ECSA Final Report

Quality Assurance 344

Jennifer le Roux

26865017

24 October 2025

Contents

Introduction……..4
Part 1: Descriptive Statistics 4
Data Overview and Inspection 4
Summary Statistics 6
Customer Data 6
Product Data 6
Sales Data 7
Missing values 7
Data Filtering and Subsetting 7
Data Visualisation 8
Exploring Relationships 10
Data correction (From section 4.3) 10
Part 3: Statistical Process Control 14
3.1 Initialisation 15
3.2 Monitoring ongoing samples 18
3.3 Process Capability Indices 19
3.4 Identifying process control issues 20
Part 4: Risk, Data correction and optimising for maximum profit 22
4.1 Probability of making a type I (Manufacture's) error 22
4.2 Probability of making a type II (Customer's) error……..22
Part 5: Optimising for Maximum Profit 24
Coffee shop 1 24
Coffee Shop 2 26
Part 6: DOE and ANOVA 29
ANOVA Results 29
Part 7: Reliability of Service 31
7.1 Estimation of the proportion of reliable service days per year 31
7.2 Optimising profit for the company 31
Conclusion……..33
References……..34

List of Figures

Figure 1 Snippet of customer_data 4
Figure 2 Snippet of product_data 4
Figure 3 Snippet of product_Headoffice data 4
Figure 4 Snippet of sales data 5
Figure 5 Cross-tabulation of ProductID prefixes against product categories, ... 5
Figure 6 Showing instances displaying conflicting information between the product_data and Products_Headoffice Datasets 6
Figure 7 Statistical summary of categorical customer data 6
Figure 8 Statistical summary of numerical customer data 6
Figure 9 Statistical Summary of categorical product data 6
Figure 10 Statistical summary of numerical product data 6
Figure 11 Statistical Summary of categorical product head office data 6
Figure 12 Statistical summary of numerical product head office data 6
Figure 13 Statistical Summary of categorical sales data 7
Figure 14 Statistical summary of numerical sales data 7
Figure 15 Subset of products with a markup greater than 20% 7
Figure 16 Mean selling price of each category 8
Figure 17 Histogram of the distribution of customer income 8
Figure 18 Line graph of the total sales in 2022 compared to 2023 9
Figure 19 Scatterplot of selling price vs markup 9
Figure 20 Bar chart of total sales quantity by city 10
Figure 20 Comparison of total 2023 sales before and after data correction 11
Figure 21 Histogram of selling price distribution of corrected vs uncorrected data 11
Figure 22 Scatter plot of selling price vs markup of the corrected vs uncorrected data 12
Figure 23 Histogram of Total Sales Value by Product Type of corrected vs uncorrected data 12
Figure 24 Demonstration of the effect of the Central Limit Theory 14
Figure 25 Control charts for mouse (MOU) 15
Figure 26 Control charts for Keyboards (KEY) 15
Figure 27 Control charts for Software (SOF) 16
Figure 28 Control charts for Cloud Subscription (CLO) 16
Figure 29 Control charts for Laptops (LAP) 17
Figure 30 Control charts for Monitors (MON) 17
Figure 31 Control charts on the full dataset for each product type 18
Figure 32 Process capability (first 1000 deliveries) 19
Figure 33 Product types violating rule A 20
Figure 34 Product types violating rule B 20
Figure 35 Product types violating rule C 21
Figure 36 Metric averages per barista count for coffee shop 1 24
Figure 37 Profit per day vs number of baristas for coffee shop 1 25
Figure 38 Average service time vs number of baristas for coffee shop 1 25
Figure 39 Reliable service vs number of baristas 26
Figure 40 Metric averages per barista count for coffee shop 2 26
Figure 41 Box plot of customer waiting time vs number of baristas for shop 2 27
Figure 42 Line graph of Daily profit vs number of baristas for shop 2 27
Figure 43 Line graph showing optimal number of baristas for maximum profit 28
Figure 44 ANOVA Table 30
Figure 45 Box plot of delivery hours of 2022 vs. 2023 30
Figure 46 Histogram of number of days vs. number of workers present 31
Figure 47 Total extra cost vs. number of workers present 32

Introduction

This report provides a comprehensive analysis of multiple datasets for businesses in the restaurant and retail sectors. It applies descriptive statistics, statistical process control (SPC), design of experiments (DOE), and optimisation methods to evaluate data accuracy, fix erroneous data, process performance and profitability. The aim for each section is to recommend strategies driven by data analytics to improve decision-making and business outcomes.

