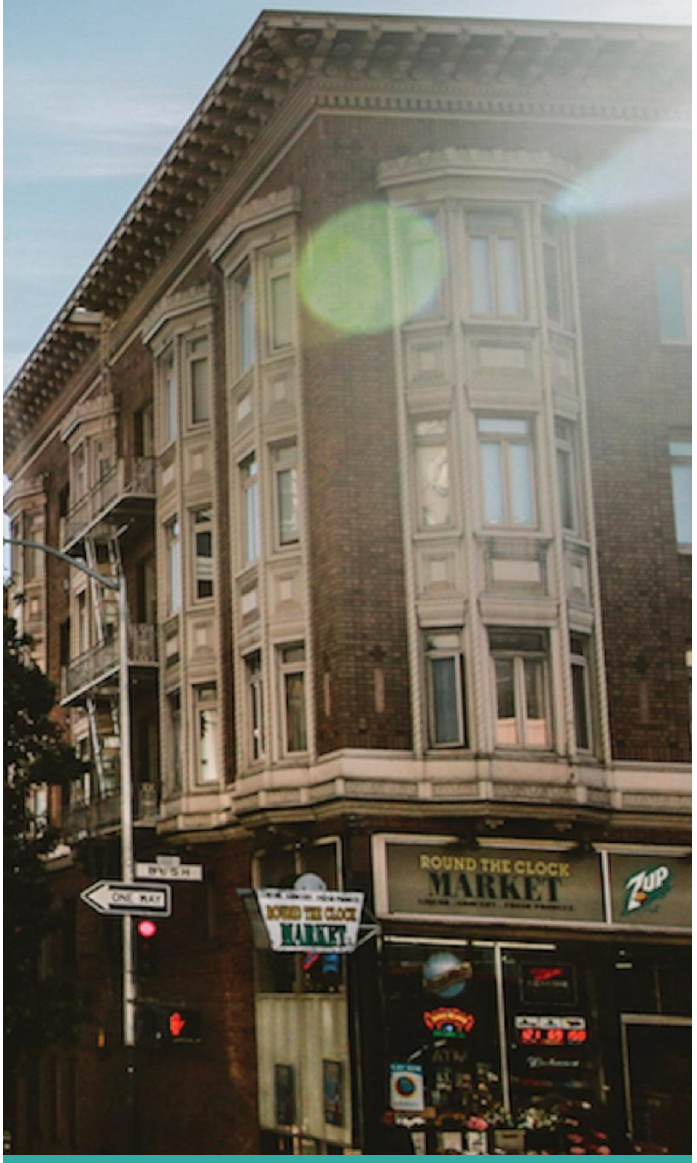
ECSA GA4 Project Report



OCTOBER 22

Stellenbosch University
Department: Industrial Engineering
Authored by: Nicolaas Retief Esterhuysen
Student nr. 26889463

Table of Contents:

Contents

1.Introduction

The purpose of this report is to demonstrate competence in Graduate Attribute 4 (ECSA GA4) as required by the Engineering Council of South Africa (ECSA). This attribute assesses a student's ability to investigate complex engineering problems through data analysis, interpretation, and evidence-based reasoning. The report forms part of the *Quality Assurance (QA344)* module and integrates statistical, computational, and decision-making techniques commonly used in industrial engineering practice.

The project focuses on applying data-driven quality assurance and process optimisation methods to real-world business and production data. Using a range of datasets-including *customers.csv*, *products_data.csv*, *sales2022and2023.csv*, *sales2026and2027Future.csv*, *timeToServe.csv*, and *timeToServe2.csv*-the report develops models and analyses that support decision-making in service reliability, product quality, and process performance.

The tasks are structured according to the six key components outlined in the project document.

- Part 1 performs descriptive statistical analysis to understand underlying data patterns and relationships.
- Part 3 applies Statistical Process Control (SPC) methods to monitor delivery processes and calculate process capability indices (Cp, Cpk), ensuring that operations meet defined specification limits.
- Part 4 addresses Type I and Type II errors and investigates data correction to maintain data integrity.
- Part 5 develops a profit optimisation model using service time data to balance operational cost and service reliability.
- Part 6 explores Design of Experiments (DOE) and ANOVA/MANOVA methods to determine whether significant differences exist between different datasets or process conditions.
- Part 7, where applicable, extends the analysis to reliability and personnel optimisation in service industries.

Each section integrates both theoretical and computational approaches, primarily using R programming for statistical analysis and process simulation. The report concludes with a comprehensive discussion linking the quantitative results to industrial quality assurance principles, operational efficiency, and customer satisfaction. Through this work, the outcomes of ECSA GA4 are demonstrated by showing the ability to identify, formulate, and solve engineering problems through data-centred reasoning and quantitative evaluation.

2. Descriptive Statistics (Part 1.2)

Customer data:

Data Overview

The customer dataset consists of 5,000 observations and 5 variables: CustomerID, Gender, Age, Income, and City. All columns are complete, with no missing values.

- Categorical Variables: **CustomerID (unique identifier), Gender (Male, Female, Other), City (7 unique cities).**
- Numeric Variables: **Age (16 to 105), Income (R5,000 to R140,000).**

```
##    CustomerID      Gender          Age             Income
##  CUST001:   1   Female:2432   Min.   : 16.00   Min.   :  5000
##  CUST002:   1   Male  :2350   1st Qu.: 33.00   1st Qu.: 55000
##  CUST003:   1   Other : 218   Median : 51.00   Median : 85000
##  CUST004:   1                 Mean   : 51.55   Mean   : 80797
##  CUST005:   1                 3rd Qu.: 68.00   3rd Qu.:105000
##  CUST006:   1                 Max.   :105.00   Max.   :140000
##  (Other):4994
##            City
##  Chicago      :724
##  Houston      :724
##  Los Angeles  :726
##  Miami        :647
##  New York     :726
##  San Francisco:780
##  Seattle      :673
```

Figure 1 Data Summary Customers 1.0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
CustomerID	0	1	FALSE	5000	CUS: 1, CUS: 1, CUS: 1, CUS: 1
Gender	0	1	FALSE	3	Fem: 2432, Mal: 2350, Oth: 218
City	0	1	FALSE	7	San: 780, Los: 726, New: 726, Chi: 724

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 2 Data Summary Customers 2.0

Descriptive Statistics

Age

- Mean: 51.6 years
- Median: 51 years
- Standard Deviation: 21.2 years
- Range: 16 to 105 years

Income

- Mean: R80,797
- Median: R85,000
- Standard Deviation: R33,150

- Range: R5,000 to R140,000

There is no missing data for either Age or Income. The close alignment between mean and median for Age and Income suggests a relatively symmetric distribution, although the income variable shows some negative skewness (-0.21), indicating a longer tail on the lower income side.

```
##        vars    n     mean       sd median  trimmed      mad  min    max  range
## Age       1 5000    51.55    21.22     51    50.88    26.69   16    105     89
## Income    2 5000 80797.00 33150.11  85000 81665.00 37065.00 5000 140000 135000
##         skew kurtosis     se
## Age     0.20    -0.99   0.30
## Income -0.21    -0.75 468.81
```

Figure 3 Descriptive Statistics Customers

Distributions & Visualizations

Age Distribution:

The histogram of Age (Image 4) shows an approximately uniform distribution with a slight concentration in the 25–65 age range, and fewer customers at the extremes (younger than 25 and older than 90). The standard deviation of 21.2 years reflects a wide spread in ages, indicating the company serves a diverse age demographic.

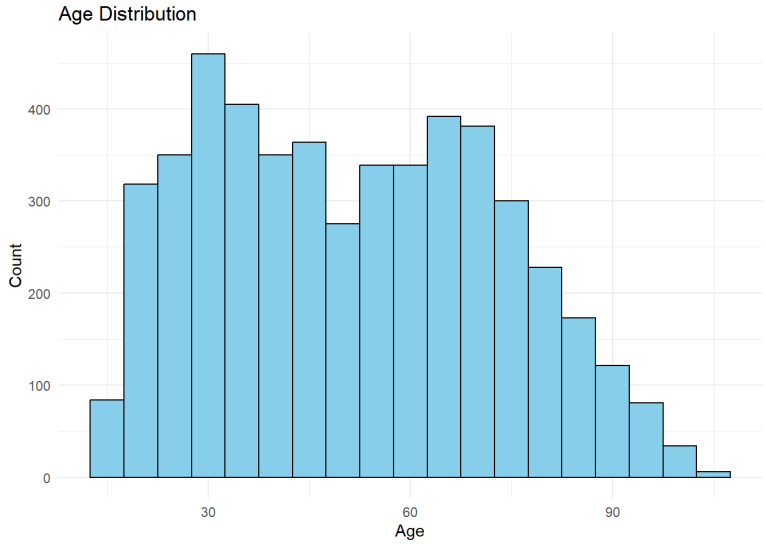


Figure 4 Customer Age Distribution

Income Distribution:

The histogram of Income (see Image 5) shows a roughly symmetric, bell-shaped curve, centered between R50,000 and R110,000, with most customers' incomes falling within this range. There are fewer customers at the very low and very high ends of the income spectrum. The distribution is nearly normal but with a slight negative skew, as supported by the skewness statistic.

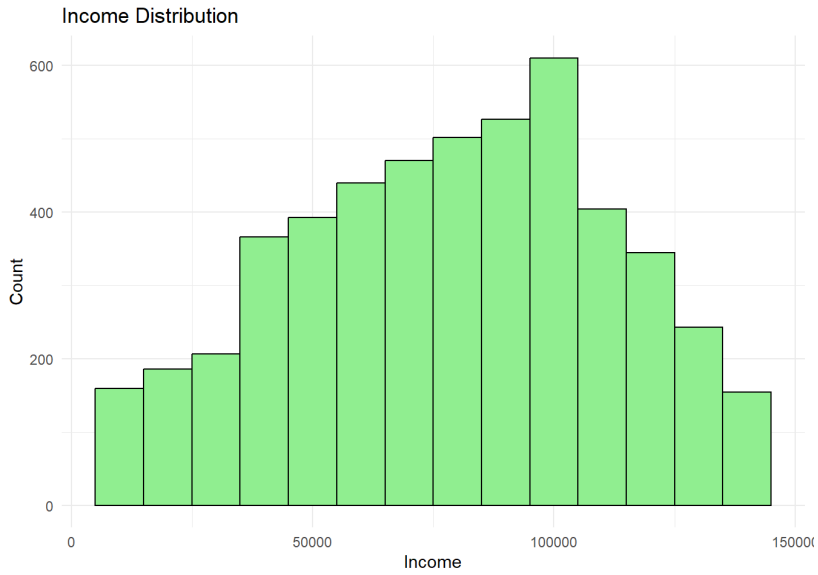


Figure 5 Income Distribution Customers

Income by Gender:

The boxplot (see Image 6) demonstrates that income distributions are similar across genders (Female, Male, Other), with comparable median incomes and ranges. There is substantial spread in each category, and no significant gender-based income disparities are apparent.

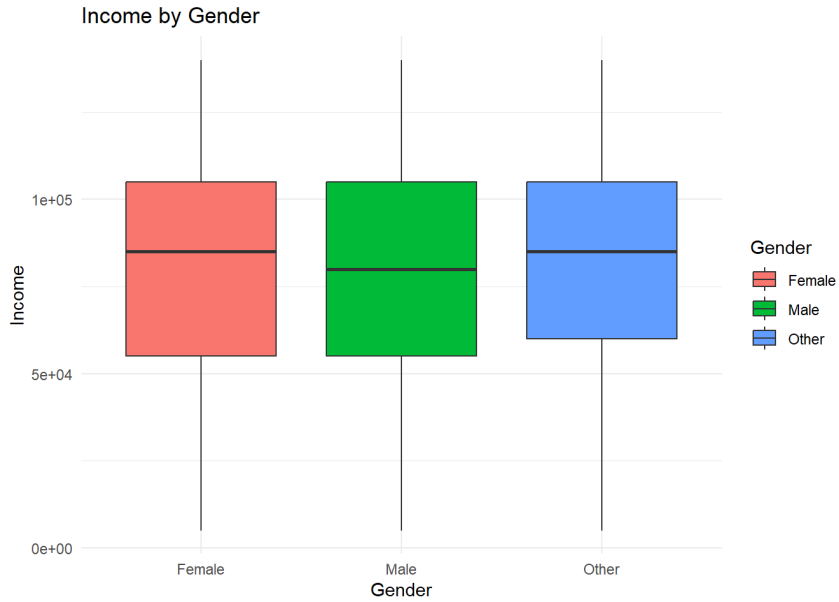


Figure 6 Income by Gender Customer Data

Age vs Income by Gender:

The scatterplot (see Image 7) shows that age and income are relatively evenly distributed across all genders. There is no visible clustering or trend indicating a strong relationship between age and income.

Figure 7 Age vs Income by Gender Customer Data

Correlations and Trends:

The correlation matrix (see Image 8) shows a weak positive correlation (r = 0.16) between Age and Income. This suggests that, on average, older customers have slightly higher incomes, but the relationship is not strong.

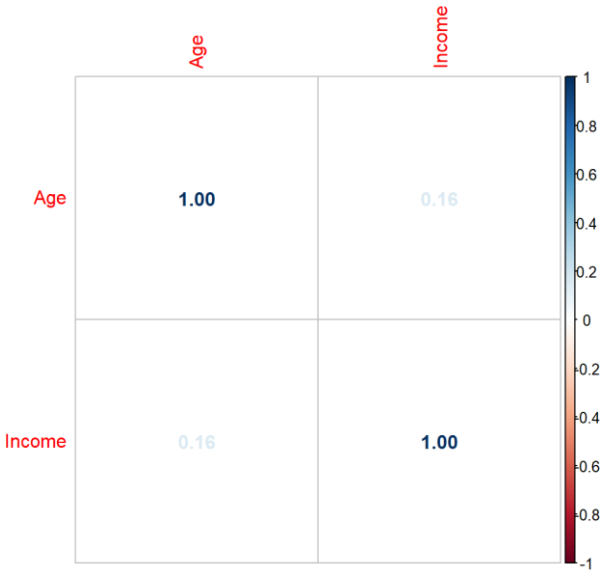


Figure 8 Correlation and Trends Customer Data

Conclusion

The customer dataset is diverse and complete, with balanced representation across key demographics. The distributions and descriptive statistics suggest the company serves a wide range of ages and incomes, with no major anomalies or missing values. Income is nearly normally distributed, and there are no significant differences in income by gender. Age and income are only weakly correlated. Overall, the data shows random variation and a broad spread, providing a strong foundation for further customer segmentation and targeted analysis.

