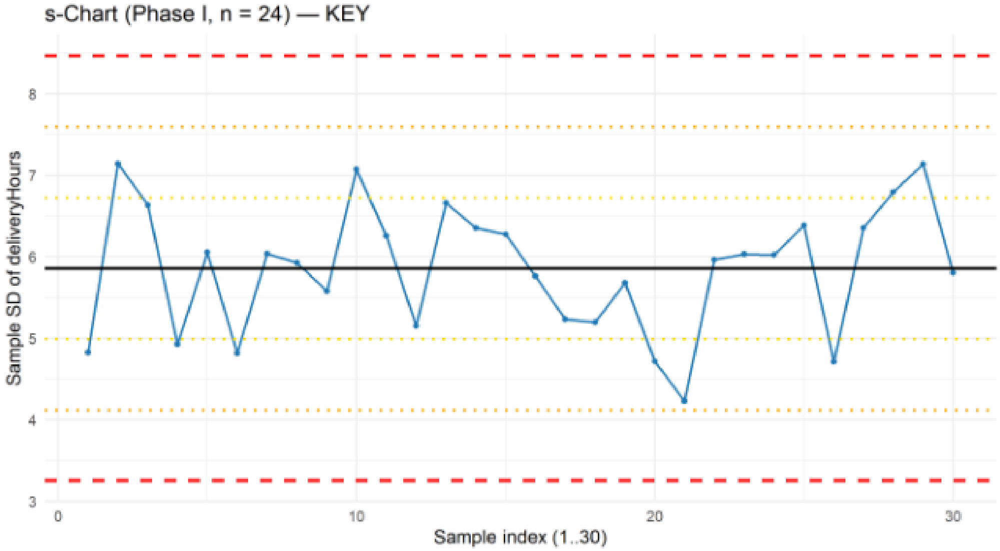


Figure 3.1.7.4: s-chart (Phase-I) for CLO

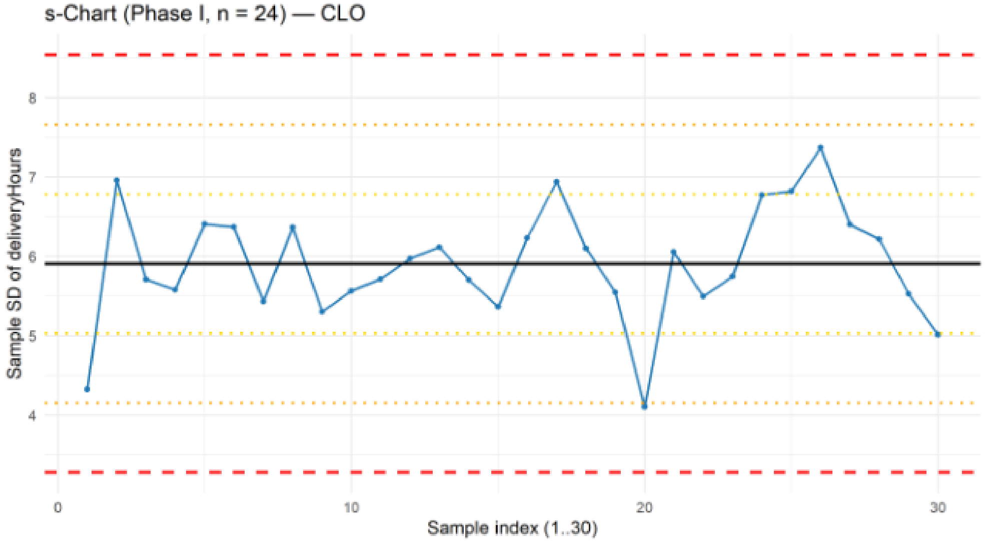


Figure 3.1.7.5: s-chart (Phase-I) for MON

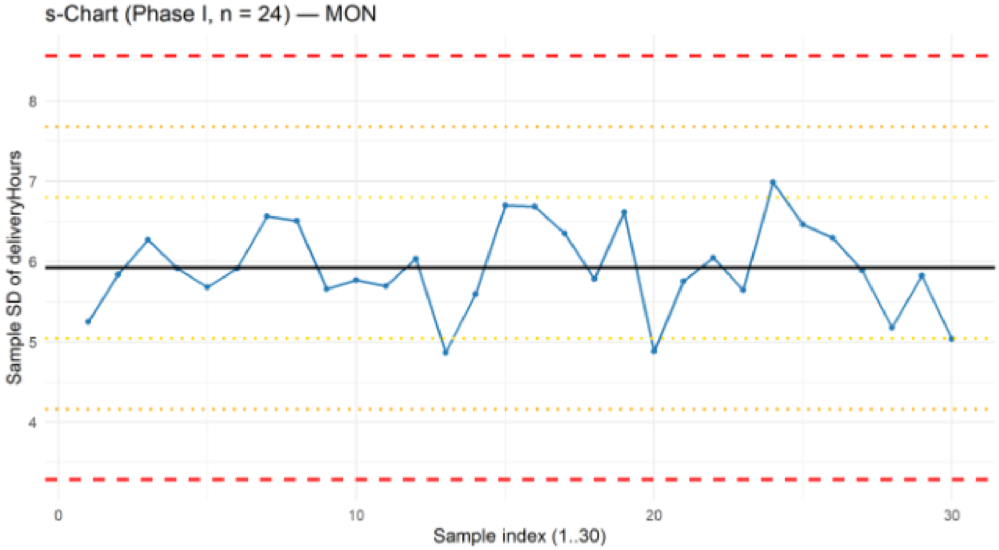


Figure 3.1.7.6: s-chart (Phase-I) for LAP

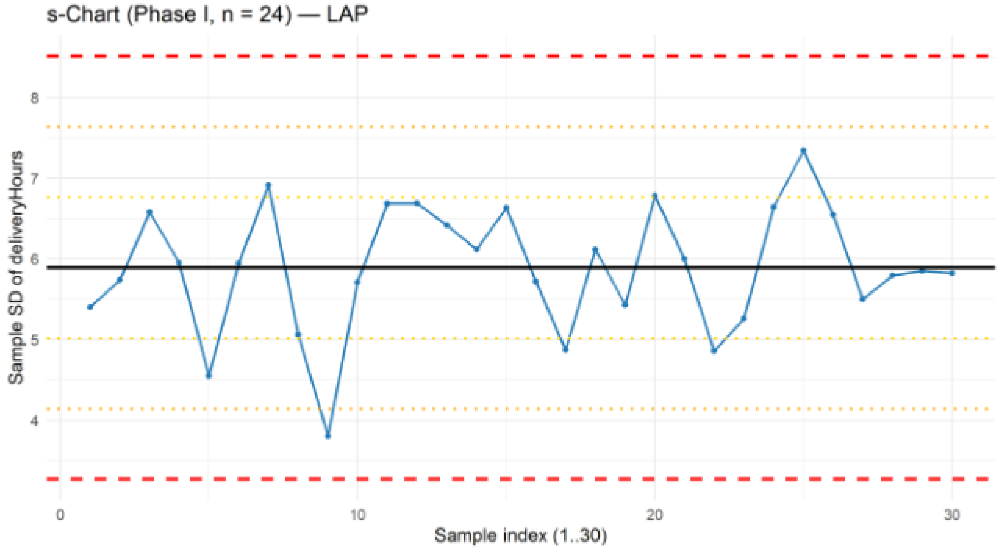


Figure 3.1.8: Faceted s-charts

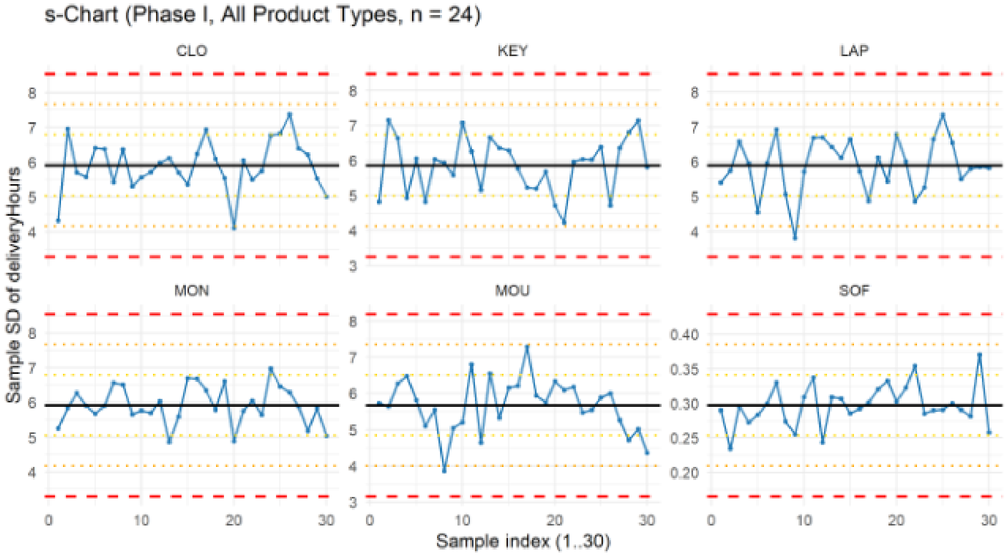