Part 1: Descriptive Statistics

Data Overview and Inspection

The data provided for this analysis includes *customer_data,* providing demographic information about the customer base including customer gender, age, income, and city of residence. The *product_data* and *products_Headoffice* dataset includes information about the categories of the products, product descriptions, selling prices and markups. Lastly, the *sales_data* covers the years 2022 and 2023, linking the customers to the products that were bought alongside quantities and delivery times.

	CustomerID <chr>	Gender <chr>	Age <int>	Income <dbl>	City <chr>
1	CUST001	Male	16	65000	New York
2	CUST002	Female	31	20000	Houston
3	CUST003	Male	29	10000	Chicago
4	CUST004	Male	33	30000	San Francisco
5	CUST005	Female	21	50000	San Francisco
6	CUST006	Male	32	80000	Miami

Figure 1 Snippet of customer_data

	ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
1	SOF001	Software	coral matt	511.53	25.05
2	SOF002	Cloud Subscription	cyan silk	505.26	10.43
3	SOF003	Laptop	burlywood marble	493.69	16.18
4	SOF004	Monitor	blue silk	542.56	17.19
5	SOF005	Keyboard	aliceblue wood	516.15	11.01
6	SOF006	Mouse	black silk	478.93	16.99

Figure 2 Snippet of product_data

	ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
1	SOF001	Software	coral silk	521.72	15.65
2	SOF002	Software	black silk	466.95	28.42
3	SOF003	Software	burlywood marble	496.43	20.07
4	SOF004	Software	black marble	389.33	17.25
5	SOF005	Software	chartreuse sandpaper	482.64	17.60
6	SOF006	Software	cornflowerblue marble	539.33	25.57

Figure 3 Snippet of product_Headoffice data

	CustomerID <chr>	ProductID <chr>	Quantity <int>	orderTime <int>	orderDay <int>	orderMonth <int>	orderYear <int>	pickingHours <dbl>	deliveryHours <dbl>
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

Figure 4 Snippet of sales data

Upon viewing the data, we see that each product ID is given a specific prefix. Using the substr() function, we are able to extract the first three characters from the product IDs. We see that the prefixes include “SOF”, “CLO”, “LAP”, “MON”, “KEY”, “MOU”, representing Software, Cloud Subsciption, Laptop, Monitor, Keyboard and Mouse, respecively. This suggests that the IDs are systematically coded, simplifying the process for filtering and grouping data. This assumption was confirmed in Rstudio by checking that the prefix of each ProductID matches the product Category, the following output was obtained:

	Cloud Subscription	Keyboard	Laptop	Monitor	Mouse	Software
CLO	2	2	1	1	2	2
KEY	2	2	1	1	2	2
LAP	1	2	2	2	2	1
MON	2	1	2	2	1	2
MOU	1	2	2	2	2	1
SOF	2	1	2	2	1	2

Figure 5 Cross-tabulation of ProductID prefixes against product categories,

The analysis reveals that the prefixes are not consistent with the product categories, as each prefix is spread across multiple categories rather than being unique to one. For this reason, prefixes can only serve as a rough guide to category and should not be treated as a definitive classification.

Moreover, it was noted that while several instances in the *products_data* and *products_Headoffice* datasets share the same *Product_ID*, they tend to differ in descrition and selling price. This is already evident when comparing the first instance of figure 2 and 3 above. The *Product_ID* “SOF001” is descibed as *coral matt* with a selling price of 511.53 in product_data but as *coral sink* and with a selling price of 521.72 in *product_Headoffice*. This erroneous data causes ambiguiting around product pricing in the *sales2022and2023* dataset, as it is not clear which product the purchase refers to. This mismatch could be due to regional variations (head office vs local branch), outdated catalogue versions, or simply data entry errors. Clarrification of business records is required in future, indicating whether sales reflect branch-level prices or head office prices, for now, these discrepencies will be flagged for review. The following figure provides a list of all conflicting instances between these two datasets, comparing the features between the local instances (*product_data*) and headquarter or “hq” instances (*products_Headoffice*).