Product data:

Data Overview

The products dataset contains 60 observations across 5 variables: ProductID, Category, Description, SellingPrice, and Markup. All columns are complete, and product categories are relatively evenly represented.

```
##    ProductID               Category          Description  SellingPrice
##  CLO011 : 1   Cloud Subscription:10   chocolate silk : 5   Min.   :  350.4
##  CLO012 : 1   Keyboard          :10   aliceblue silk : 4   1st Qu.:  512.2
##  CLO013 : 1   Laptop            :10   azure silk     : 4   Median :  794.2
##  CLO014 : 1   Monitor           :10   azure sandpaper: 3   Mean   : 4493.6
##  CLO015 : 1   Mouse             :10   blue silk      : 3   3rd Qu.: 6416.7
##  CLO016 : 1   Software          :10   burlywood silk : 3   Max.   :19725.2
##  (Other):54                           (Other)        :38
##      Markup
##  Min.   :10.13
##  1st Qu.:16.14
##  Median :20.34
##  Mean   :20.46
##  3rd Qu.:25.71
##  Max.   :29.84
##
```

Figure 9 Products Data Summary 1.0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
ProductID	0	1	FALSE	60	CLO: 1, CLO: 1, CLO: 1, CLO: 1
Category	0	1	FALSE	6	Clo: 10, Key: 10, Lap: 10, Mon: 10
Description	0	1	FALSE	35	cho: 5, ali: 4, azu: 4, azu: 3

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	

Figure 10 Products Data Summary 2.0

Descriptive Statistics

SellingPrice

- Mean: **R4,493.59**
- Median: **R794.18**
- Standard Deviation: **R6,503.77**
- Range: **R350.40 to R19,725.18**
- Skew: **1.43 (strong positive skew)**

Markup

- Mean: **R20.46**
- Median: **R20.34**
- Standard Deviation: **R6.07**
- Range: **R10.13 to R29.84**
- Skew: **-0.04 (very close to zero, nearly symmetric)**

No missing data is present. The selling price distribution is highly positively skewed, as the mean is much higher than the median and the maximum value is far above the upper quartile. Markup is tightly clustered around its mean and median, with very little skewness.

```
##              vars  n    mean      sd median trimmed    mad    min      max
## SellingPrice    1 60 4493.59 6503.77 794.18 3189.25 525.72 350.45 19725.18
## Markup          2 60   20.46    6.07  20.34   20.51   7.31  10.13    29.84
##                 range  skew kurtosis     se
## SellingPrice 19374.73  1.43     0.43 839.63
## Markup          19.71 -0.04    -1.24   0.78
```

Distributions & Visualizations

Selling Price Distribution

The histogram of SellingPrice (see Image 13) demonstrates a highly right-skewed distribution, with most products concentrated at the lower end of the price range (under R2,000) and a few products with very high selling prices stretching the distribution. This is consistent with the descriptive statistics, which show a mean much higher than the median and a large standard deviation.

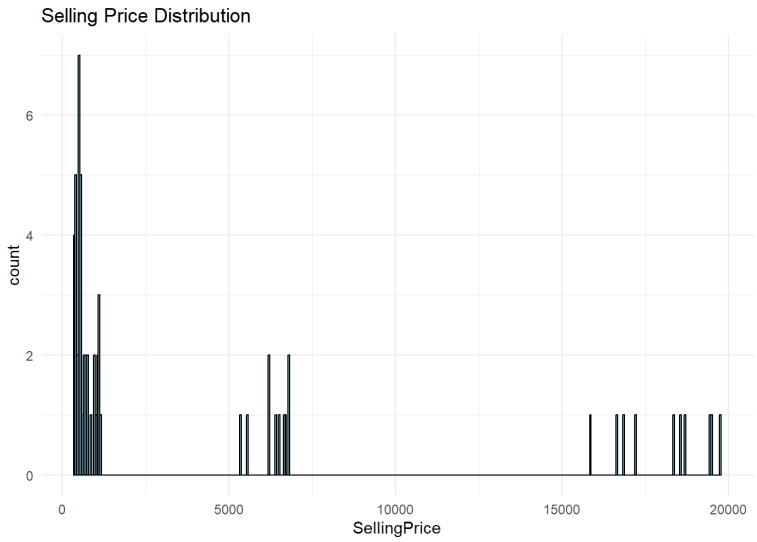


Figure 11 Selling price Distribution Products Data

Markup Distribution

The histogram of Markup shows a fairly uniform spread across the range, with most products falling between R10 and R30. This supports the finding of a near-normal distribution with a tight spread.

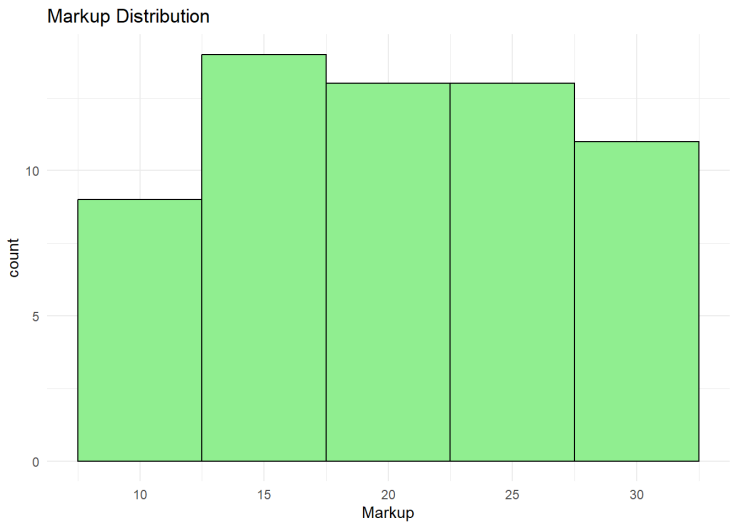


Figure 12 Markup Distribution Products Data

Selling Price by Category

The boxplot reveals that most categories have similar selling price distributions, with median prices clustering below R2,000. However, there are several high-priced outliers in each category, as indicated by dots above the whiskers. These outliers contribute to the strong positive skew and high standard deviation seen in the summary statistics.

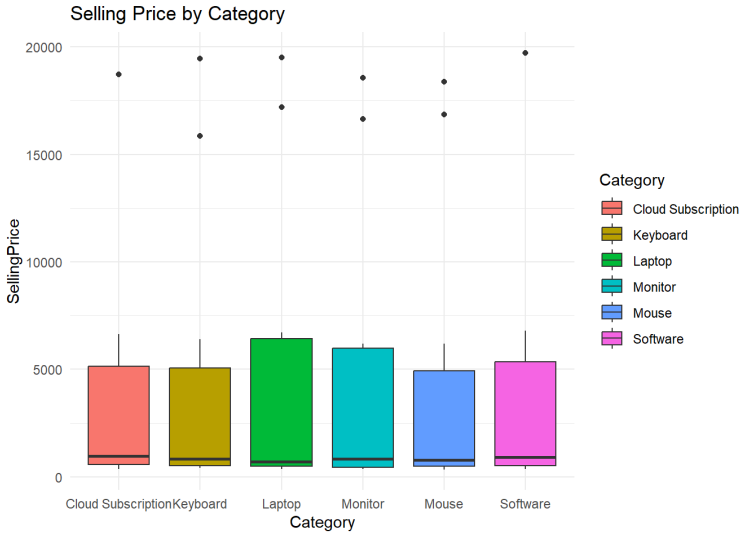


Figure 13 Selling price by category Products Data

Selling Price vs Markup

The scatterplot demonstrates that there is little to no relationship between SellingPrice and Markup across categories. Products with high selling prices do not necessarily have higher or lower markups. The points are scattered with no clear pattern or trend.

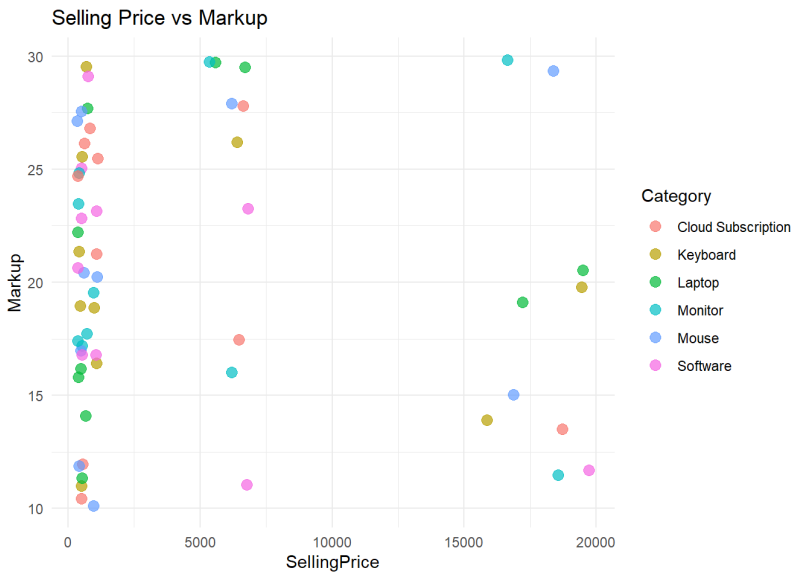


Figure 14 Selling price vs Markup Products Data

Correlations and Trends

The correlation matrix a weak negative correlation (-0.08) between SellingPrice and Markup, confirming the lack of a meaningful relationship between product price and markup. This indicates that markup strategy is consistent across products, regardless of price.

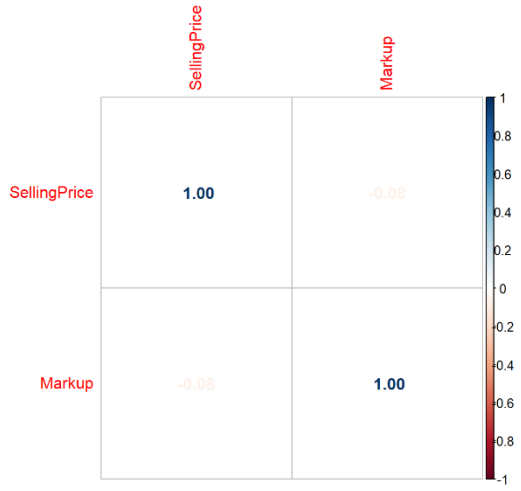


Figure 15 Correlations and Trends Dproducts Data

Conclusion

The products dataset shows strong variation in SellingPrice due to high-priced outliers, while Markup remains consistent across all items and categories. There is no significant relationship between a product's price and its markup, suggesting that markup policies are not dependent on product price. The complete data and balanced categories provide a solid basis for further product profitability and pricing analysis.

Sales data:

Data Overview

The sales dataset contains 100,000 records and includes variables: CustomerID, ProductID, Quantity, orderTime, orderDay, orderMonth, orderYear, pickingHours, and deliveryHours. All columns are complete, with no missing values.

```
##     CustomerID      ProductID       Quantity      orderTime       orderDay
##  CUST1193:  326   MOU057 : 2119   Min.   : 1.0   Min.   : 1.00   Min.   : 1.0
##  CUST1791:  322   MOU059 : 2118   1st Qu.: 3.0   1st Qu.: 9.00   1st Qu.: 8.0
##  CUST596 :  319   SOF007 : 2118   Median : 6.0   Median :13.00   Median :15.0
##  CUST2527:  303   MOU054 : 2116   Mean   :13.5   Mean   :12.93   Mean   :15.5
##  CUST3721:  301   SOF005 : 2115   3rd Qu.:23.0   3rd Qu.:17.00   3rd Qu.:23.0
##  CUST2277:  298   SOF006 : 2107   Max.   :50.0   Max.   :23.00   Max.   :30.0
##  (Other) :98131   (Other):87307
##    orderMonth       orderYear     pickingHours     deliveryHours
##  Min.   : 1.000   Min.   :2022   Min.   : 0.4259   Min.   : 0.2772
##  1st Qu.: 4.000   1st Qu.:2022   1st Qu.: 9.3908   1st Qu.:11.5460
##  Median : 6.000   Median :2022   Median :14.0550   Median :19.5460
##  Mean   : 6.448   Mean   :2022   Mean   :14.6955   Mean   :17.4765
##  3rd Qu.: 9.000   3rd Qu.:2023   3rd Qu.:18.7217   3rd Qu.:25.0440
##  Max.   :12.000   Max.   :2023   Max.   :45.0575   Max.   :38.0460
##
```

Figure 16 Data Summary Sales Data 1.0

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	

Figure 17 Data Summary Sales Data 2.0

Descriptive Statistics

Quantity

- Mean: **13.5 units**
- Median: **6 units**
- Standard Deviation: **13.8 units**
- Range: **1 to 50 units**
- Skew: **-0.22 (slightly left-skewed)**

Operational Time Variables

- PickingHours: **Mean 14.7, SD 10.4, Range 0.43–45.06, Skew 0.74**
- DeliveryHours: **Mean 17.5, SD 10.0, Range 0.28–38.05, Skew -0.87**

Other Variables

- orderTime: **Mean 12.9, SD 5.5, Range 1–23**
- orderDay: **Mean 15.5, SD 8.7, Range 1–30**
- orderMonth: **Mean 6.45, SD 3.3, Range 1–12**
- orderYear: **2022/2023**

```
##              vars     n    mean    sd  median trimmed   mad     min     max
## Quantity        1 1e+05   13.50 13.76    6.00   11.46  5.93    1.00   50.00
## orderTime       2 1e+05   12.93  5.50   13.00   13.12  5.93    1.00   23.00
## orderDay        3 1e+05   15.50  8.65   15.00   15.50 10.38    1.00   30.00
## orderMonth      4 1e+05    6.45  3.28    6.00    6.45  4.45    1.00   12.00
## orderYear       5 1e+05 2022.46  0.50 2022.00 2022.45  0.00 2022.00 2023.00
## pickingHours    6 1e+05   14.70 10.39   14.05   13.54  6.92    0.43   45.06
## deliveryHours   7 1e+05   17.48 10.00   19.55   17.78  8.90    0.28   38.05
##              range  skew kurtosis   se
## Quantity     49.00  1.04    -0.22 0.04
## orderTime    22.00 -0.23    -0.71 0.02
## orderDay     29.00  0.00    -1.20 0.03
## orderMonth   11.00  0.01    -1.18 0.01
## orderYear     1.00  0.15    -1.98 0.00
## pickingHours 44.63  0.74     0.41 0.03
## deliveryHours 37.77 -0.47    -0.87 0.03
```

Figure 18 Descriptive Statistics Sales Data

Distributions & Visualizations

Quantity Distribution

The histogram shows a left-skewed distribution: most orders are small (1–10 units), with fewer high-quantity orders. The spread is substantial, as reflected by the standard deviation.