Figure 3.1.9:1 X-bar chart (phase I) for SOF

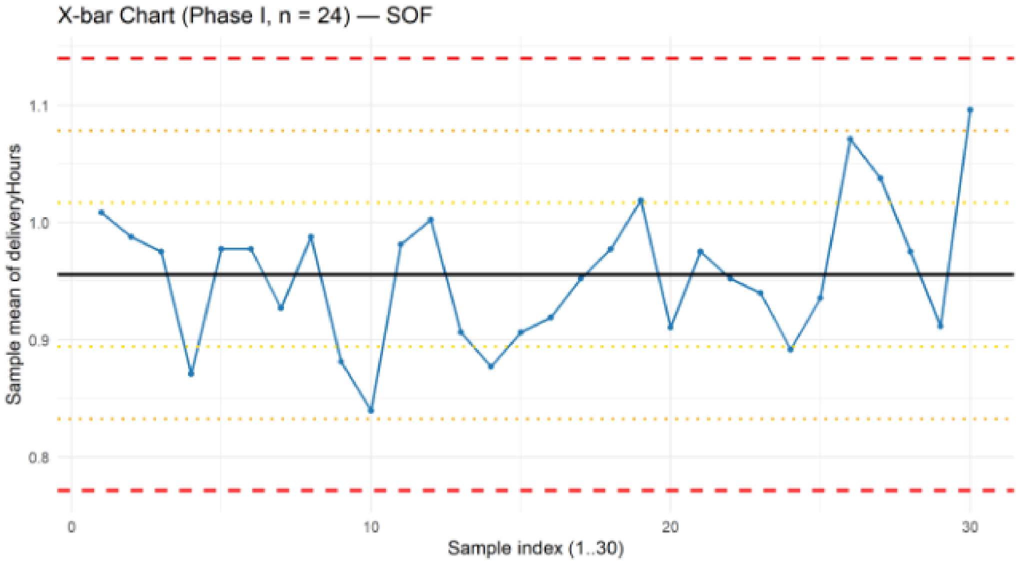


Figure 3.1.9.2: X-bar chart (phase I) for MOU

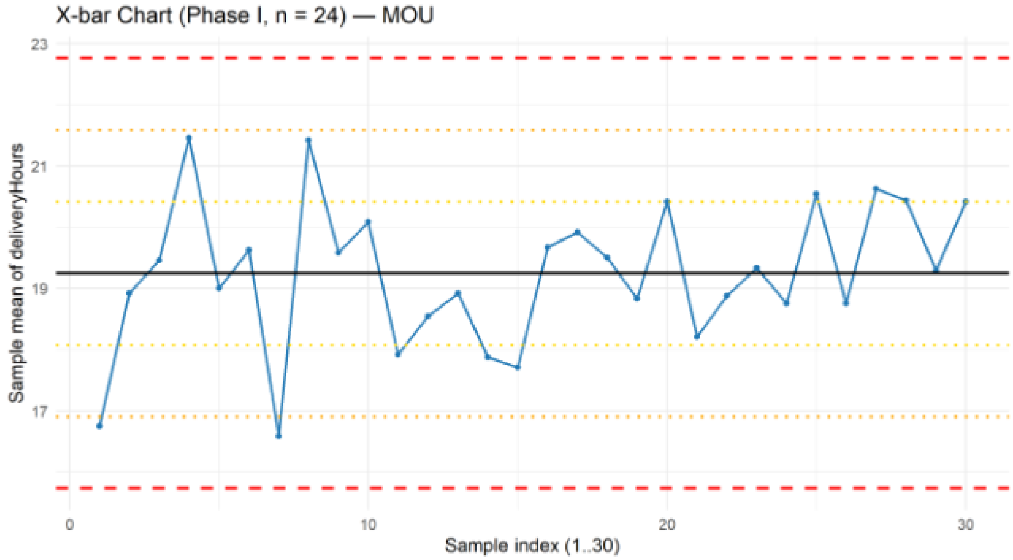


Figure 3.1.9.3: X-bar chart (phase I) for KEY

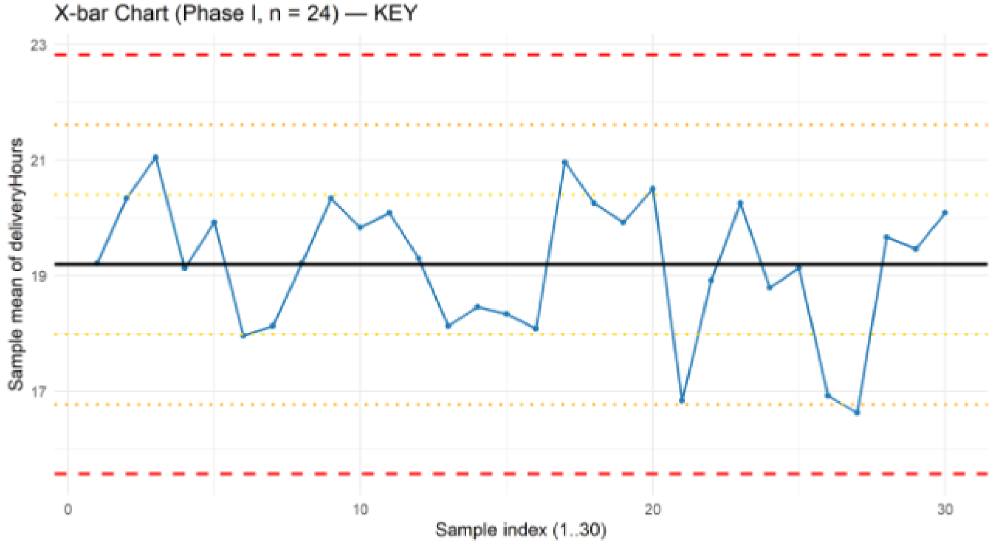


Figure 3.1.9.4: X-bar chart (phase I) for CLO

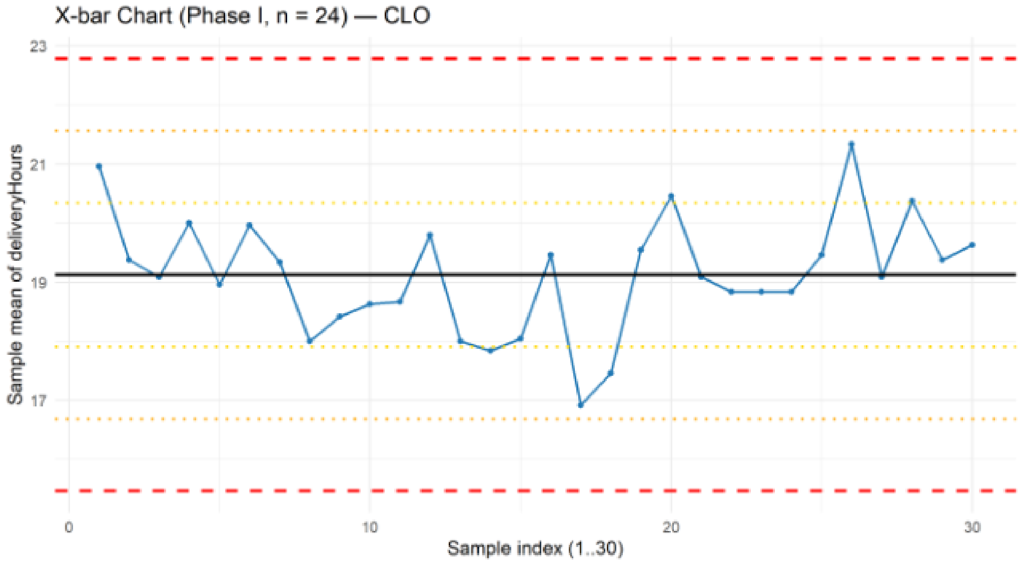