	ProductID	Category_local	Description_local	SellingPrice_local	Markup_local	Prefix	Category_hq	Description_hq	SellingPrice_hq	Markup_hq
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<dbl>	<dbl>
1	SOF001	Software	coral matt	511.53	25.05	SOF	Software	coral silk	521.72	15.65
2	SOF002	Cloud Subscription	cyan silk	505.26	10.43	SOF	Software	black silk	466.95	28.42
3	SOF003	Laptop	burlywood marble	493.69	16.18	SOF	Software	burlywood marble	496.43	20.07
4	SOF004	Monitor	blue silk	542.56	17.19	SOF	Software	black marble	389.33	17.25
5	SOF005	Keyboard	aliceblue wood	516.15	11.01	SOF	Software	chartreuse sandpaper	482.64	17.60
6	SOF006	Mouse	black silk	478.93	16.99	SOF	Software	cornflowerblue marble	539.33	25.57

Figure 6 Showing instances displaying conflicting information between the product_data and Products_Headoffice Datasets

Summary Statistics

Customer Data

	skim_variable	n_missing	complete_rate	min	m...	empty	complete_rate	min	max	empty	n_unique	whitespace
	<chr>	<int>	<dbl>	<int>	<int>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>
1	CustomerID	0	1	7	8	0	1	7	8	0	5000	0
2	Gender	0	1	4	6	0	1	4	6	0	3	0
3	City	0	1	5	13	0	1	5	13	0	7	0

Figure 7 Statistical summary of categorical customer data

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	Age	0	1	51.5538	21.2161	16	33	51	68	105	▇▇▇...
2	Income	0	1	80797.0000	33150.1067	5000	55000	85000	105000	140000	▂▇▁...

Figure 8 Statistical summary of numerical customer data

The customer dataset is complete, having no missing values. The ages of the customers range from 16 to 105 years old, averaging at about 52 years, suggesting that most of the customer base is middle-aged. There is a substantial variation in customer income levels, ranging from 5000 to 140000, with a mean of about 80800. A strong variation in customer purchasing power is indicated by the wide standard deviation of customer income levels (approximately 33150), highlighting the contrast between the high- and low-income groups. It is important that this variability is considered when designing products, marketing and implementing promotions.

Product Data

	skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
	<chr>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>
1	ProductID	0	1	6	6	0	60	0
2	Category	0	1	5	18	0	6	0
3	Description	0	1	9	21	0	35	0

Figure 9 Statistical Summary of categorical product data

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	SellingPrice	0	1	4493.59283	6503.770150	350.45	512.1825	794.185	6416.6600	19725.18	▇▁▁▁▁
2	Markup	0	1	20.46167	6.072598	10.13	16.1400	20.335	25.7075	29.84	▇▇▁...

Figure 10 Statistical summary of numerical product data

	skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
	<chr>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>
1	ProductID	0	1	5	6	0	110	0
2	Category	0	1	5	18	0	6	0
3	Description	0	1	9	24	0	60	0

Figure 11 Statistical Summary of categorical product head office data

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	SellingPrice	0	1	4410.9619	6463.822788	290.52	495.9375	797.215	5843.332	22420.14	▇▁▁▁▁
2	Markup	0	1	20.3855	5.665949	10.06	15.8400	20.580	24.845	30.00	▇▇▁...

Figure 12 Statistical summary of numerical product head office data

Both the *product_data* and *products_Headoffice* datasets are complete with no missing values. An extreme variation is seen in the selling price of both datasets, with *product_data,* ranging from approximately 350 to 19725, with a mean of about 4494, and *Products_Headoffice* data has a range of 291 to 22420 and a mean of about 4411. This suggests that only a small number of very high-priced items in the catalogue. Markups are more stable, averaging around 20% with an interquartile range of 16-25%. As before mentioned, the differences between the two datasets suggest some inconsistencies which need to be resolved to ensure accuracy when linking sales to product pricing.

Sales Data

	skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
	<chr>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>
1	CustomerID	0	1	7	8	0	5000	0
2	ProductID	0	1	6	6	0	60	0

Figure 13 Statistical Summary of categorical sales data

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p7	p100	hist
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dt	<dbl>	<chr>
1	Quantity	0	1	13.50347	13.7601316	1.0000000	3.000000	6.000	23.000(	50.0000	▇▂▁▁▁
2	orderTime	0	1	12.93230	5.4951268	1.0000000	9.000000	13.000	17.000(	23.0000	▃▃▆▇▇
3	orderDay	0	1	15.49683	8.6465055	1.0000000	8.000000	15.000	23.000(	30.0000	▇▇▇▇▇
4	orderMonth	0	1	6.44813	3.2834460	1.0000000	4.000000	6.000	9.000(	12.0000	▇▇▇▇▇
5	orderYear	0	1	2022.46273	0.4986115	2022.0000000	2022.000000	2022.000	2023.000(	2023.0000	▇▁▁▁▆
6	pickingHours	0	1	14.69547	10.3873345	0.4258889	9.390833	14.055	18.721(	45.0575	▇▆▂▁▁
7	deliveryHours	0	1	17.47646	9.9999440	0.2772000	11.546000	19.546	25.044(	38.0460	▇▇▅▂▁