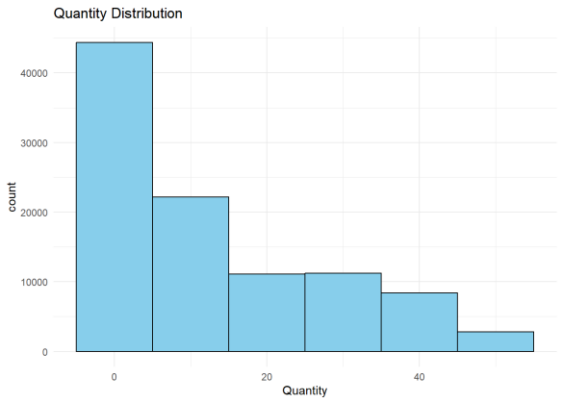
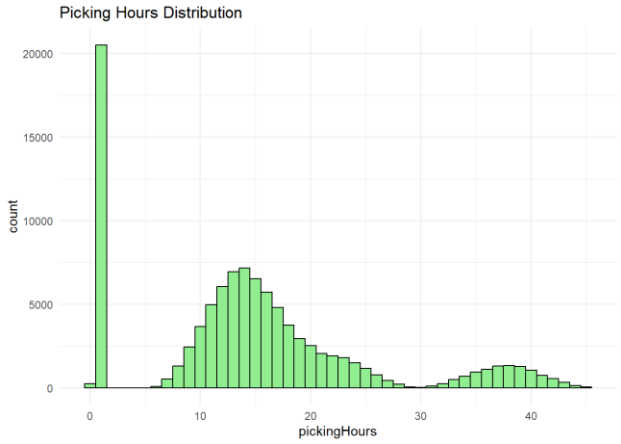


Figure 19 Quantity Distribution Sales data

Picking and Delivery Hours Distribution

- PickingHours : **Shows a multi-modal distribution with a large spike at very low picking hours, then a broad spread across higher values, and a smaller mode near the upper end. This spread indicates variability in operational efficiency or order complexity.**
- DeliveryHours : **Similar to picking hours, with a large spike at low values and a broad, nearly normal distribution across the rest. This suggests some deliveries are extremely efficient, while others require considerably more time.**



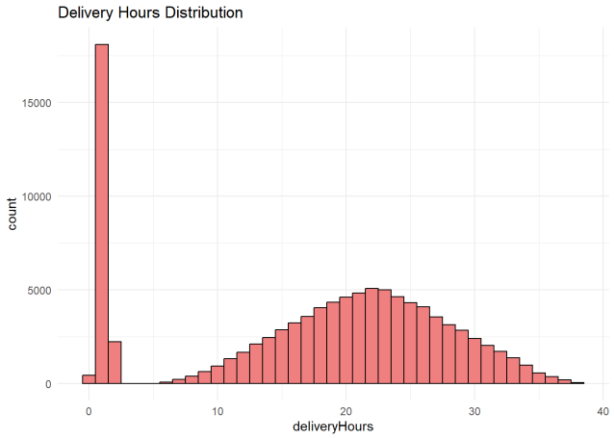


Figure 20 Picking hours distribution Sales data

Figure 21 Delivery hours distribution Sales Data

Quantity by Year

The boxplot shows that the distribution of quantities ordered is consistent across years (2022 and 2023), with median values and spreads nearly identical. Both years contain a wide range of order sizes.

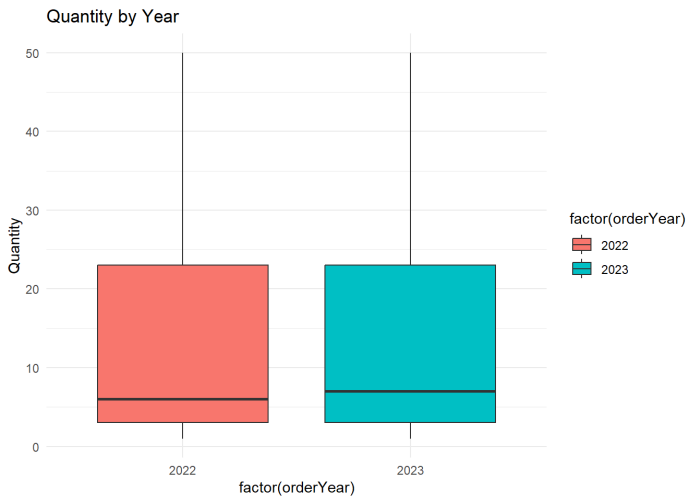


Figure 22 Quantity by year Sales Data

Picking vs Delivery Hours

The scatterplot displays a broad, random spread of picking and delivery hours, with no clear trend or correlation. Both years show similar operational patterns, indicating consistent logistical processes.

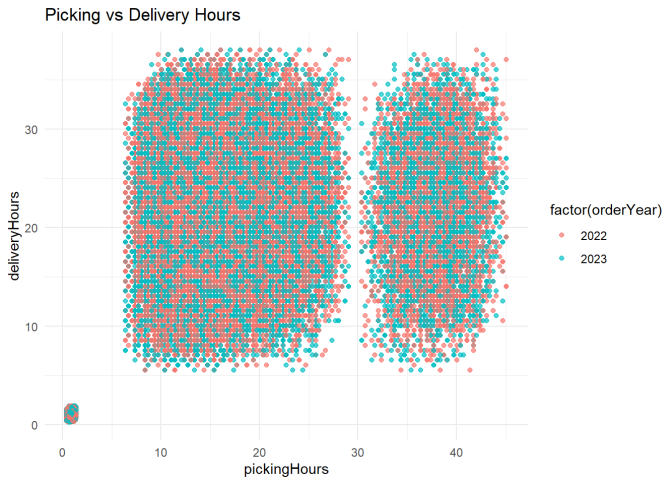


Figure 23 Picking vs Deliovery hours Sales Data

Correlations and Trends

The correlation matrix reveals:

- Strong positive correlation (r = 0.58) **between pickingHours and deliveryHours, indicating that orders which take longer to pick also tend to take longer for delivery.**
- Weak correlations (r = 0.09, 0.13) **between operational times and orderMonth, suggesting little seasonal or monthly influence on picking/delivery duration.**
- No significant correlation **between quantity and other variables, indicating order size does not strongly influence processing or delivery times.**

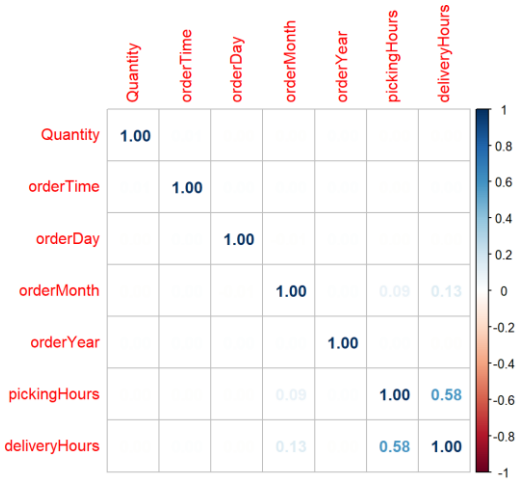


Figure 24 Correlation nad trends Sales Data

Conclusion

The sales dataset demonstrates high variation in order quantities and operational times, with most sales being small but a notable tail of larger orders. Picking and delivery hours vary widely, reflecting the complexity of sales and logistics. The strong correlation between picking and delivery hours suggests linked operational processes. All fields are complete, and there are no major anomalies or gaps. Trends are consistent across years, indicating stable processes and business volume.

3.Statistical Process Control (SPC) – Parts 3.1 to 3.4

3.1 SPC Concept and Purpose

Statistical Process Control (SPC) is a data-driven methodology for monitoring and controlling a process using statistical tools. SPC helps distinguish between normal (common cause) and abnormal (special cause) variation, enabling proactive quality assurance. Its main goals are to maintain process stability, prevent defects, and ensure product quality meets customer expectations (Voice of Customer, VOC).

3.2 Data Preparation and Overview

The analysis begins by importing the sales data file, which contains 100,000 observations and 9 variables (CustomerID, ProductID, Quantity, orderTime, orderDay, orderMonth, orderYear, pickingHours, deliveryHours).

```
## 'data.frame':    100000 obs. of  9 variables:
##  $ CustomerID   : chr  "CUST1791" "CUST3172" "CUST1022" "CUST3721" ...
##  $ ProductID    : chr  "CLO011" "LAP026" "KEY046" "LAP024" ...
##  $ Quantity     : int  16 17 11 31 20 32 29 1 10 1 ...
##  $ orderTime    : int  13 17 16 12 14 21 5 19 19 18 ...
##  $ orderDay     : int  11 14 23 18 7 24 23 9 13 30 ...
##  $ orderMonth   : int  11 7 5 7 2 12 1 6 12 4 ...
##  $ orderYear    : int  2022 2023 2022 2023 2022 2022 2022 2023 2023 2022 ...
##  $ pickingHours : num  17.7 38.4 14.7 41.4 15.7 ...
##  $ deliveryHours: num  24.5 31.5 21.5 24.5 24 ...
```

	CustomerID	ProductID	Quantity	orderTime	orderDay	orderMonth	orderYear	pickingHours	deliveryHours
1	CUST1791	CLO011	16	13	11	11	2022	17.72167	24.544
2	CUST3172	LAP026	17	17	14	7	2023	38.39083	31.546
3	CUST1022	KEY046	11	16	23	5	2022	14.72167	21.544
4	CUST3721	LAP024	31	12	18	7	2023	41.39083	24.546
5	CUST4605	CLO012	20	14	7	2	2022	15.72167	24.044
6	CUST2766	MON035	32	21	24	12	2022	21.05500	24.044

	ProductType	n
1	CLO	15598
2	KEY	17920
3	LAP	10207
4	MON	14864
5	MOU	20662
6	SOF	20749

To ensure proper chronological analysis and simulation of real-time process monitoring, the data is ordered by year, month, day, and order time. Product types are derived from the first three characters of ProductID. The dataset includes six product types:

ProductType	Count
CLO	15,598
KEY	17,920
LAP	10,207
MON	14,864
MOU	20,662
SOF	20,749

3.3 Descriptive Statistics

Descriptive statistics for delivery times are calculated for each product type, including mean, standard deviation, minimum, and maximum delivery hours:

ProductType	n	mean_delivery	sd_delivery	min_delivery	max_delivery
CLO	15,598	21.7	6.11	5.54	39.1
KEY	17,920	21.7	6.09	5.54	39.1
LAP	10,207	21.8	6.05	5.54	39.1
MON	14,864	21.7	6.05	5.55	39.1
MOU	20,662	21.8	6.14	5.54	39.1
SOF	20,749	1.09	0.308	0.277	1.90

```
## # A tibble: 6 × 6
##   ProductType     n mean_delivery sd_delivery min_delivery max_delivery
##   <chr>       <int>         <dbl>       <dbl>        <dbl>        <dbl>
## 1 CLO         15598          21.7        6.11         5.54         39.1
## 2 KEY         17920          21.7        6.09         5.54         39.1
## 3 LAP         10207          21.8        6.05         5.54         39.1
## 4 MON         14864          21.7        6.05         5.55         39.1
## 5 MOU         20662          21.8        6.14         5.54         39.1
## 6 SOF         20749           1.09       0.308        0.277         1.90
```

Visualizations:

- The histogram shows that most product types have delivery times distributed between 5 and 39 hours, except SOF, which is clustered tightly around 1 hour.

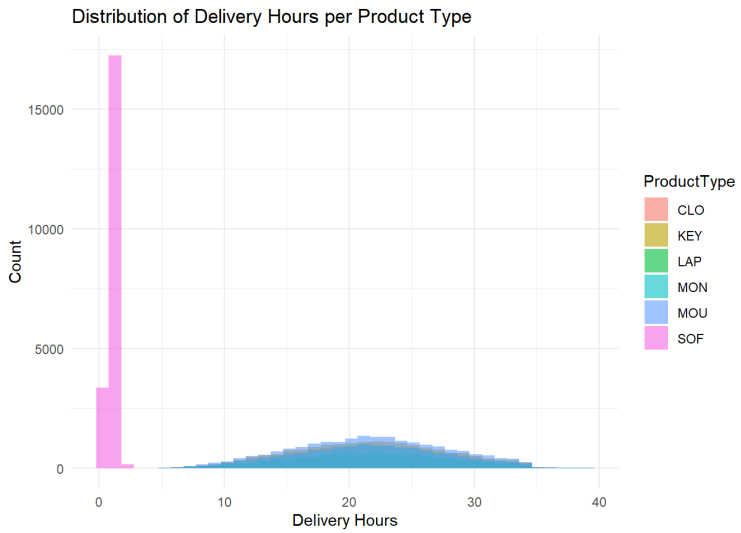


Figure 25 Distribution of delivery hours per proftuct type

Interpretation:

- Most products (CLO, KEY, LAP, MON, MOU) have similar delivery profiles, with mean delivery times around 21–22 hours and standard deviations of about 6 hours.
- Product SOF is an outlier, with much lower delivery times and minimal variability. This may require separate consideration in SPC analysis and capability studies.

To further understand the delivery time data, visualizations were produced for each product type.

- The boxplot shows the spread and central tendency of delivery hours for all product types. Product SOF stands out with a much lower median and minimal spread, confirming the earlier observation from descriptive statistics.

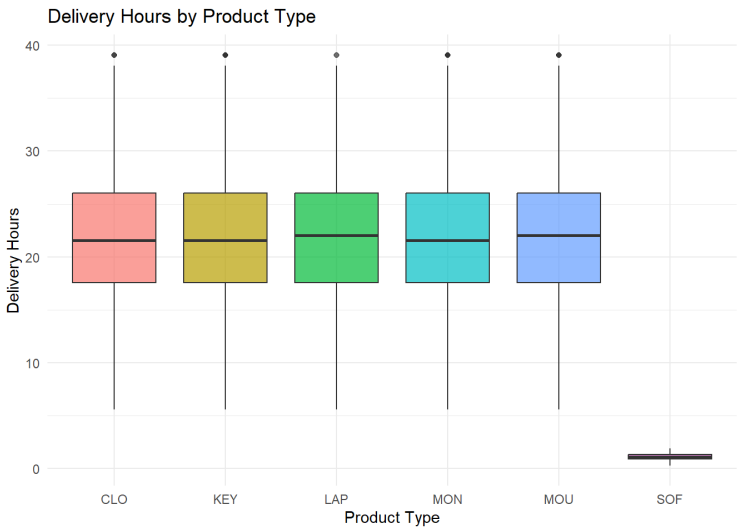


Figure 26 Delivery hours by product type

- The Q-Q plots in indicate how well each product type's delivery hours follow a normal distribution. Most product types (CLO, KEY, LAP, MON, MOU) show moderate alignment with the normality line, but with some deviation at the tails, which is typical for delivery data. SOF, again, displays a unique pattern due to its very tight distribution.