Figure 3.1.9.5: X-bar chart (phase I) for MON

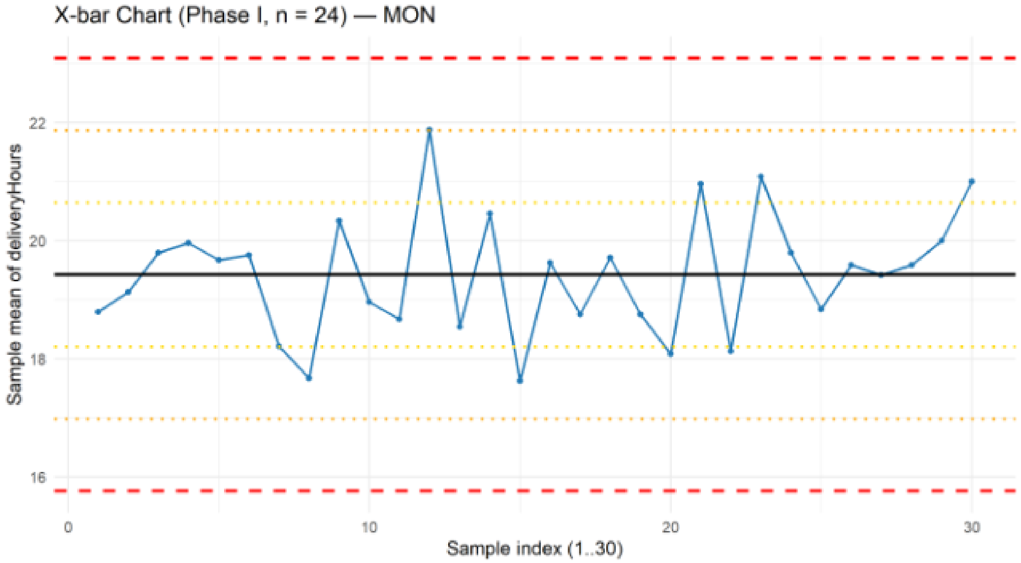


Figure 3.1.9.6: X-bar chart (phase I) for LAP

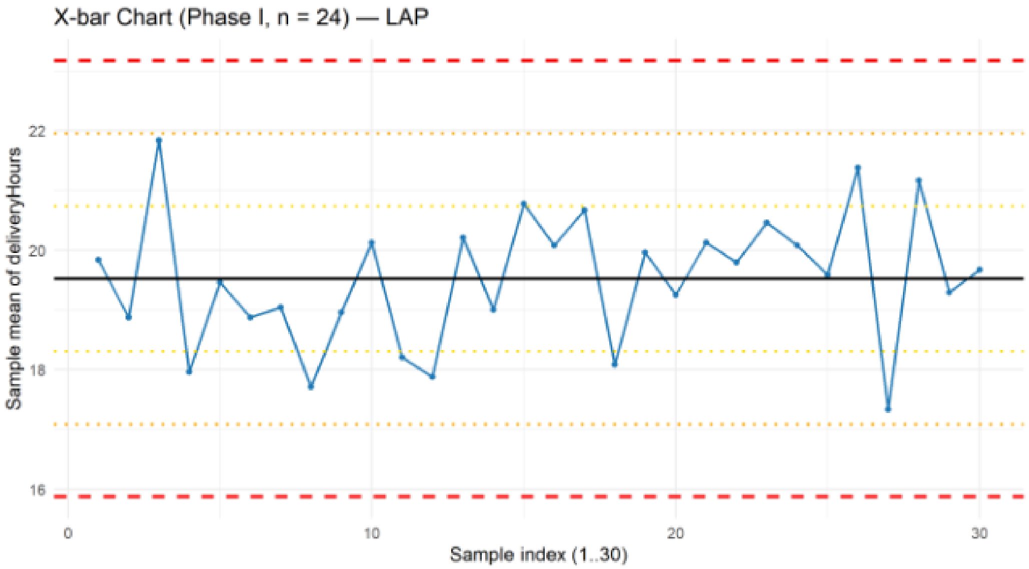


Figure 3.1.10: Faceted X-bar charts

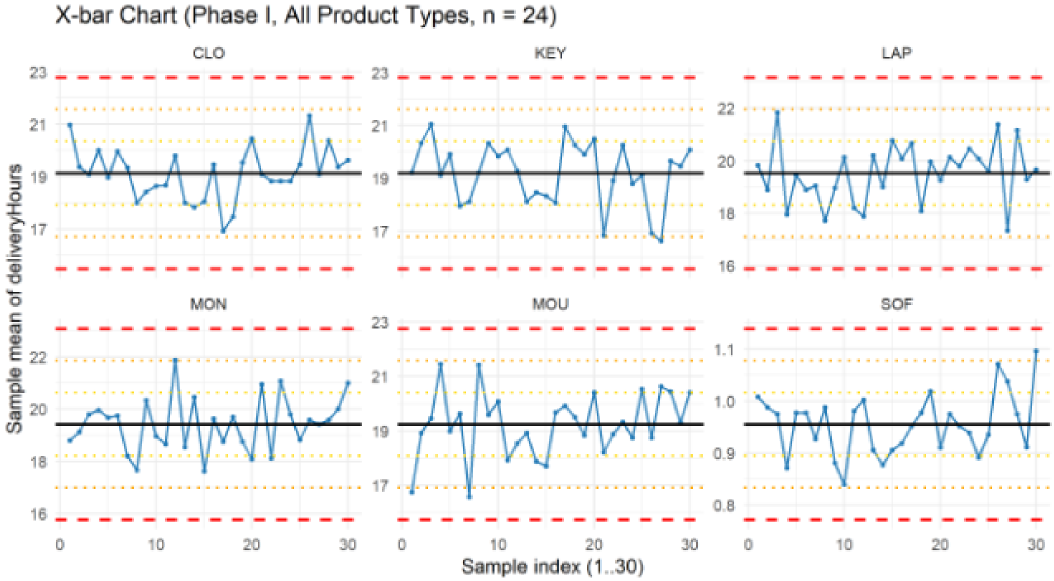


Figure 3.2.3: Faceted Phase-II X-bar (All product types, free y-scales, Rule-1 highlights)