7 rows | 1-10 of 11 columns

Figure 14 Statistical summary of numerical sales data

The sales dataset is also complete, with no missing values. A wide variation is seen in order quantities, ranging from purchases of a single unit up to 50 units, with a mean of 13.5 and a median of 6, indicating a right-skewed distribution that is dominated by smaller orders. Looking at the time related features, one can see that the data represents 2022 and 2023 equally. *PickingHours* and *DeliveryHours* average around 14.7 and 17.5 respectively, while both show a significant variation with some picking times taking up to 45 hours and some delivery times taking up to 38 hours. This large range may indicate inefficiencies in the process or variations in the logistics that should be investigated further.

Missing values

Across all four datasets, the summary statistics show a complete rate of 1 for each variable and no missing values. Because of this, the data preparation process is simplified, and no imputation or removal of instances or features are required. Without missing data, the dataset is more reliable and ensures that observations made on the dataset are not influenced by gaps in the dataset.

Data Filtering and Subsetting

Data filtering and subsetting aids in the segmenting, cleaning and analysing of data to answer meaningful business questions. Products with a markup of more than 20% are analysed below. This subset helps to identify high-margin products. With these insights we can perform profitability analyses, which aid in the decision-making process with regards to promotions and inventory prioritisation.

	ProductID	Category	Description	SellingPrice	Markup	Prefix
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<chr>
1	SOF001	Software	coral matt	511.53	25.05	SOF
10	SOF010	Monitor	chocolate sandpaper	396.72	23.47	SOF
14	CLO014	Cloud Subscription	burlywood silk	1083.11	21.25	CLO
15	CLO015	Laptop	azure silk	728.26	27.70	CLO
18	CLO018	Mouse	chocolate matt	1105.66	20.23	CLO
19	CLO019	Software	aliceblue silk	1092.07	23.14	CLO

Figure 15 Subset of products with a markup greater than 20%

The filtered data shows that products with a mark-up greater than 20% are spread across several categories, including Software, Monitor, Cloud Subscription, Laptop and Mouse. This indicates that higher-margin items are not limited to a single product type. With that, the selling prices vary widely from about 397 for the monitor and 1100 for mouse and software, this suggests that the markup percentage does not directly depend on the absolute price of the product. Instead, it is likely that is reflects strategic pricing decisions aimed at maximising profitability in different product lines. Subsets like this are useful for identifying categories which contribute most to profit margins and for guiding strategies for promotions and pricing.

Category <chr>	MeanSellingPrice <dbl>
Laptop	5217.545
Monitor	5014.170
Keyboard	4638.172
Mouse	4585.465
Software	4181.323
Cloud Subscription	3691.861

Figure 16 Mean selling price of each category

An analysis of the mean selling price across all categories shows that Laptops have the highest average price of approximately 5217.55. Cloud Subscription has the lowest mean of 3691.86. These results indicate that hardware products generally have higher selling prices compared to software, which could have an influence on inventory, marketing and pricing strategies.