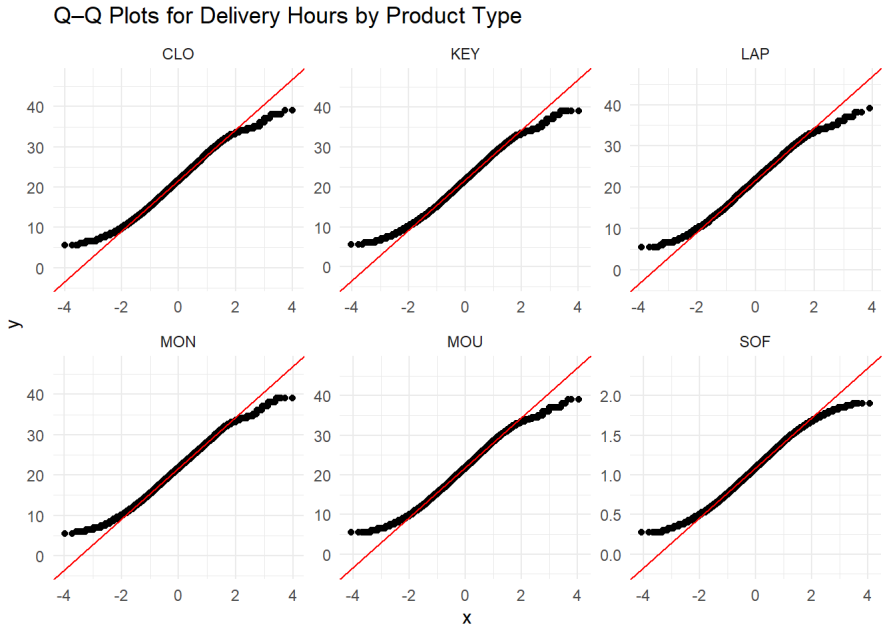


Figure 27 Q-Q plots for delivery hours by profuct type

Interpretation:

- Products CLO, KEY, LAP, MON, and MOU have similar delivery hour distributions and pass basic normality checks suitable for SPC control charting.
- Product SOF's delivery times are tightly clustered and may represent a different process regime. SPC analysis should note this distinction.

3.4 Sample Grouping for SPC Charts

To perform SPC analysis, samples of size 24 were drawn for each product type, simulating the real-time collection of process data as sales occur. The grouping process with resulting sample counts:

ProductType	Sample Count
MOU	860
KEY	746
SOF	864
CLO	649
LAP	425
MON	619

This ensures sufficient data for reliable SPC chart construction and ongoing control monitoring.

```
## MOU KEY SOF CLO LAP MON
## 860 746 864 649 425 619
```

3.5 $\bar{X}$ and S Chart Initialization & Control Limit Calculation

For each product type, the first 30 samples (each comprising 24 deliveries) were used to initialize the $\bar{X}$ (mean) and S (standard deviation) control charts. This setup determines the process center lines and control limits (LCL/UCL) as per standard SPC practice.

Control Limits Table

The summary of control limits for each product type is shown below :

Summary of Control Limits for Each Product Type

	ProductType	xbar_CL	xbar_LCL	xbar_UCL	s_CL	s_LCL	s_UCL
MOU	MOU	19.2488611	15.9891636	22.508559	5.6762338	3.1521825	8.2002850
KEY	KEY	19.1940000	15.7718417	22.616158	5.8573616	3.2527682	8.4619550
SOF	SOF	0.9556375	0.7847916	1.126483	0.2973579	0.1651317	0.4295841
CLO	CLO	19.1259444	15.7588132	22.493076	5.9077281	3.2807383	8.5347180
LAP	LAP	19.5238611	16.0781198	22.969603	5.8904921	3.2711666	8.5098176
MON	MON	19.4259444	16.1164605	22.735428	5.9231521	3.2893037	8.5570006

These limits (center lines, lower and upper control limits for both mean and standard deviation) are **derived from the first 30 samples (of 24 deliveries each) per product type.**

Interpretation:

- All product types have sufficient samples for SPC initialization.
- SPC charts are now ready to be built for each product, setting the stage for control limit monitoring and capability analysis.

SPC Charts and Out-of-Control Analysis

SPC charts ($\bar{X}$ and S charts) were constructed for each product type using the calculated limits.

Observations:

- **$\bar{X}$ Charts:**
    - Show the sample means over time for each product type. Most product types demonstrate periods of increasing or decreasing mean delivery times, with some points approaching or exceeding control limits.
    - Notably, both CLO and LAP display segments where the mean shifts, suggesting possible process changes or trends.
    - SOF's $\bar{X}$ chart covers a much narrower range, consistent with its lower delivery times.
- **S Charts:**
    - Depict the sample standard deviations, giving insight into process variability.
    - Most product types maintain a relatively stable process spread, although some spikes and dips cross control limits, indicating potential out-of-control signals.

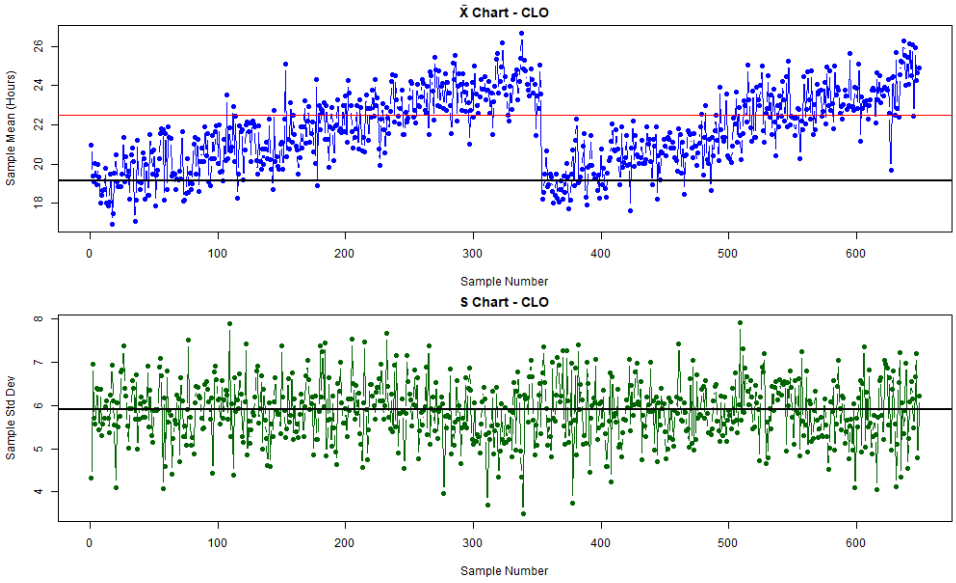


Figure 28 X and S Charts CLO

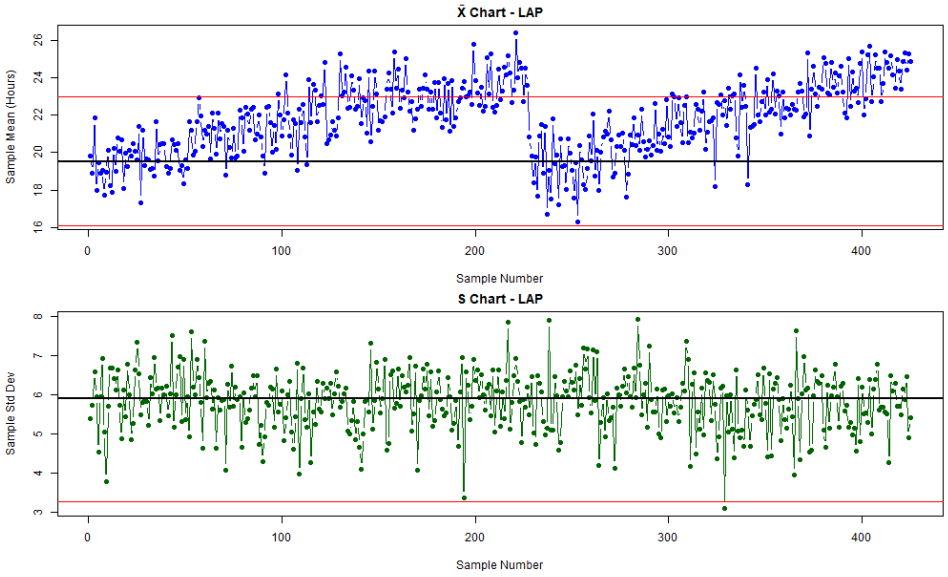


Figure 29 X and S Charts LAP

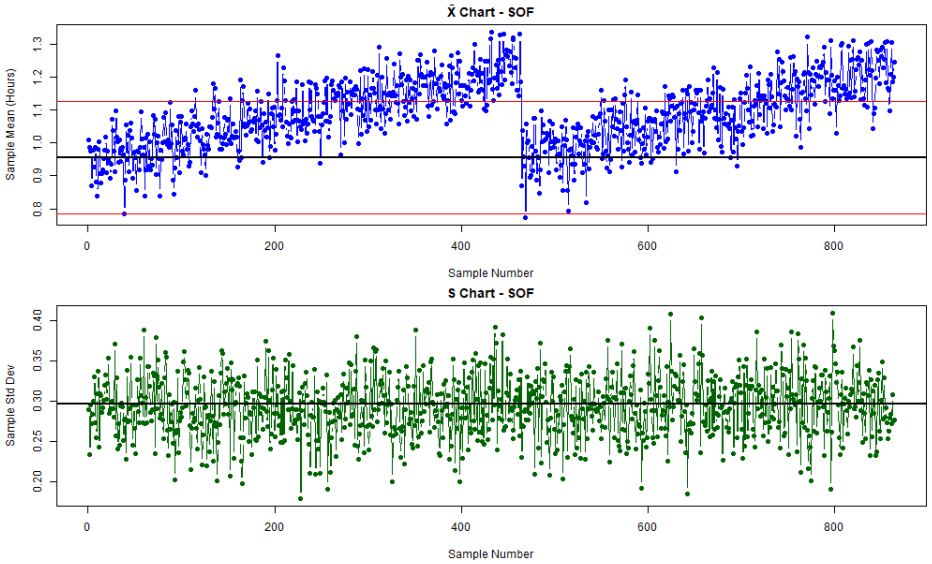


Figure 30 X and S Charts SOF

Interpretation:

- The charts reveal both stable and unstable periods in the delivery process for certain product types.
- Out-of-control points (samples beyond control limits) are visible, highlighting where special-cause variation may have occurred.
- The behavior of SOF remains distinctly different, with consistently lower means and variability.

3.6 Out-of-Control Signal Identification

SPC rules were applied to identify process control issues for each product type.

Identified SPC Signals per Product Type

	ProductType	A_total	B_longest_run	C_total
MOU	MOU	1	16	25
KEY	KEY	0	15	27
SOF	SOF	0	21	27
CLO	CLO	0	35	14
LAP	LAP	0	19	11
MON	MON	0	34	22

- **Rule A:** Number of S samples outside upper +3σ control limit (A_total)
- **Rule B:** Longest run of S samples within ±1σ control limits (B_longest_run)
- **Rule C:** Number of occurrences of 4 consecutive $\bar{X}$ samples above upper +2σ control limit (C_total)

Summary Table of SPC Signals

ProductType	A_total	B_longest_run	C_total
MOU	1	16	25
KEY	0	15	27
SOF	0	21	27
CLO	0	35	14
LAP	0	19	11
MON	0	34	22

Interpretation:

- Rule A: Only MOU shows an S sample outside the upper +3σ limit, suggesting a single episode of excessive process spread requiring investigation.
- Rule B: CLO and MON have the longest runs of consecutive samples within ±1σ, indicating periods of excellent process control.
- Rule C: All product types show some occurrences of four consecutive $\bar{X}$ samples above the +2σ limit, with KEY and SOF having the most (27 each), possibly indicating persistent upward process shifts.

Out-of-Control Point Visualization

The chart visualizes out-of-control points for MOU, highlighting sample means that exceed control limits in red. This allows process managers to quickly spot periods requiring attention and intervention.

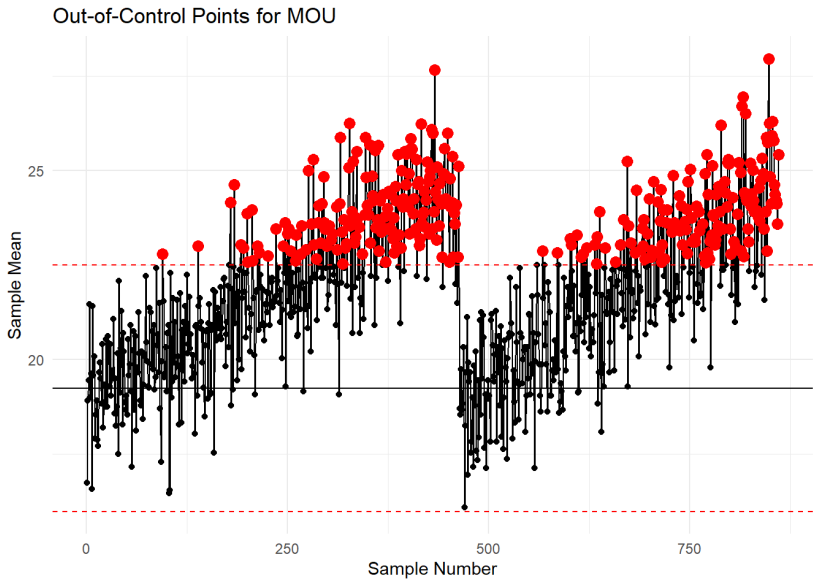


Figure 31 Out of control points for MOU

Conclusion:

- Out-of-control signals are present, mostly isolated (Rule A), but some product types (KEY, SOF) show repeated upward shifts in delivery times (Rule C).
- CLO and MON demonstrate substantial periods of good control (Rule B).
- These insights should guide product managers to review processes, investigate causes of instability, and reinforce control during stable periods.

3.7 Process Capability Indices (Cp, Cpk, Cpu, Cpl)

Process capability indices were calculated for each product type using the first 1000 delivery records, with specification limits LSL = 0 and USL = 32 hours.