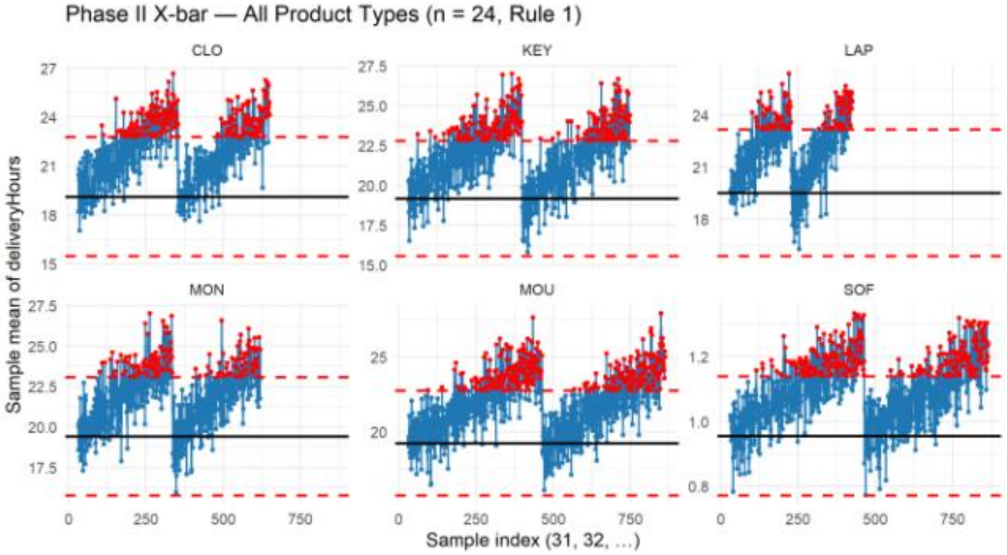


Figure 3.2.4: Faceted Phase-II s-chart (All product types, free y-scales, Rule-1 highlights)

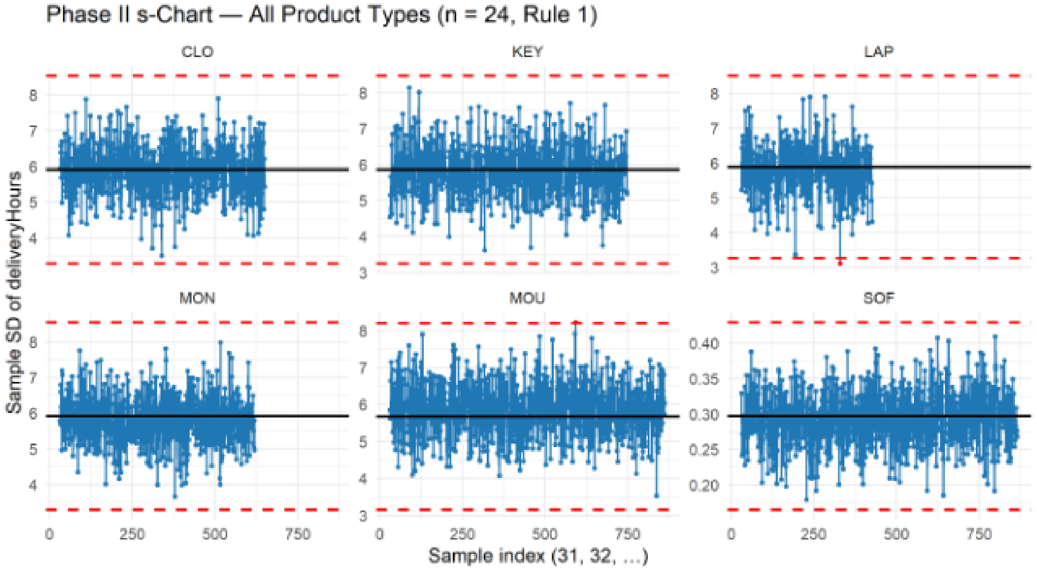
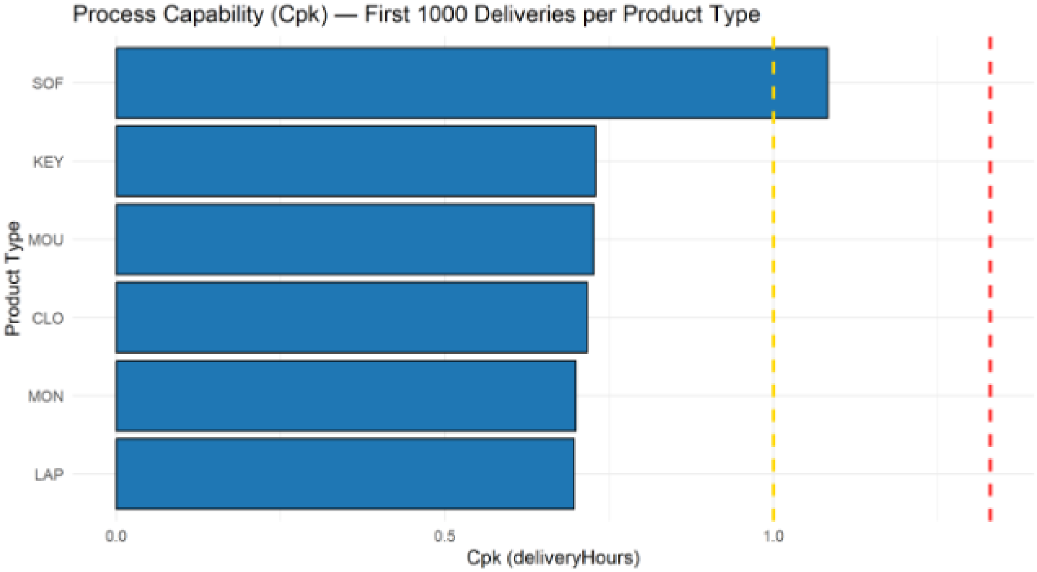


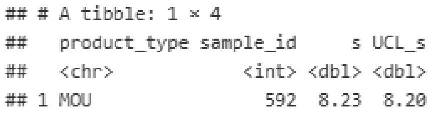
Figure 3.3.2: Visual: Cpk by product type



APPENDIX C:

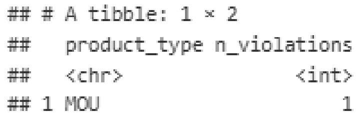
C1: Tables

Table 4.1A1: 1 s-sample outside +3sigma



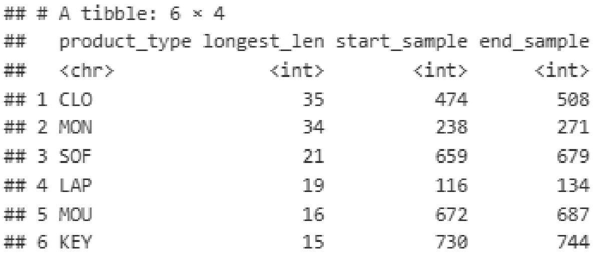
```
## # A tibble: 1 × 4
##   product_type sample_id     s UCL_s
##   <chr>            <int> <dbl> <dbl>
## 1 MOU                592  8.23  8.20
```

Table 4.1A2: 1 s-sample outside +3sigma



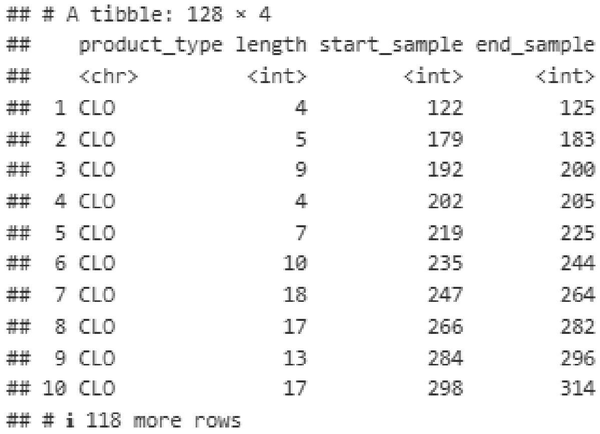
```
## # A tibble: 1 × 2
##   product_type n_violations
##   <chr>               <int>
## 1 MOU                     1
```

Table 4.1B: Longest consecutive run with s within plus or minus one sigma



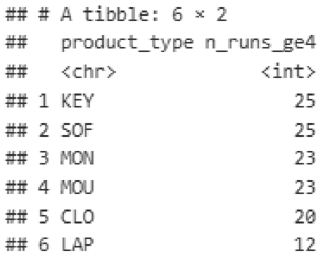
```
## # A tibble: 6 × 4
##   product_type longest_len start_sample end_sample
##   <chr>              <int>        <int>      <int>
## 1 CLO                   35          474        508
## 2 MON                   34          238        271
## 3 SOF                   21          659        679
## 4 LAP                   19          116        134
## 5 MOU                   16          672        687
## 6 KEY                   15          730        744
```