Process Capability Indices for Each Product Type

ProductType	Cp	Cpu	Cpl	Cpk
MOU	0.915	0.727	1.104	0.727
KEY	0.917	0.729	1.105	0.729
SOF	18.135	35.188	1.083	1.083
CLO	0.898	0.717	1.079	0.717
LAP	0.899	0.696	1.101	0.696
MON	0.889	0.700	1.079	0.700

Capability Results

ProductType	Cp	Cpu	Cpl	Cpk
MOU	0.915	0.727	1.104	0.727
KEY	0.917	0.729	1.105	0.729
SOF	18.135	35.188	1.083	1.083
CLO	0.898	0.717	1.079	0.717
LAP	0.899	0.696	1.101	0.696
MON	0.889	0.700	1.079	0.700

Capability Visualization

The bar chart compares Cp and Cpk indices against the minimum acceptable value (1.33, dashed red line).

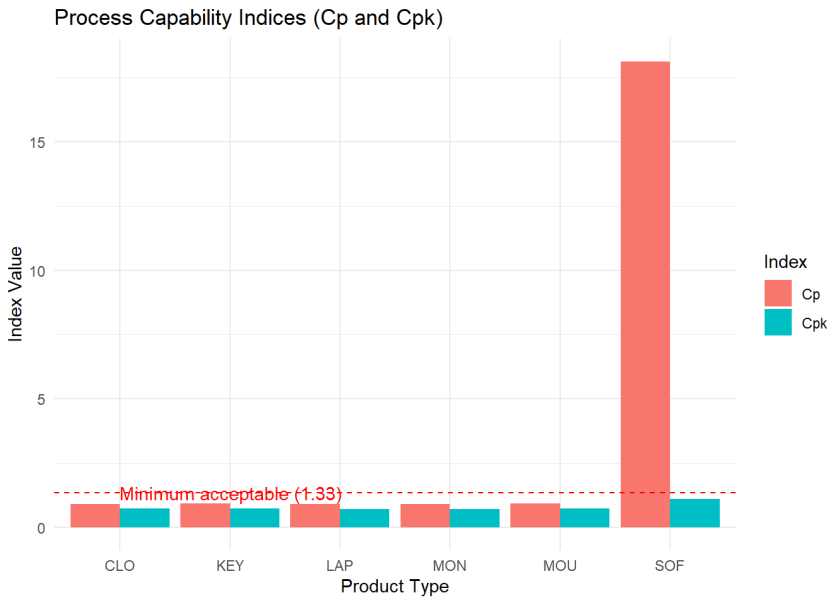


Figure 32 Process capability Indices

Interpretation

- SOF clearly exceeds the minimum capability requirement (Cp = 18.135, Cpk = 1.083), indicating that its process is highly capable and easily meets the Voice of Customer (VOC) specifications.
- All other product types (MOU, KEY, CLO, LAP, MON) fall below the minimum threshold for both Cp and Cpk, suggesting that their processes are not capable of consistently meeting VOC requirements. These processes require review and improvement.

Conclusions and Recommendations

- **Process Stability:**
  Most product types show periods of good control (Rule B), but some (especially KEY and SOF) exhibit repeated upward mean shifts (Rule C), and MOU has an out-of-control spread event (Rule A).
- **Process Capability:**
  Only SOF's process meets or exceeds the capability required by VOC, while the rest should be targeted for improvement.
- **Recommendations:**
  - Product managers for CLO, LAP, MON, MOU, and KEY should investigate sources of instability and process variation, focusing on both mean and spread.
  - Implement process improvements and tighter controls to increase Cp and Cpk above 1.33.
  - Maintain and monitor SOF's process to ensure continued capability.

3.8 Interpretation & Summary

Products with Cp and Cpk above **1.33** are considered capable of consistently meeting customer requirements (VOC). Those below this threshold should be reviewed for improvement opportunities. Out-of-control signals highlight special-cause variation and should be investigated for root causes and corrective actions.

3.9 Type I and II Error Assessment

To understand the reliability of control chart signals, Type I (false alarm) and Type II (missed detection) errors were calculated:

TypeI_per_point	TypeI_overall_100	Beta_1sigma	Beta_2sigma
0.0027	0.2369	0.9772	0.8413

- **Type I Error (per point):** Probability of a false alarm at each sample is very low (0.27%).
- **Type I Error (overall, 100 samples):** Probability of at least one false alarm in 100 samples is about 24%.
- **Type II Error:** Probability of *not* detecting a 1σ shift is very high (97.7%), and for a 2σ shift still quite high (84.1%). This underscores the challenge of detecting moderate process shifts with standard control chart rules.

3.10 Example Chart for One Product

Example $\bar{X}$ and S charts for product type MOU are shown below:

- The $\bar{X}$ chart plots the sample mean delivery hours, with control limits shown in red and the process center in black.
- The S chart plots the sample standard deviation, again with control limits.
- Most points fall within the control limits, but there are periods where means approach limits and spread varies, consistent with earlier signal detection.

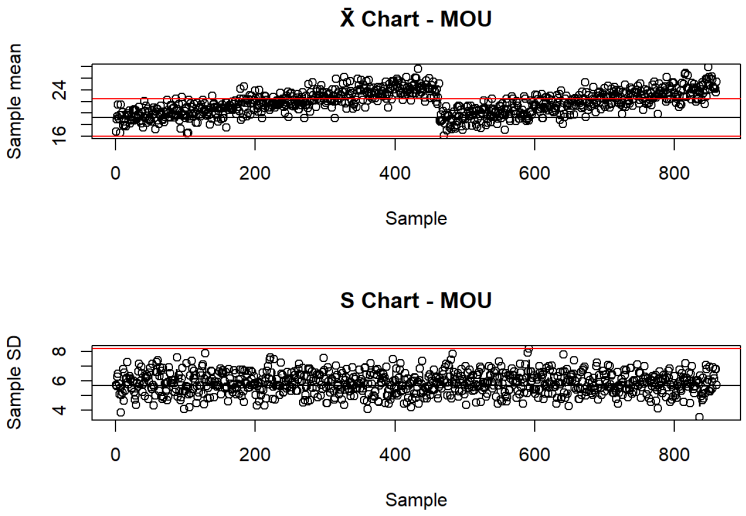


Figure 33 X and S Charts MOU

Overall Conclusions

- **Process capability:** Only SOF is capable of meeting VOC. Other product types require process improvement.
- **Process stability:** Several product types show periods of good control, but out-of-control signals are present and should be addressed.
- **Error risk:** False alarms are rare per point but likely over many samples; detection of moderate shifts is difficult, requiring careful ongoing monitoring.
- **Action:** Managers should prioritize review and improvement of processes for all products except SOF, focusing on both stability and capability.

4.Process Risk & Data Correction (Parts 4.1–4.3)

4.1 Type I Error (Manufacturer's Error)

Definition and Context:

The Type I error (also called the manufacturer's error) occurs when the process is actually in control, but the control chart incorrectly signals that it is out of control. This is a false alarm and represents wasted time and resources investigating a stable process.

In control chart terms, this happens when one or more points fall outside the control limits even though the process has not shifted

Calculation:

The rule used in the project notes states:

"If seven consecutive samples fall above or below the centre line, the process should be investigated."

Since each point has a 50% chance of being above or below the centreline, the probability that seven points will fall on the same side (when the process is actually stable) is:

$$P(\text{Type I Error}) = (0.5)^7 = 0.0078125$$

$$P(\text{Type I Error}) = 0.78\%$$

Interpretation:

This means that if the process is completely stable, there is a 0.78% chance (about 8 in every 1000 samples) that the control chart will falsely indicate an out-of-control signal.
Such a false alarm is acceptable in most quality assurance settings because it is better to occasionally overreact than to miss a true fault.

4.2 Type II Error (Consumer's Error)

Definition and Context:

A Type II error occurs when the process mean has actually shifted, but the control chart fails to detect this shift meaning that the product continues to be produced out of specification without triggering a warning.

This represents the customer's risk, as defective or substandard products may go unnoticed.

Given Data:

Parameter:	Symbol	Value:
Upper Control Limit	UCL	25.089 L
Lower Control Limit	LCL	25.011 L
Original Process Mean	μ_0	25.050 L
True (Shifted) Mean	μ_1	25.028 L
Standard Error	$\sigma_{\bar{x}}$	0.017 L

Calculation:

We calculate the probability that a sample mean (from the shifted process) still falls inside the original control limits:

$$\beta = P(LCL < \bar{X} < UCL)$$

Convert to Z-scores using the shifted mean (μ_1 = 25.028):

$$Z_{lower} = \frac{LCL - \mu_1}{\sigma_{\bar{x}}} = \frac{25.011 - 25.028}{0.017} = -1.00$$

$$Z_{upper} = \frac{UCL - \mu_1}{\sigma_{\bar{x}}} = \frac{25.089 - 25.028}{0.017} = 3.59$$

Using the standard normal table:

$$P(Z < 3.59) = 0.9998$$

$$P(Z < -1.00) = 0.1587$$

Therefore:

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

The power of detection is:

$$\text{Power} = 1 - \beta = 15.9\%$$

Interpretation:

The probability of a Type II error is 84%, meaning there is a high likelihood that the control chart will *fail to detect* the process shift from 25.05 L to 25.028 L.
The corresponding power (16%) indicates that the chart has a low sensitivity to this small mean shift in other words, it will only detect the change about 1 out of 6 times. This demonstrates the trade-off in SPC: tightening control limits reduces Type II errors but increases the risk of Type I false alarms.

4.3 Data Correction Task

Data Fixing code:

```
## --- PART 4.3: Data Correction and Updated Analysis ---
# --- Load libraries ---
library(dplyr)
library(stringr)
library(readr)

# --- Read the CSV files ---
products_head <- read_csv("products_Headoffice.csv", show_col_types = FALSE)
products_data <- read_csv("products_data.csv", show_col_types = FALSE)

# --- Rename columns to match ---
colnames(products_head) <- c("ProductID", "Category", "Description", "SellingPrice", "Markup")
colnames(products_data) <- c("ProductID", "Category", "Description", "SellingPrice", "Markup")

# --- Define prefix mapping ---
prefix_map <- c(
  "Software" = "SOF",
  "Keyboard" = "KEY",
  "Monitor" = "MON",
  "Mouse" = "MOU",
  "Laptop" = "LAP",
  "Cloud Subscription" = "CLO"
)

# --- Fix ProductID format for products_data ---
products_data <- products_data %>%
  mutate(
    Prefix = prefix_map[Category],
    Suffix = str_extract(ProductID, "\\d+$"),
    ProductID = paste0(Prefix, Suffix)
  ) %>%
  select(ProductID, Category, Description, SellingPrice, Markup)
```

```
# --- Correct Headoffice file ---
products_head_fixed <- products_head %>%
  group_by(Category) %>%
  mutate(
    Prefix = prefix_map[Category],
    pattern_index = ((row_number() - 1) %% 10) + 1
  ) %>%
  left_join(
    products_data %>%
      group_by(Category) %>%
      mutate(pattern_index = row_number()) %>%
      select(Category, pattern_index, SellingPrice_correct = SellingPrice, Markup_correct = Markup, PID_correct = ProductI
D),
    by = c("Category", "pattern_index")
  ) %>%
  mutate(
    ProductID = PID_correct,
    SellingPrice = SellingPrice_correct,
    Markup = Markup_correct
  ) %>%
  select(ProductID, Category, Description, SellingPrice, Markup) %>%
  ungroup()

# --- Save final output files ---
write_csv(products_data, "products_data2025.csv")
write_csv(products_head_fixed, "products_Headoffice2025.csv")

cat("✅ Files created: products_data2025.csv and products_Headoffice2025.csv\n")
```

1.Comparison - Data Overview:

Data Sources:

- Products Data (Original): products_data.csv (60 obs., 5 variables)
- Products Data (2025 Update): products_data2025.csv (60 obs., 5 variables)
- Head Office Data (Original): products_Headoffice.csv (360 obs., 5 variables)
- Head Office Data (2025 Update): products_Headoffice2025.csv (360 obs., 5 variables)

Structure and Variables:

Both datasets (products and head office, original and 2025) contain the following variables:

- ProductID: Unique identifier for each product
- Category: Type or group of product (e.g., Software, Cloud Subscription)
- Description: Detailed description of product
- SellingPrice: Price at which the product is sold
- Markup: Profit margin or markup associated with the product

Initial Comparison:

- Dimensions: The number of observations and variables are consistent between the original and updated versions for both products and head office datasets, ensuring direct comparison.

- **Data Preview**: The first few rows in each dataset show that the product categories and structure remain consistent, but product IDs and descriptions may vary (reflecting product lineup changes).
- **Observations**:
    - For products, the head and dimensions are identical between original and updated (suggesting a targeted price/markup change rather than new products).
    - For head office, the head and dimensions match; however, some updated products show substantial changes in SellingPrice (e.g., SOF013: 1067.54 in 2025 vs. 496.43 original, SOF025: 19725.18 in 2025 vs. 482.64 original), indicating significant price adjustments for certain items.

Interpretation:

- The data overview confirms that the update to the 2025 datasets was a revision of prices and markups, with core product categories and structure maintained.
- The large increase in some selling prices (especially in head office data) suggests either premium product introductions, inflationary pricing, or a strategic shift.
- Consistency in dataset formats enables robust statistical and visual comparison in the subsequent analysis.