Table 4.1C1: 4 consecutive X bar samples above 2 sigma



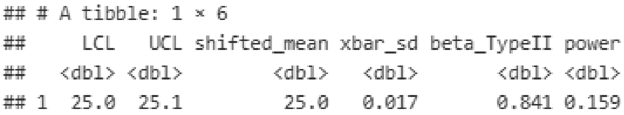
```
## # A tibble: 128 × 4
##    product_type length start_sample end_sample
##    <chr>         <int>        <int>      <int>
##  1 CLO               4          122        125
##  2 CLO               5          179        183
##  3 CLO               9          192        200
##  4 CLO               4          202        205
##  5 CLO               7          219        225
##  6 CLO              10          235        244
##  7 CLO              18          247        264
##  8 CLO              17          266        282
##  9 CLO              13          284        296
## 10 CLO              17          298        314
## # ℹ 118 more rows
```

Table 4.1C2: 4 consecutive X bar samples above 2 sigma



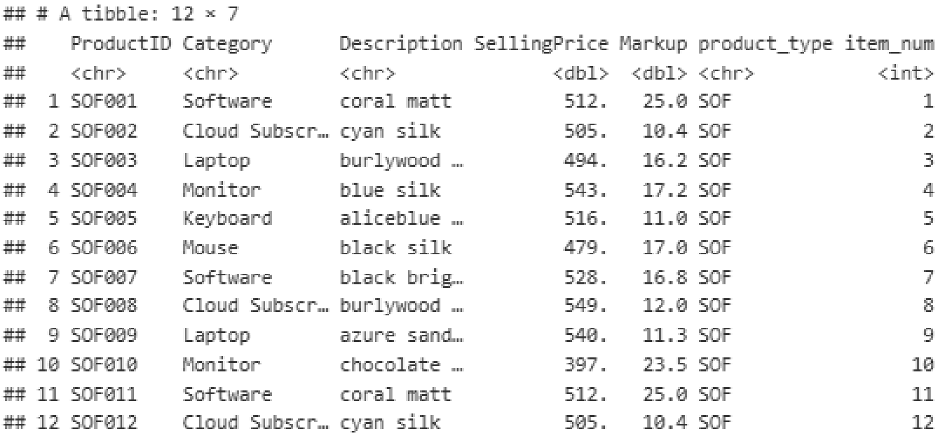
```
## # A tibble: 6 × 2
##   product_type n_runs_ge4
##   <chr>             <int>
## 1 KEY                  25
## 2 SOF                  25
## 3 MON                  23
## 4 MOU                  23
## 5 CLO                  20
## 6 LAP                  12
```

Table 4.2: Type II (consumer’s) error for bottle filling



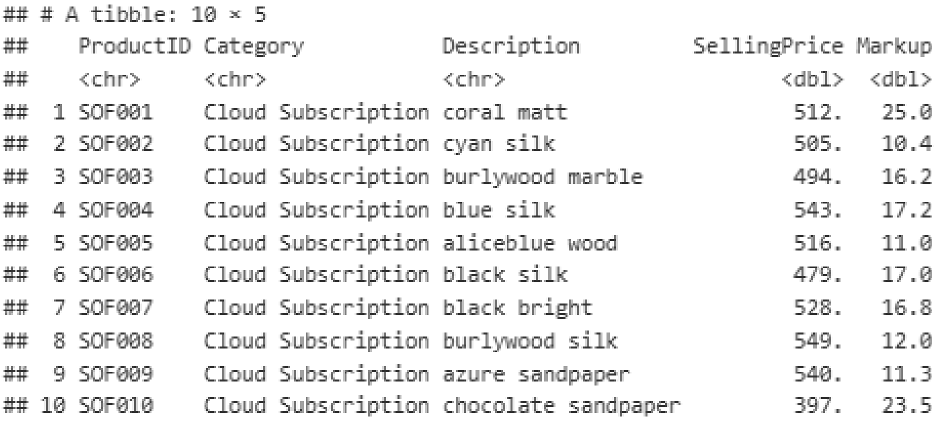
```
## # A tibble: 1 × 6
##     LCL   UCL shifted_mean xbar_sd beta_TypeII power
##   <dbl> <dbl>        <dbl>   <dbl>       <dbl> <dbl>
## 1  25.0  25.1         25.0   0.017       0.841 0.159
```

Table 4.3.2: Rebuild corrected Head Office catalog



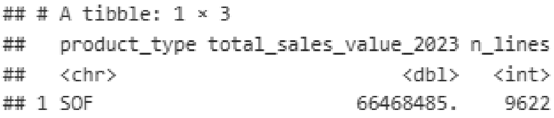
```
## # A tibble: 12 × 7
##    ProductID Category      Description SellingPrice Markup product_type item_num
##    <chr>     <chr>         <chr>              <dbl>  <dbl> <chr>           <int>
##  1 SOF001    Software      coral matt          512.   25.0 SOF                 1
##  2 SOF002    Cloud Subscr… cyan silk           505.   10.4 SOF                 2
##  3 SOF003    Laptop        burlywood …         494.   16.2 SOF                 3
##  4 SOF004    Monitor       blue silk           543.   17.2 SOF                 4
##  5 SOF005    Keyboard      aliceblue …         516.   11.0 SOF                 5
##  6 SOF006    Mouse         black silk          479.   17.0 SOF                 6
##  7 SOF007    Software      black brig…         528.   16.8 SOF                 7
##  8 SOF008    Cloud Subscr… burlywood …         549.   12.0 SOF                 8
##  9 SOF009    Laptop        azure sand…         540.   11.3 SOF                 9
## 10 SOF010    Monitor       chocolate …         397.   23.5 SOF                10
## 11 SOF011    Software      coral matt          512.   25.0 SOF                11
## 12 SOF012    Cloud Subscr… cyan silk           505.   10.4 SOF                12
```

Table 4.3.3: Fix local products_data category to correspond with ProductID



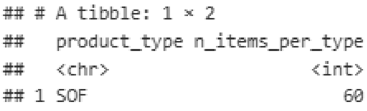
```
## # A tibble: 10 × 5
##    ProductID Category           Description         SellingPrice Markup
##    <chr>     <chr>              <chr>                      <dbl>  <dbl>
##  1 SOF001    Cloud Subscription coral matt                  512.   25.0
##  2 SOF002    Cloud Subscription cyan silk                   505.   10.4
##  3 SOF003    Cloud Subscription burlywood marble            494.   16.2
##  4 SOF004    Cloud Subscription blue silk                   543.   17.2
##  5 SOF005    Cloud Subscription aliceblue wood              516.   11.0
##  6 SOF006    Cloud Subscription black silk                  479.   17.0
##  7 SOF007    Cloud Subscription black bright                528.   16.8
##  8 SOF008    Cloud Subscription burlywood silk              549.   12.0
##  9 SOF009    Cloud Subscription azure sandpaper             540.   11.3
## 10 SOF010    Cloud Subscription chocolate sandpaper         397.   23.5
```

Table 4.3.4: Re-Run week 1 outcome: Total 2023 sales value per product type



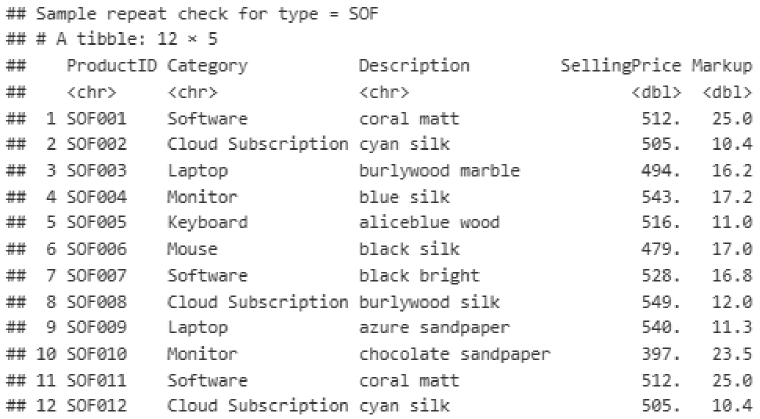
```
## # A tibble: 1 × 3
##   product_type total_sales_value_2023 n_lines
##   <chr>                         <dbl>   <int>
## 1 SOF                       66468485.    9622
```

Table 4.3.5.1: Cross-check and summaries



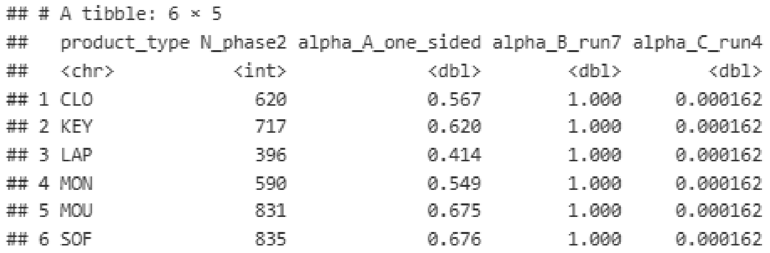
```
## # A tibble: 1 × 2
##   product_type n_items_per_type
##   <chr>                   <int>
## 1 SOF                        60
```

Table 4.3.5.2: Cross-check and summaries



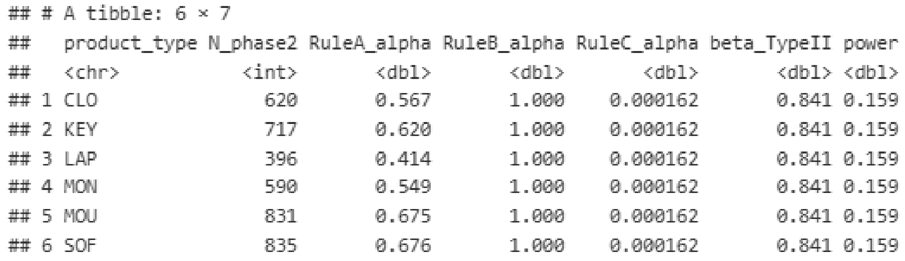
```
## Sample repeat check for type = SOF
## # A tibble: 12 × 5
##    ProductID Category           Description         SellingPrice Markup
##    <chr>     <chr>              <chr>                      <dbl>  <dbl>
##  1 SOF001    Software           coral matt                  512.   25.0
##  2 SOF002    Cloud Subscription cyan silk                   505.   10.4
##  3 SOF003    Laptop             burlywood marble            494.   16.2
##  4 SOF004    Monitor            blue silk                   543.   17.2
##  5 SOF005    Keyboard           aliceblue wood              516.   11.0
##  6 SOF006    Mouse              black silk                  479.   17.0
##  7 SOF007    Software           black bright                528.   16.8
##  8 SOF008    Cloud Subscription burlywood silk              549.   12.0
##  9 SOF009    Laptop             azure sandpaper             540.   11.3
## 10 SOF010    Monitor            chocolate sandpaper         397.   23.5
## 11 SOF011    Software           coral matt                  512.   25.0
## 12 SOF012    Cloud Subscription cyan silk                   505.   10.4
```

Table 4.4: Type I (manufacturer’s) error



```
## # A tibble: 6 × 5
##   product_type N_phase2 alpha_A_one_sided alpha_B_run7 alpha_C_run4
##   <chr>           <int>             <dbl>        <dbl>        <dbl>
## 1 CLO               620             0.567        1.000     0.000162
## 2 KEY               717             0.620        1.000     0.000162
## 3 LAP               396             0.414        1.000     0.000162
## 4 MON               590             0.549        1.000     0.000162
## 5 MOU               831             0.675        1.000     0.000162
## 6 SOF               835             0.676        1.000     0.000162
```

Table 4.5: Comparison summary of type I vs type II

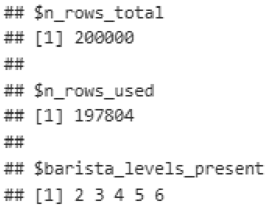


```
## # A tibble: 6 × 7
##   product_type N_phase2 RuleA_alpha RuleB_alpha RuleC_alpha beta_TypeII power
##   <chr>           <int>       <dbl>       <dbl>       <dbl>       <dbl> <dbl>
## 1 CLO               620       0.567       1.000    0.000162       0.841 0.159
## 2 KEY               717       0.620       1.000    0.000162       0.841 0.159
## 3 LAP               396       0.414       1.000    0.000162       0.841 0.159
## 4 MON               590       0.549       1.000    0.000162       0.841 0.159
## 5 MOU               831       0.675       1.000    0.000162       0.841 0.159
## 6 SOF               835       0.676       1.000    0.000162       0.841 0.159
```

APPENDIX D:

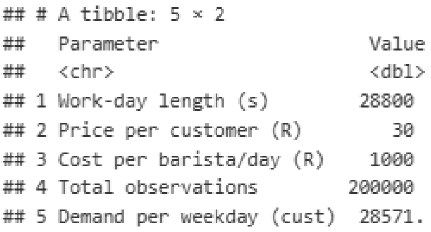
D1: Tables

Table 5.1: Data preparation



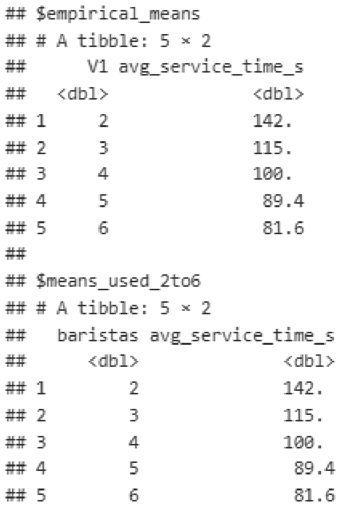
```
## $n_rows_total
## [1] 200000
##
## $n_rows_used
## [1] 197804
##
## $barista_levels_present
## [1] 2 3 4 5 6
```

Table 5.2: Model parameters



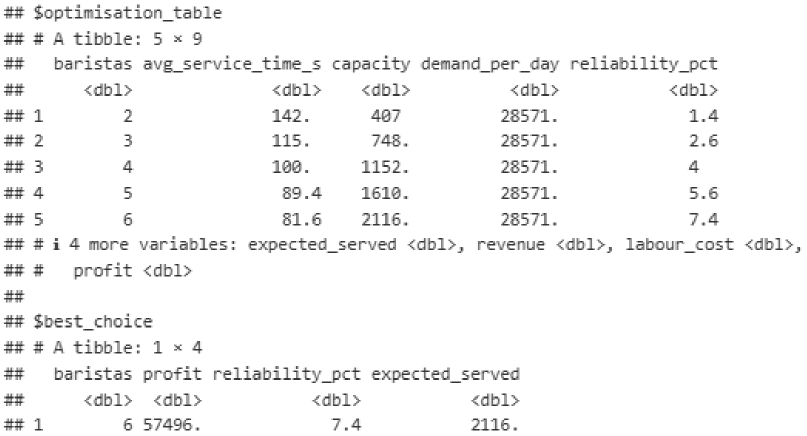
```
## # A tibble: 5 × 2
##   Parameter                  Value
##   <chr>                      <dbl>
## 1 Work-day length (s)        28800
## 2 Price per customer (R)        30
## 3 Cost per barista/day (R)    1000
## 4 Total observations        200000
## 5 Demand per weekday (cust)  28571.
```