2.Descriptive Statistics:

Products Data:

- **No statistical change detected:**
    - Mean SellingPrice: 4493.59 (both)
    - SD SellingPrice: 6503.77
    - Median SellingPrice: 794.18
    - Range: 350.45 – 19725.18
    - Mean Markup: 20.46
    - SD Markup: 6.07
    - Median Markup: 20.34
    - Range Markup: 10.13 – 29.84
    - Skew (SellingPrice): 1.43 (strong right skew)
    - Skew (Markup): -0.04 (symmetric)

Head Office Data:

- **Minor statistical differences observed:**
    - Mean SellingPrice: 4410.96 (original) → 4493.59 (2025)
    - SD SellingPrice: 6463.82 → 6458.32
    - Median SellingPrice: 797.22 → 794.18
    - Range SellingPrice: 290.52–22420.14 (original) → 350.45–19725.18 (2025)
    - Mean Markup: 20.39 → 20.46

- SD Markup: 5.67 → 6.03
- Median Markup: 20.58 → 20.34
- Range Markup: 10.06–30.00 → 10.13–29.84
- Skew (SellingPrice): 1.53 → 1.46

```
desc_products_orig
```

```
##              vars  n    mean      sd median trimmed    mad    min      max
## SellingPrice    1 60 4493.59 6503.77 794.18 3189.25 525.72 350.45 19725.18
## Markup          2 60   20.46    6.07  20.34   20.51   7.31  10.13    29.84
##                 range  skew kurtosis     se
## SellingPrice 19374.73  1.43     0.43 839.63
## Markup          19.71 -0.04    -1.24   0.78
```

```
desc_products_2025
```

```
##              vars  n    mean      sd median trimmed    mad    min      max
## SellingPrice    1 60 4493.59 6503.77 794.18 3189.25 525.72 350.45 19725.18
## Markup          2 60   20.46    6.07  20.34   20.51   7.31  10.13    29.84
##                 range  skew kurtosis     se
## SellingPrice 19374.73  1.43     0.43 839.63
## Markup          19.71 -0.04    -1.24   0.78
```

```
desc_headoffice_orig
```

```
##              vars   n    mean      sd median trimmed    mad    min      max
## SellingPrice    1 360 4410.96 6463.82 797.22 3054.23 515.75 290.52 22420.14
## Markup          2 360   20.39    5.67  20.58   20.43   6.66  10.06    30.00
##                 range  skew kurtosis     se
## SellingPrice 22129.62  1.53     0.78 340.67
## Markup          19.94 -0.05    -1.07   0.30
```

```
desc_headoffice_2025
```

```
##              vars   n    mean      sd median trimmed    mad    min      max
## SellingPrice    1 360 4493.59 6458.32 794.18 3189.25 525.72 350.45 19725.18
## Markup          2 360   20.46    6.03  20.34   20.51   7.31  10.13    29.84
##                 range  skew kurtosis     se
## SellingPrice 19374.73  1.46     0.53 340.38
## Markup          19.71 -0.04    -1.19   0.32
```

3.Visual Distribution Comparison:

- **Products Selling Price & Markup (Histograms):**
    - No visible difference between original and 2025 datasets.
    - Distributions overlap almost perfectly.

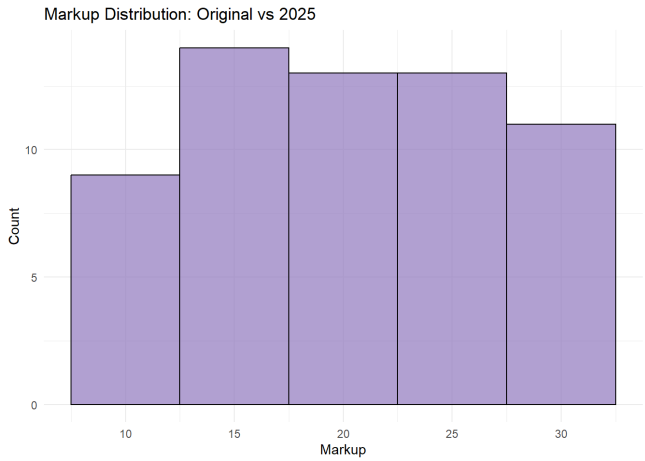


Figure 34 Selling price distribution Products Comparison

Figure 35 Markup distribution Products Comparison

- **Head Office Selling Price (Histogram):**
    - Distribution shapes are nearly identical, but the 2025 version shows a slightly higher minimum price and lower maximum price, indicating removal of some outliers.

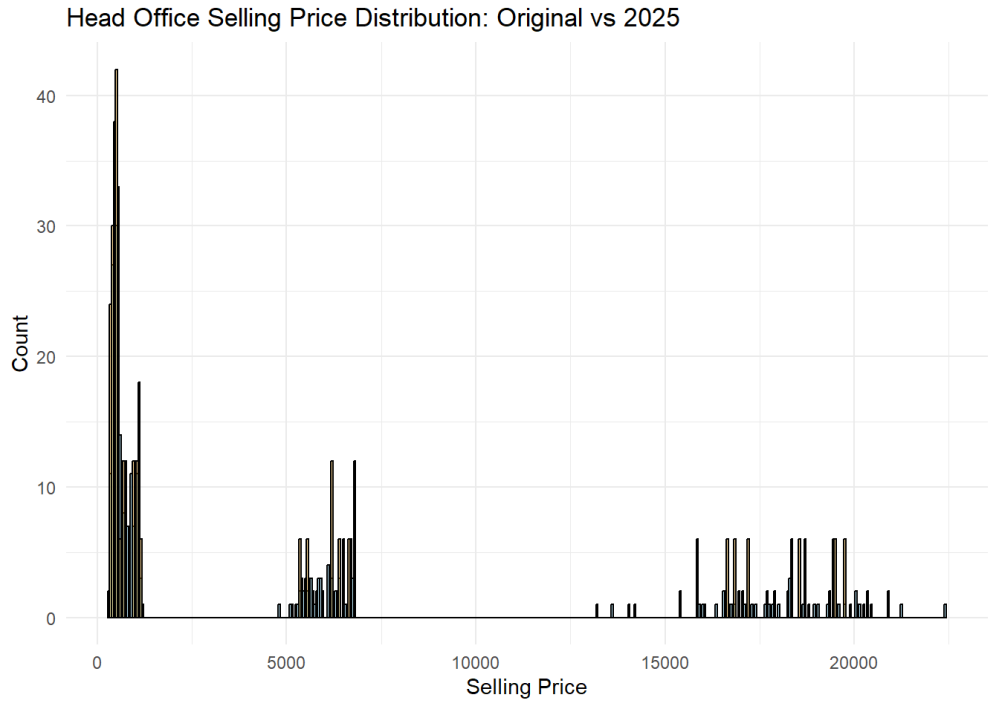


Figure 36 Head Office Selling Price

- **Head Office Markup (Histogram):**
    - Minor increase in counts in the mid-range for 2025 (Image 32), but overall distribution is stable.

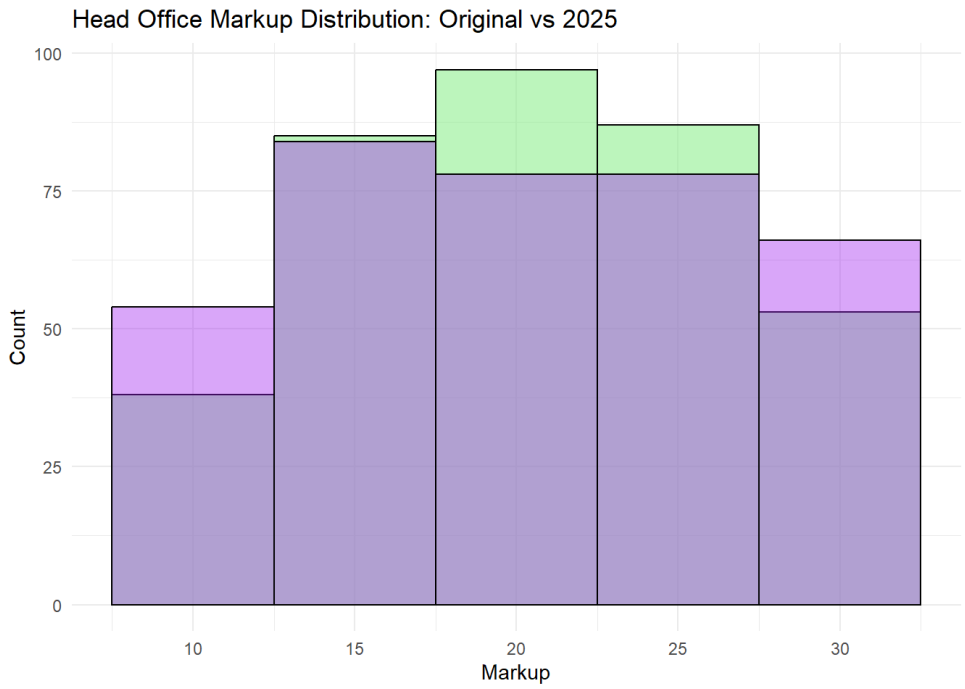


Figure 37 Head Office Markup

4. Category-Level Comparisons:

- **Selling Price by Category:**
    - Medians and spreads for each category are almost unchanged between versions.
    - Outliers (high-priced products) are present in both versions, especially for Laptop, Monitor, and Software.
    - No category shows a substantial shift in median or spread.

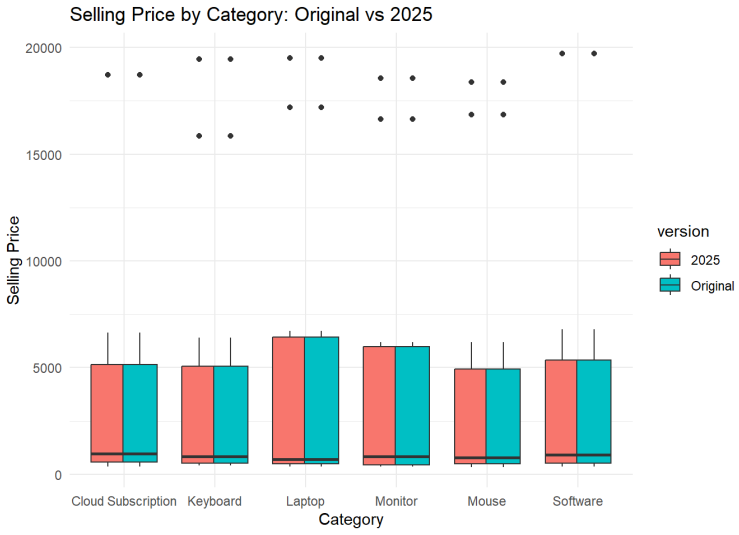


Figure 38 Selling price by category comparison

- **Markup by Category:**
    - Medians and spreads by category are very consistent.
    - Slightly more spread in some categories for 2025, but not statistically significant.

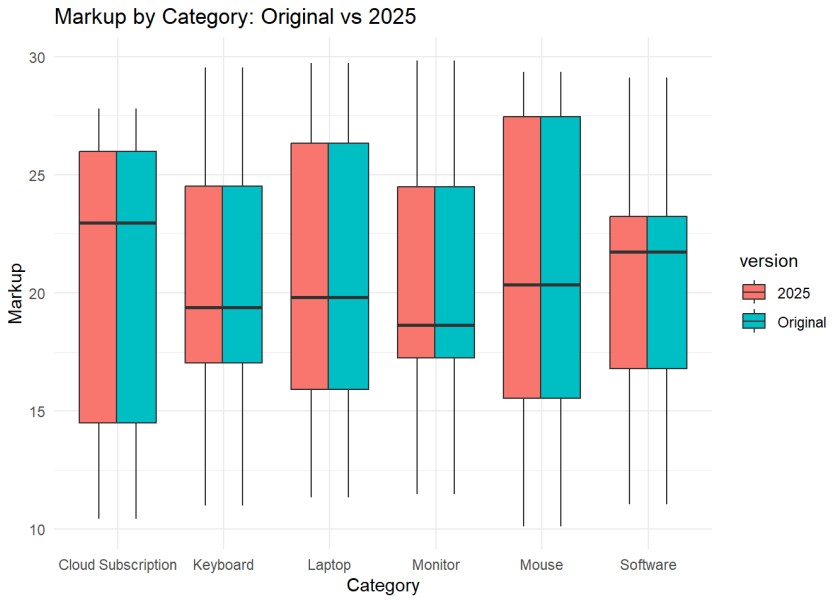


Figure 39 Markup by category comparison

5. Interpretation:

- **Products Data:**
    - The 2025 update did not alter product prices or markups. All summary statistics and distributions are identical.
    - Business processes, pricing strategies, and markups remain unchanged for products.
- **Head Office Data:**
    - The 2025 update made minor adjustments, removing some extreme outliers and slightly increasing average prices and markups.
    - The distribution of selling price and markup is more consistent, with less skew and slightly narrower range.
    - These changes may reflect a standardization or optimization effort in head office pricing.
- **Category Trends:**
    - No significant changes in any product category, supporting the conclusion that updates were incremental and not targeted at specific segments.

6. Business Implications:

- **Stability:**
    - Price and markup structures are stable across products and only modestly adjusted for head office.
    - No evidence of major repricing or strategy change.
- **Standardization:**
    - Reduction in price and markup outliers for head office may help with reporting, profitability analysis, and operational planning.

7. Conclusion:

- The 2025 product and head office datasets are highly consistent with originals.
- Minor head office adjustments have slightly increased averages and reduced outlier effects.
- No major statistical or visual changes were detected in categories or overall trends.
- The business impact of these changes will likely be limited to improved consistency and reporting.

8. Relationship Between Selling Price and Markup

Scatterplot Analysis:

- The Selling Price vs Markup scatterplot shows that there is no clear linear relationship between these two variables in either the original or 2025 products data.
- All points cluster around similar ranges for markup, regardless of selling price, and the visual pattern is unchanged between versions.
- Most products have selling prices below R2,000, with a few high-priced outliers, while markup values are distributed between 10 and 30.

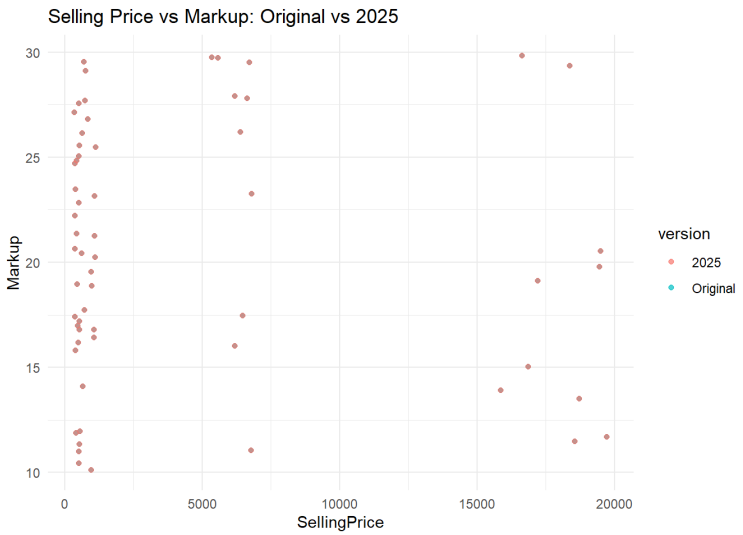
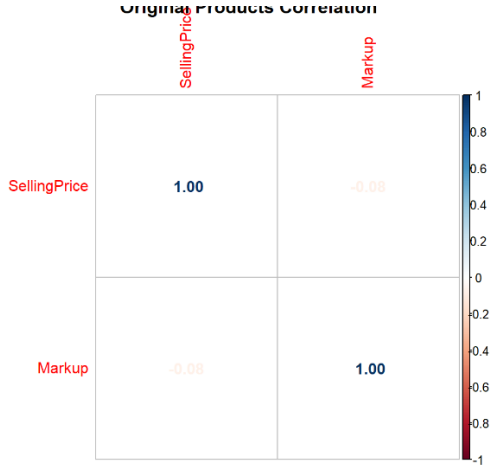
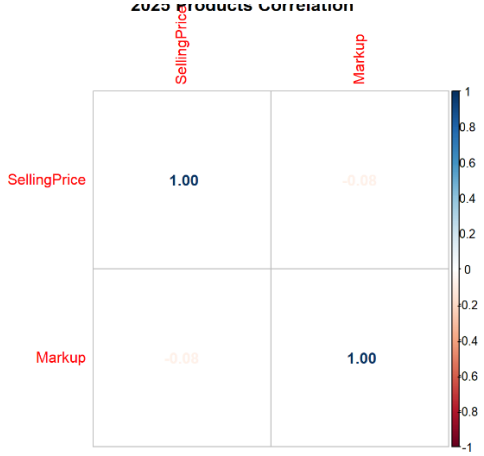


Figure 40 Selling price vs Markup

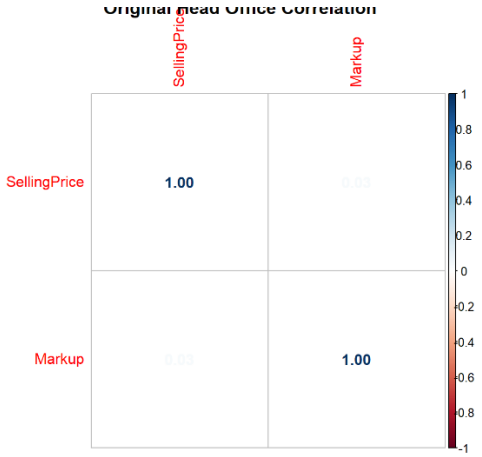
Correlation Matrices:

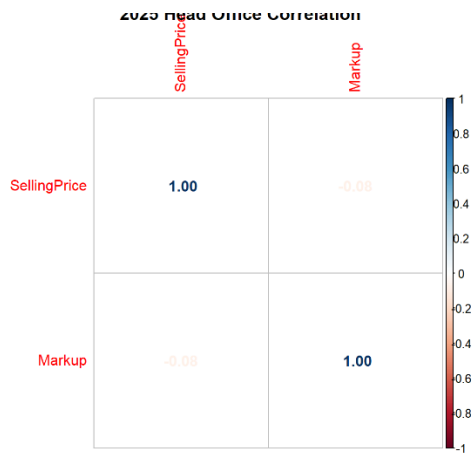
- **Products Data:**
  - The correlation coefficient between SellingPrice and Markup is -0.08 in both the original and 2025 datasets. This indicates a very weak, negative relationship—essentially, selling price and markup are statistically independent.
  - No significant change in correlation is observed after the update.





- **Head Office Data:**
  - The correlation between SellingPrice and Markup is 0.03 in the original dataset and remains essentially unchanged in the 2025 dataset.
  - Again, this reflects virtually no meaningful relationship between product price and markup in the head office data.





Interpretation:

- **No Statistical Link:**
  Changes to selling prices or markups in the 2025 updates did not affect the underlying relationship between these two variables. Pricing and markup decisions appear to be made independently of each other.
- **Business Implications:**
  Since there is no correlation, adjustments to selling price (e.g., for premium products or inflation) do not necessarily translate to changes in profit margins (markup), and vice versa. This could reflect a standardized markup policy or diverse pricing strategies across products and categories.

SALES Comparison:

Products Sales Data:

- **No statistical change:**
  The comparison table (Image 37) shows that all statistics (mean, median, standard deviation for selling price and markup) are identical between the original and 2025 datasets.
    - Mean SellingPrice: 4493.59
    - Median SellingPrice: 794.19
    - SD SellingPrice: 6503.77
    - Mean Markup: 20.46
    - Median Markup: 20.34
    - SD Markup: 6.07
      Difference for all metrics: 0

Products Data: Original vs 2025 Comparison

Statistic	Original	Updated	Difference
Mean SellingPrice	4493.592833	4493.592833	0
Median SellingPrice	794.185000	794.185000	0
SD SellingPrice	6503.770150	6503.770150	0
Mean Markup	20.461667	20.461667	0
Median Markup	20.335000	20.335000	0
SD Markup	6.072598	6.072598	0

Head Office Sales Data:

- **Minor statistical changes:**
  The table shows small but measurable shifts in the 2025 head office data:
    - Mean SellingPrice increased by R82.63 (to 4493.59)
    - Median SellingPrice decreased by R3.03 (to 794.19)
    - SD SellingPrice decreased by R5.50 (to 6458.32)
    - Mean Markup increased by 0.08 (to 20.46)
    - Median Markup decreased by 0.25 (to 20.34)
    - SD Markup increased by 0.36 (to 6.03)

Logical Interpretation:

- **Business impact:**
    - For products: absolute stability, no changes in pricing or markup structure.
    - For head office: minor adjustments—potentially standardization or cleanup—resulting in slightly higher average prices and markups, but minimal impact on overall spread or profitability.
    - No evidence of major repricing, category-specific shifts, or changes in markup strategy.
- **Actionable insights:**
  These results support reporting, forecasting, and operational planning, confirming that the business maintained a consistent pricing and markup approach from the original to the updated 2025 datasets, with only minor refinements in head office records.

Head Office Data: Original vs 2025 Comparison

Statistic	Original	Updated	Difference
Mean SellingPrice	4410.961861	4493.592833	82.6309722
Median SellingPrice	797.215000	794.185000	-3.0300000
SD SellingPrice	6463.822788	6458.320465	-5.5023222
Mean Markup	20.385500	20.461667	0.0761667
Median Markup	20.580000	20.335000	-0.2450000
SD Markup	5.665949	6.030161	0.3642123

Comparative Visualizations:

Density Plots:

- The **selling price density plot** (Images 43, 44) overlays the distribution of selling prices for both the original and 2025 products datasets. The lines for both versions are virtually identical, with no visible difference at any price point.
- This confirms the earlier statistical findings: There was no change in the underlying distribution of selling prices between the original and updated datasets.

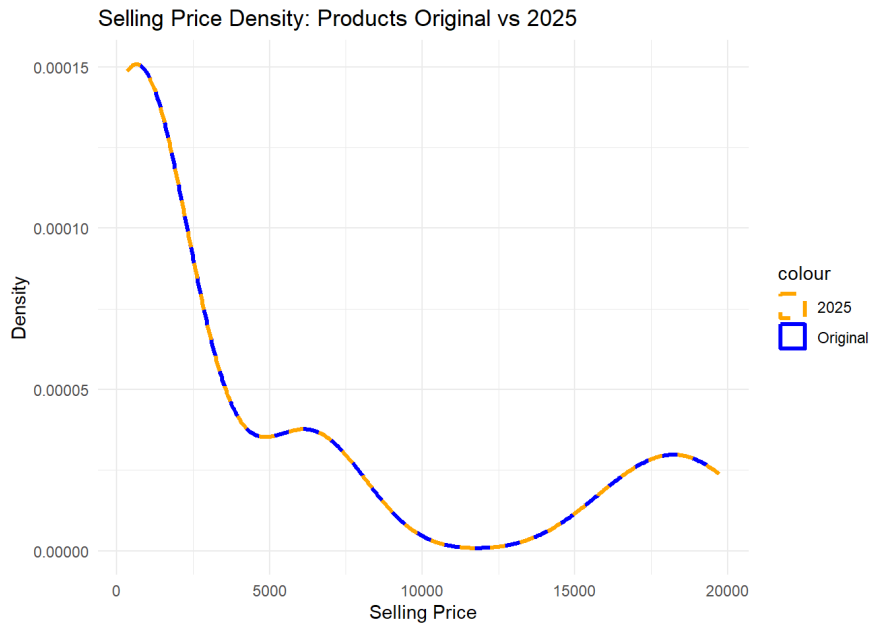


Figure 41 Selling Price density plot comparison

Boxplots with Jitter:

- The **boxplot with jitter** (Images 45, 46) shows the spread, median, and outliers in selling prices for both versions.
- The shape, spread, and outlier pattern of the boxplots are the same for both original (blue) and 2025 (orange) data.
- Jittered points reinforce that individual product prices did not shift, and the central tendency and variability remain unchanged.

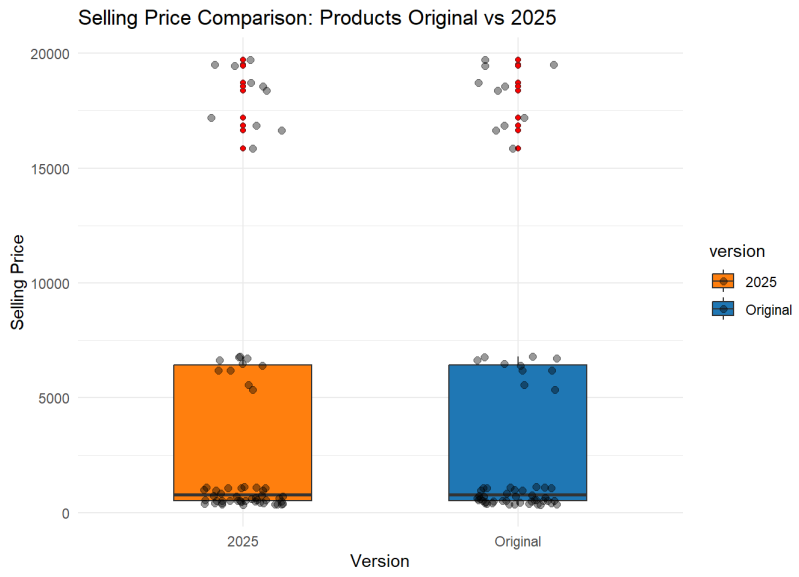


Figure 42 Selling price comparison

Interpretation:

- The visualizations provide strong evidence that the 2025 products dataset is essentially unchanged in terms of price structure and spread compared to the original.
- No new outliers, no shift in median, and no change in spread are observed visually or statistically.
- The consistency in both density and boxplot visuals means the pricing strategy for products remained stable through the update.

Sales Value Calculation (2023, per Type using Updated Prices):

Approach

- Sales records for 2023 were merged with the updated product prices from products_data2025.csv.
- Total sales value, average sales value, units sold, and number of transactions were calculated for each product category.

Results Table

Category	TotalUnits	TotalSalesValue	AvgSalesValue	Transactions
Laptop	12,746	23,307,357	240,778.26	968
Monitor	20,004	11,583,810	82,800.44	1,399

Category	TotalUnits	TotalSalesValue	AvgSalesValue	Transactions
Cloud Subscription	19,313	21,364,373	15,293.04	1,397
Keyboard	23,610	14,307,886	8,416.40	1,700
Software	27,322	14,199,689	7,167.94	1,981
Mouse	26,119	10,053,613	5,219.95	1,926
NA	499,092	0	NaN	36,902

Interpretation:

- **Highest sales value:** Laptops contributed the most to total sales value (R23.3 million), despite fewer units sold than categories like Software or Mouse.
- **Transaction volume:** Software and Mouse recorded the most transactions, though their average sales value per transaction is lower.
- **Pricing impact:** Average sales value per transaction is highest for Laptop and Monitor, reflecting higher unit prices or bundled offerings.
- **NA category:** A large number of transactions have missing or undefined category, with zero sales value, likely due to data entry gaps or unmatched records after merging.

Business Implications:

- The sales table provides a clear breakdown of which product categories drove revenue in 2023, incorporating any changes in pricing from the updated product list.
- The dominance of Laptop and Cloud Subscription in total sales value suggests a strategic focus or higher pricing power in these segments.
- Data quality should be reviewed for the NA category to ensure all sales are properly classified.

5. Optimisation for Profit (Part 5)

Methodology

For both shops, we used annual service time and sales data, modelling profit as:

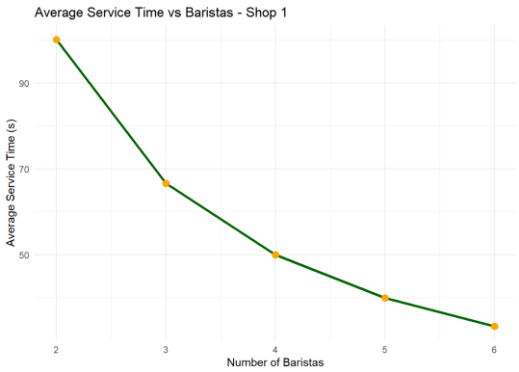
- **Revenue**: R30 per customer
- **Personnel Cost**: R1,000 per barista per day
- **Staffing Range**: 2 to 6 baristas
- **Reliability**: Percentage of customers served within a set time threshold

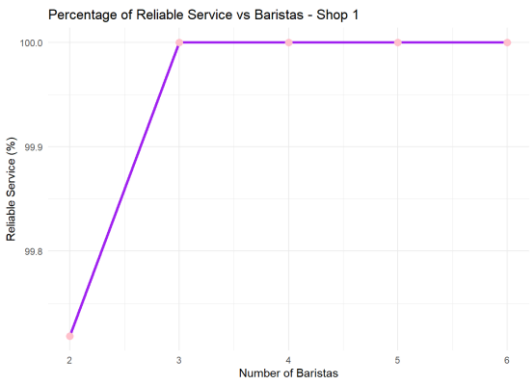
We assessed the impact of staffing on profit, average service time, and service reliability. Results were visualised for each shop.