Table 5.3: Average service time by barista level



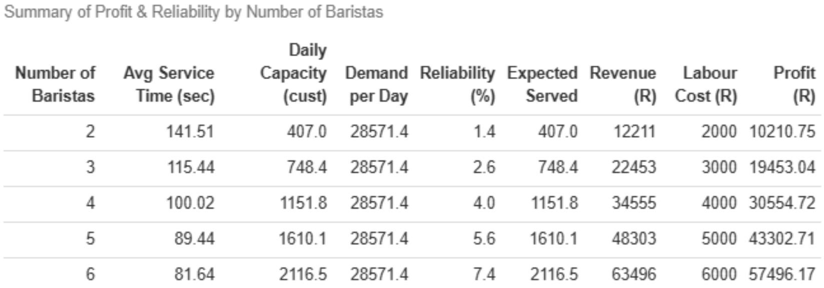
```
## $empirical_means
## # A tibble: 5 × 2
##      V1 avg_service_time_s
##   <dbl>              <dbl>
## 1     2              142.
## 2     3              115.
## 3     4              100.
## 4     5               89.4
## 5     6               81.6
##
## $means_used_2to6
## # A tibble: 5 × 2
##   baristas avg_service_time_s
##      <dbl>              <dbl>
## 1        2              142.
## 2        3              115.
## 3        4              100.
## 4        5               89.4
## 5        6               81.6
```

Table 5.4: Results: Optimal staffing (Generic day)



```
## $optimisation_table
## # A tibble: 5 × 9
##   baristas avg_service_time_s capacity demand_per_day reliability_pct
##      <dbl>              <dbl>    <dbl>          <dbl>           <dbl>
## 1        2              142.      407          28571.             1.4
## 2        3              115.      748.         28571.             2.6
## 3        4              100.     1152.         28571.             4
## 4        5               89.4    1610.         28571.             5.6
## 5        6               81.6    2116.         28571.             7.4
## # ℹ 4 more variables: expected_served <dbl>, revenue <dbl>, labour_cost <dbl>,
## #   profit <dbl>
##
## $best_choice
## # A tibble: 1 × 4
##   baristas profit reliability_pct expected_served
##      <dbl>  <dbl>           <dbl>           <dbl>
## 1        6 57496.             7.4           2116.
```

Table 5.6: Summary



Summary of Profit & Reliability by Number of Baristas

Number of Baristas	Avg Service Time (sec)	Daily Capacity (cust)	Demand per Day	Reliability (%)	Expected Served	Revenue (R)	Labour Cost (R)	Profit (R)
2	141.51	407.0	28571.4	1.4	407.0	12211	2000	10210.75
3	115.44	748.4	28571.4	2.6	748.4	22453	3000	19453.04
4	100.02	1151.8	28571.4	4.0	1151.8	34555	4000	30554.72
5	89.44	1610.1	28571.4	5.6	1610.1	48303	5000	43302.71
6	81.64	2116.5	28571.4	7.4	2116.5	63496	6000	57496.17

D2: Figures

Figure 5.5.1: Visualization of results: Profit by Weekday

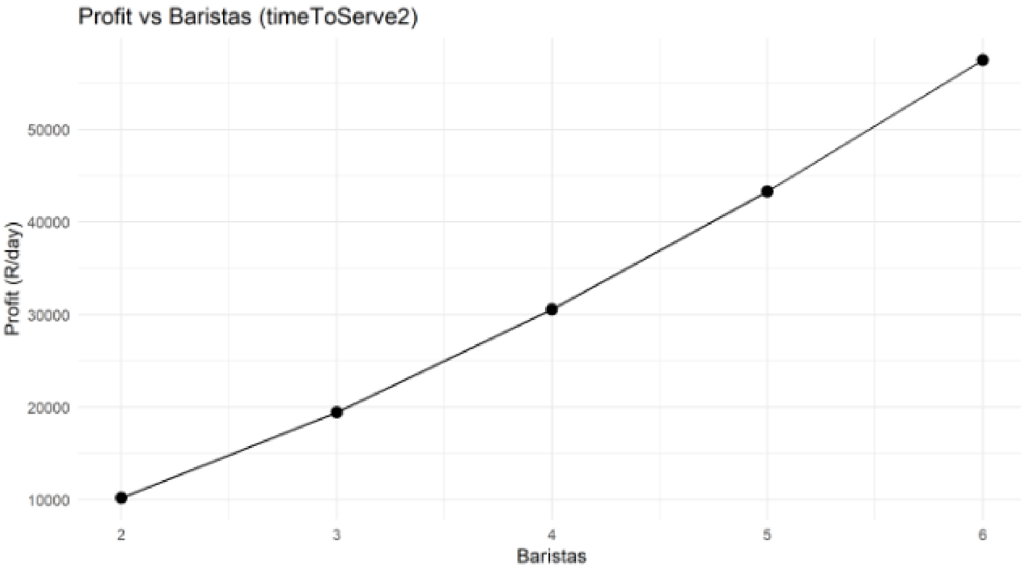
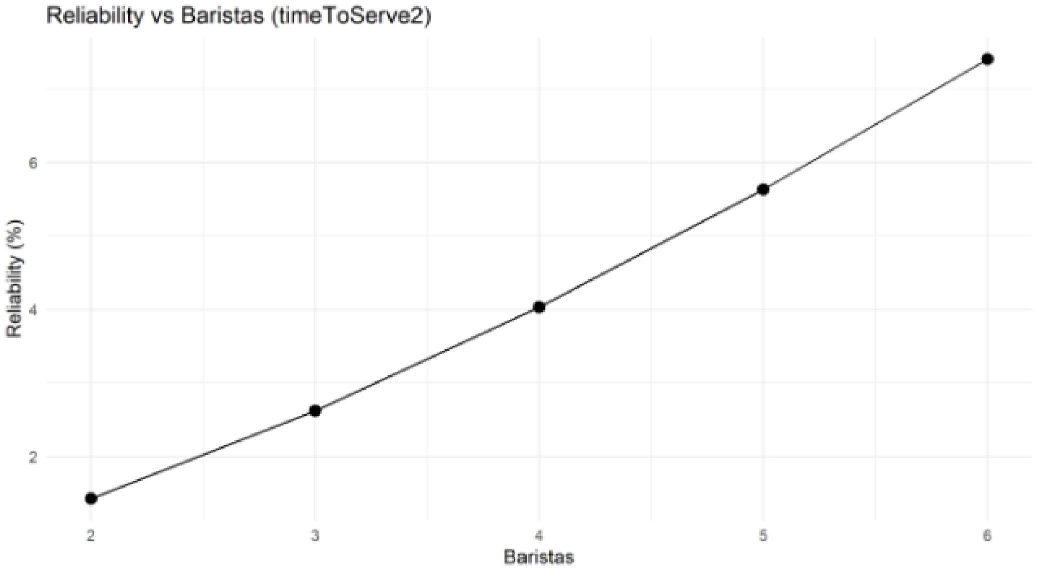


Figure 5.5.2: Reliability vs Baristas


Reliability vs Baristas (timeToServe2)
Reliability (%)
2
4
6
Baristas
2
3
4
5
6

APPENDIX E:

E1: Figures

Figure 6.1: Q-Q plot

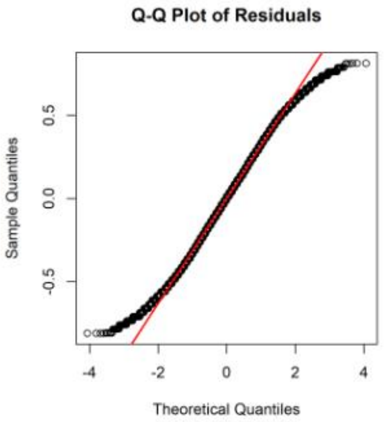
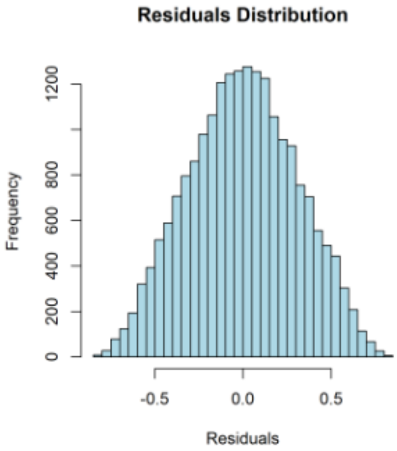