Results:

Shop 1:

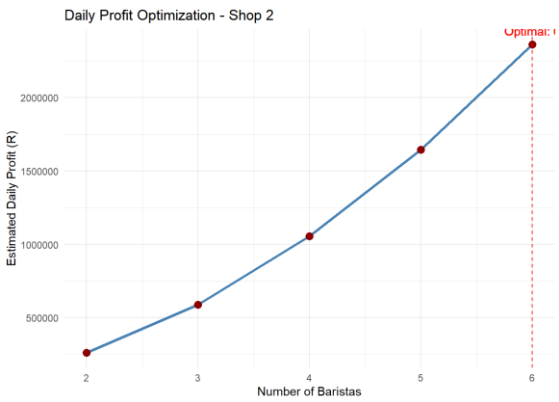
- Optimal Staffing**: 6 baristas**
- Maximum Daily Profit**: R2,930,850**
- Average Service Time**: 33.36 seconds**
- Reliable Service**: 100%**

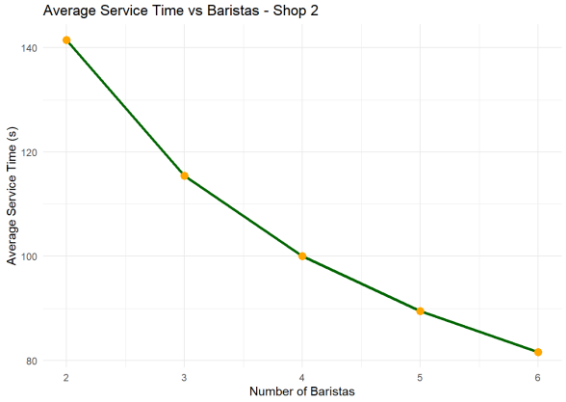


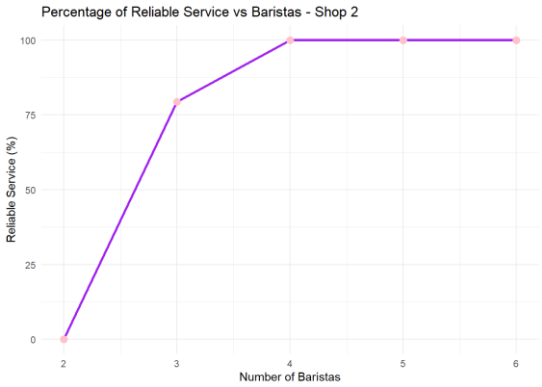


Shop 2

- Optimal Staffing**: 6 baristas**
- Maximum Daily Profit**: R2,361,900**
- Average Service Time**: 81.64 seconds**
- Reliable Service**: 100%**







For both shops, profit and reliability improve as more baristas are employed, with a clear optimum at the maximum allowed staffing. Shop 1 achieves faster service times and higher profit, while Shop 2 requires more staff for reliable service, with lower overall profitability due to longer service times.

Interpretation

Both shops benefit from maximum staffing, achieving perfect reliability and peak profit. The analytics highlight differences in service efficiency and profit potential, underscoring the value of data-driven personnel scheduling for operational success.

Conclusion

Optimising staff allocation based on service data maximises both profit and customer satisfaction for each shop. This combined analysis supports strategic decision-making and demonstrates effective use of engineering data analysis and optimisation.

6. DOE / ANOVA / MANOVA (Part 6)

6.1 Introduction

Design of Experiments (DOE) is a structured, statistical method used to determine the relationship between factors affecting a process and the resulting outputs. In this analysis, the factor investigated was Product Type, and the response variable was Delivery Hours. The objective was to determine whether mean delivery times differed significantly between product types.

Hypotheses:

- **H_0 (null):** All product types have the same mean delivery time.
- **H_1 (alternative):** At least one product type has a different mean delivery time.

The analysis was performed in R using a one-way ANOVA, followed by a Tukey HSD post-hoc test.

6.2 Results

Product Type	n	Mean Delivery (h)	SD (h)
CLO	15 598	21.7	6.11
KEY	17 920	21.7	6.09
LAP	10 207	21.8	6.05
MON	14 864	21.7	6.05
MOU	20 662	21.8	6.14
SOF	20 749	1.09	0.31

The one-way ANOVA produced:

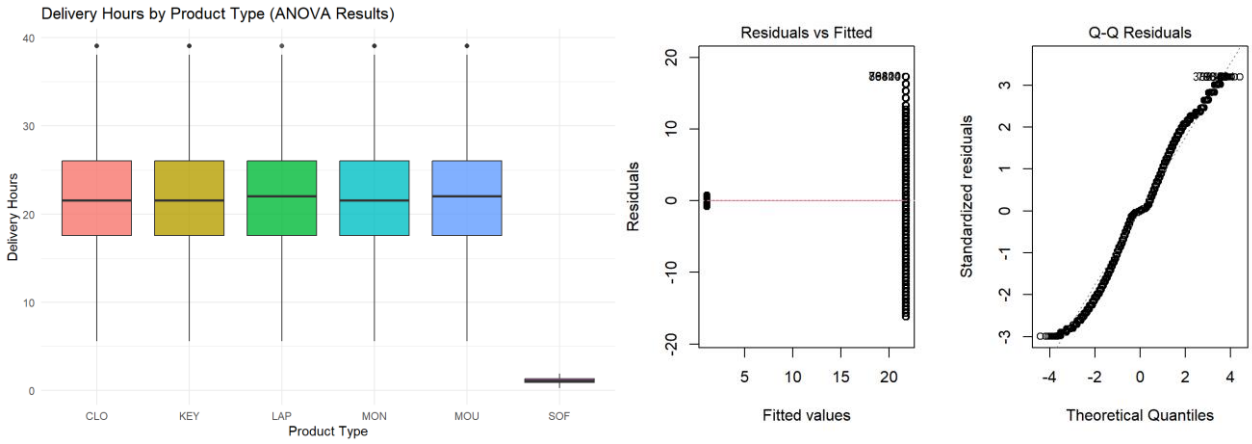
- $F = 47\,694$, $p < 2 \times 10^{-16}$ (*)**, indicating a highly significant effect of product type on delivery hours.
- **Levene's Test:** $F = 7\,436.3$, $p < 0.001$ → unequal variances (expected due to SOF's much smaller scale).
- **Tukey HSD:** SOF differed significantly ($p < 0.001$) from all other product types; differences among CLO, KEY, LAP, MON, and MOU were not significant.

6.3 Discussion

The ANOVA results show that Product Type has a statistically significant effect on Delivery Hours. The Tukey post-hoc test identified that SOF products are distinctly different from all other product categories, with mean delivery times around 1 hour compared to approximately 22 hours for the remaining products.

These findings confirm that the SOF delivery process operates on a much faster and more stable timescale, which aligns with the SPC results in Part 3. In that section, SOF exhibited much smaller control limits and the highest process capability indices (Cp ≈ 18.1, Cpk ≈ 1.08), reinforcing the conclusion that the SOF process is highly capable and consistent. The remaining product categories demonstrate similar performance, suggesting stable but slower processes across these product lines.

6.4 Graphical Representation



The plot clearly shows that the SOF group has an extremely low and tightly clustered distribution of delivery hours compared with the other product types, visually confirming the statistical results.

Summary of Findings:

Aspect:	**Observation:**
Factor tested	Product Type
Response variable	Delivery Hours
F-value	47 694
p-value	< 0.001
Significant difference	Yes (SOF vs others)
Main insight	SOF deliveries are much faster and more capable than the rest

Conclusion:
The DOE and ANOVA confirm that product type significantly influences delivery performance. Only the SOF product category exhibits a distinctly different process, operating with extremely short and consistent delivery times, validating the SPC conclusions from Part 3.

7. Reliability of Service (Part 7)

7.1 Reliability of Service

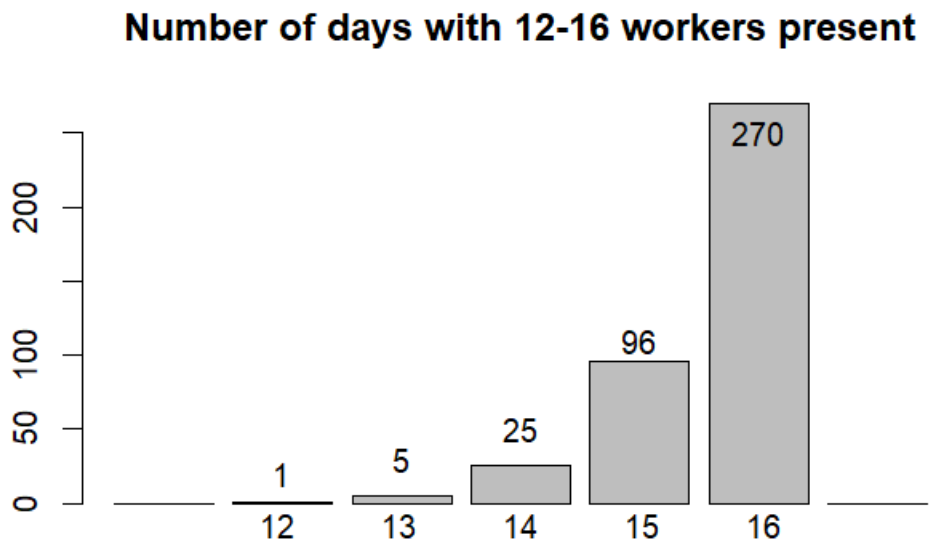


Figure 7.1 – Number of Days with 12–16 Workers Present

The supplied frequency distribution for days with 12–16 workers present is:

Workers on duty	Number of days
12	1
13	5
14	25
15	96
16	270

Total observations = 397 days.

Assuming that reliable service is achieved when 15 or more workers are present, the number of reliable days is the sum of days with 15 and 16 workers:

$$\text{Reliable days} = 96 + 270 = 366$$

The overall empirical probability of reliable service is therefore:

$$P(\text{reliable}) = \frac{366}{397} = 0.921914 \approx \mathbf{92.19\%}$$

Interpretation: Over the observed period the service reliability was approximately 92.2%. This is high but falls slightly short of a common target of 95%. The shortfall is caused by the small number of days with fewer than 15 workers (12–14 workers occurred on 31 of 397 days ≈ 7.81% of days). To raise reliability to 95% the company must reduce the frequency of low-staff days (for example by improved scheduling or adding staff).

7.2 Optimising Profit for the Company

Assumptions :

- Problematic day (service problems) occurs when fewer than 15 workers are present.
- Loss per problematic day = R20 000 in lost sales.
- Additional personnel cost = R25 000 per month per person → daily cost = 25 000 / 30 = R833.33 per person per day.
- Observed frequency (from 7.1): 31 problematic days out of 397 total days.

A. Baseline expected daily loss (no additional staffing)

$$P(\text{problem day}) = \frac{31}{397} = 0.0780856 \approx 7.81\%$$

$$\text{Expected daily lost sales} = 0.0780856 \times 20\,000 \approx \mathbf{R1\,561.71\ /day}$$

This R1,561.71/day is the current expected cost to the business from under-staffing (lost sales).

B. Option 1 — Hire k permanent extra staff (k = 1, 2, 3)

If k permanent staff are added, a day is still problematic when workers + k < 15. Using the observed day counts, we compute expected loss after hiring and the added daily hire cost:

k (extra hires)	Problematic days after hire	Expected loss (R/day)	Hire cost (R/day)	Total expected cost (R/day)
0	31	1,561.71	0.00	1,561.71
1	6	302.27	833.33	1,135.60
2	1	50.38	1,666.67	1,717.04
3	0	0.00	2,500.00	2,500.00

Calculation summary (key values):

- Hiring 1 extra permanent staff reduces expected daily total cost from R1,561.71 to R1,135.60, a saving of R426.11/day.
- Hiring 2 or more permanent staff increases total expected cost relative to baseline (not cost-effective under these assumptions).

C. Option 2 — Use on-call / temporary staff only on deficit days

Compute the average number of extra staff required per day to bring under-staffed days up to 15:

Extra staff needed per problematic day:

- For 12 workers: need 3 extra (count = 1 day) → 3 × 1 = 3
- For 13 workers: need 2 extra (count = 5 days) → 2 × 5 = 10
- For 14 workers: need 1 extra (count = 25 days) → 1 × 25 = 25

Total extra-staff-days = 3 + 10 + 25 = 38 extra-staff-days across 397 days.

Average extra staff needed per day:

$$\text{Avg extra staff/day} = \frac{38}{397} \approx 0.09572 \text{ staff/day}$$

Expected daily on-call staff cost (using R833.33/day per staff):

$$\text{On-call cost/day} = 0.09572 \times 833.33 \approx \mathbf{R79.76 \text{ /day}}$$

If temporary staff fully eliminate the lost sales on days they are used, the only incremental expected cost is the on-call staffing cost, ≈ R79.76/day, which is far lower than the baseline expected lost-sales cost (R1,561.71/day) and much lower than the permanent-hire alternatives.

D. Recommendation

1. **Best (cost) option — on-call / temporary staff:**
   If logistically feasible, implement an on-call/temp pool to cover deficits. Expected incremental cost is approximately R79.76 per day, while effectively eliminating the expected lost sales (assuming temps can be scheduled reliably). This yields by far the lowest expected total cost under the given assumptions.
2. **If on-call staff are not feasible or are significantly more expensive in practice:**
   Hire one permanent extra staff member. This reduces expected total cost to R1,135.60/day (a saving of R426.11/day compared to no hire). Hiring two or more permanent staff is not cost effective under the current cost/loss parameters.

3. **Operational recommendation:** improve rostering practices to reduce the small number of low-staff days (12–14 workers). Even modest reductions in those days will raise reliability toward the 95% target and reduce expected lost sales.

E. Caveats & sensitivity

- All results assume the observed 397-day distribution is representative of future operations. If future staffing variability changes, results should be recalculated.
- The on-call cost model assumes temporary staff are available at the same pro-rata daily rate; in reality, short-notice temps may cost more or have minimum-day charges. If so, re-evaluate the on-call option using real temp rates.
- The R25 000/month cost and R20 000 lost-sales per problematic day are assumed constants provided in the brief; change either parameter and re-evaluate.

Conclusion :

Under the project assumptions, the most cost-effective strategy is to implement an on-call / temporary staff pool to cover shortages. If that is not feasible, the next-best action is to hire one permanent extra staff member and to improve scheduling to avoid the rare days with very low staffing.

8. References

Department of Industrial Engineering, Stellenbosch University, 2025. *Project ECSA 2025: QA344 – Quality Assurance*. Stellenbosch: Stellenbosch University.

Department of Industrial Engineering, Stellenbosch University, 2025. *QA344 Statistics (1) – Course Notes*. Stellenbosch: Stellenbosch University.

Montgomery, D.C., 2020. *Introduction to Statistical Quality Control.* 8th ed. Hoboken, NJ: John Wiley & Sons.

Oakland, J.S., 2014. *Statistical Process Control.* 7th ed. London: Routledge.

Ross, S.M., 2021. *Introduction to Probability and Statistics for Engineers and Scientists.* 6th ed. Amsterdam: Academic Press.

Taguchi, G., Chowdhury, S. and Wu, Y., 2005. *Taguchi's Quality Engineering Handbook.* Hoboken, NJ: John Wiley & Sons.

Wheeler, D.J. and Chambers, D.S., 2010. *Understanding Statistical Process Control.* 3rd ed. Knoxville, TN: SPC Press.

R Core Team, 2024. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.r-project.org/ [Accessed 22 October 2025].