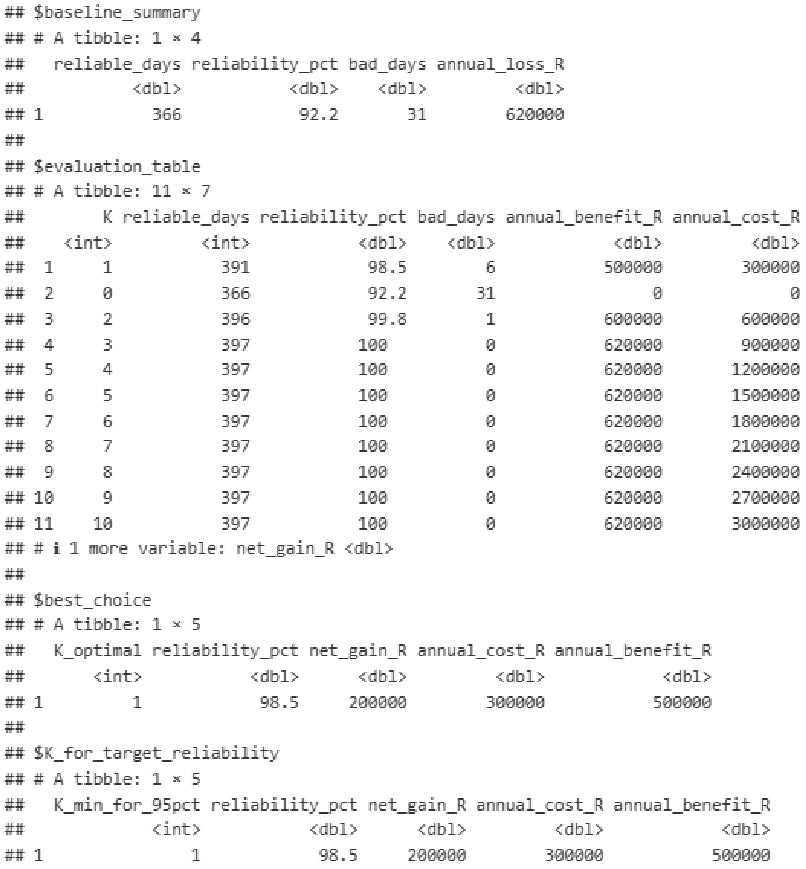
Figure 6.2: Residuals distribution



APPENDIX F:

F1: Tables

Table 7: Reliability and profit optimization (car rental)



```
## $baseline_summary
## # A tibble: 1 × 4
##   reliable_days reliability_pct bad_days annual_loss_R
##           <dbl>           <dbl>    <dbl>         <dbl>
## 1           366            92.2       31        620000
##
## $evaluation_table
## # A tibble: 11 × 7
##        K reliable_days reliability_pct bad_days annual_benefit_R annual_cost_R
##    <int>         <int>           <dbl>    <dbl>            <dbl>         <dbl>
##  1     1           391            98.5        6           500000        300000
##  2     0           366            92.2       31                0             0
##  3     2           396            99.8        1           600000        600000
##  4     3           397           100          0           620000        900000
##  5     4           397           100          0           620000       1200000
##  6     5           397           100          0           620000       1500000
##  7     6           397           100          0           620000       1800000
##  8     7           397           100          0           620000       2100000
##  9     8           397           100          0           620000       2400000
## 10     9           397           100          0           620000       2700000
## 11    10           397           100          0           620000       3000000
## # ℹ 1 more variable: net_gain_R <dbl>
##
## $best_choice
## # A tibble: 1 × 5
##   K_optimal reliability_pct net_gain_R annual_cost_R annual_benefit_R
##       <int>           <dbl>      <dbl>         <dbl>            <dbl>
## 1         1            98.5     200000        300000           500000
##
## $K_for_target_reliability
## # A tibble: 1 × 5
##   K_min_for_95pct reliability_pct net_gain_R annual_cost_R annual_benefit_R
##             <int>           <dbl>      <dbl>         <dbl>            <dbl>
## 1               1            98.5     200000        300000           500000
```

F2: Figures

Figure 7.1: Reliability vs. added personnel

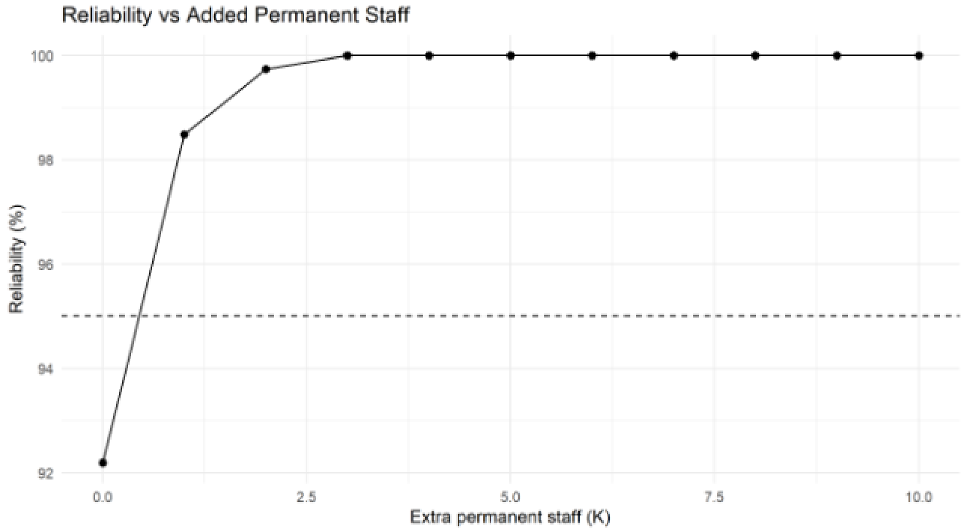
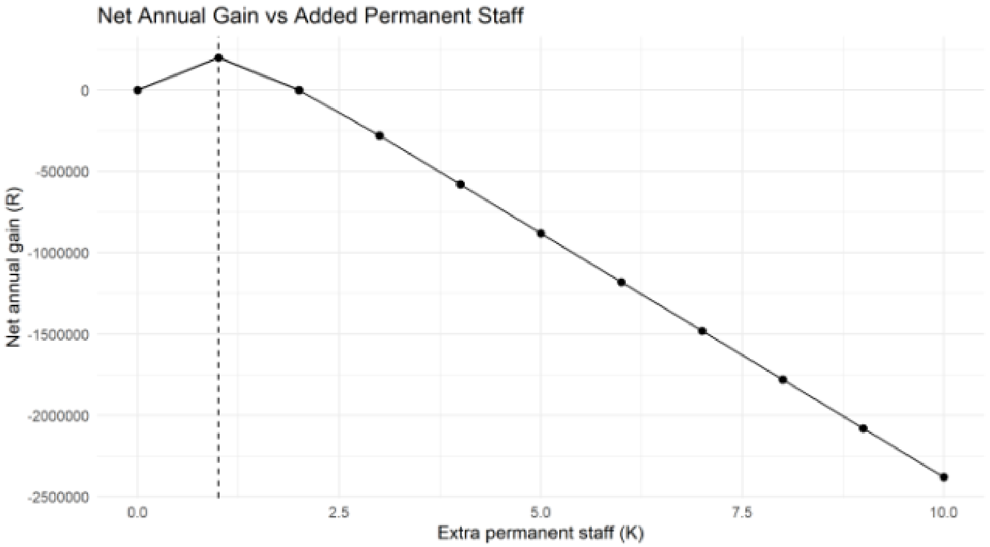


Figure 7.2: Net annual gain vs. added personnel


Net Annual Gain vs Added Permanent Staff
Net annual gain (R)
0
-500000
-1000000
-1500000
-2000000
-2500000
0.0
2.5
5.0
7.5
10.0
Extra permanent staff (K)