Data Visualisation

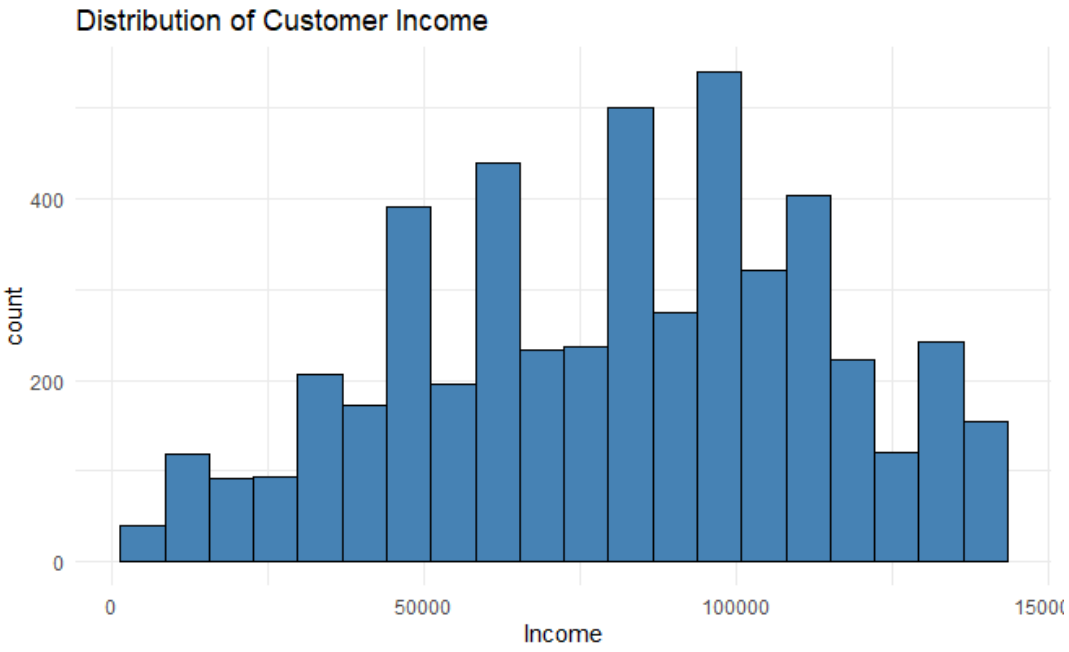


Figure 17 Histogram of the distribution of customer income

From the histogram below, it is clear that most of customers have an income level arround 100000 with only about 150 customers earning near to 150 000 and only a small portion earning less than

25000. From this distribution it would be sensible to market towards middle to higher income earners as the majority of the customer base falls into this category.

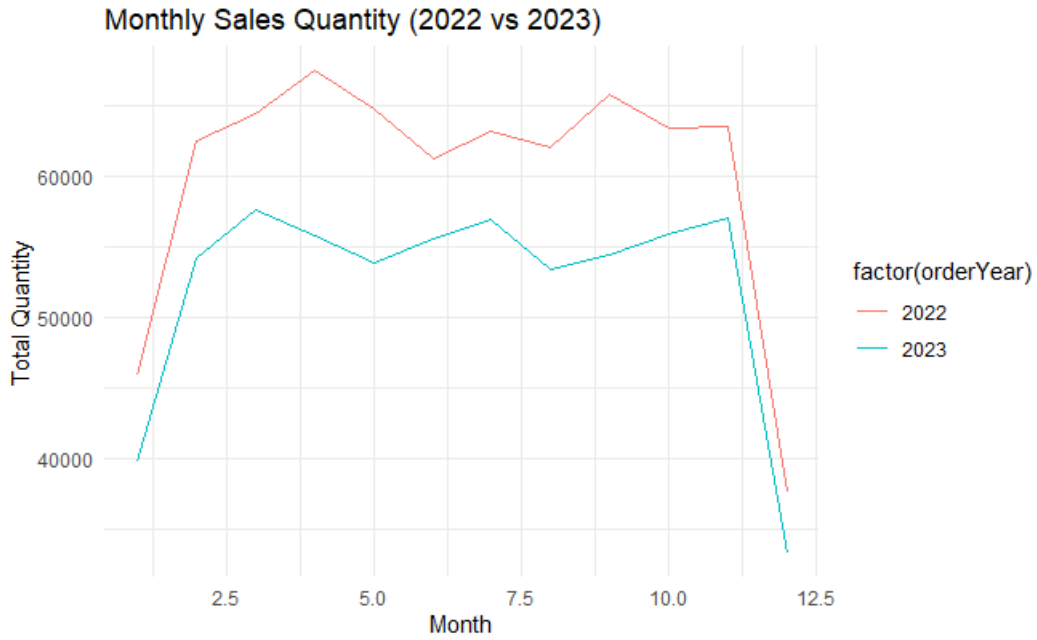


Figure 18 Line graph of the total sales in 2022 compared to 2023

From the line graph above the total quantity of sales was much lower in 2023 than in 2022. While sales decreased, similar seasonal patterns were observed over the months of both years, indicating months where sales spiked (March to April and September to October). 2022 consistently outperformed the sales of 2023, suggesting that the overall demand or customer engagement declined in 2023 compared to the previous year, despite following a similar trend over time.

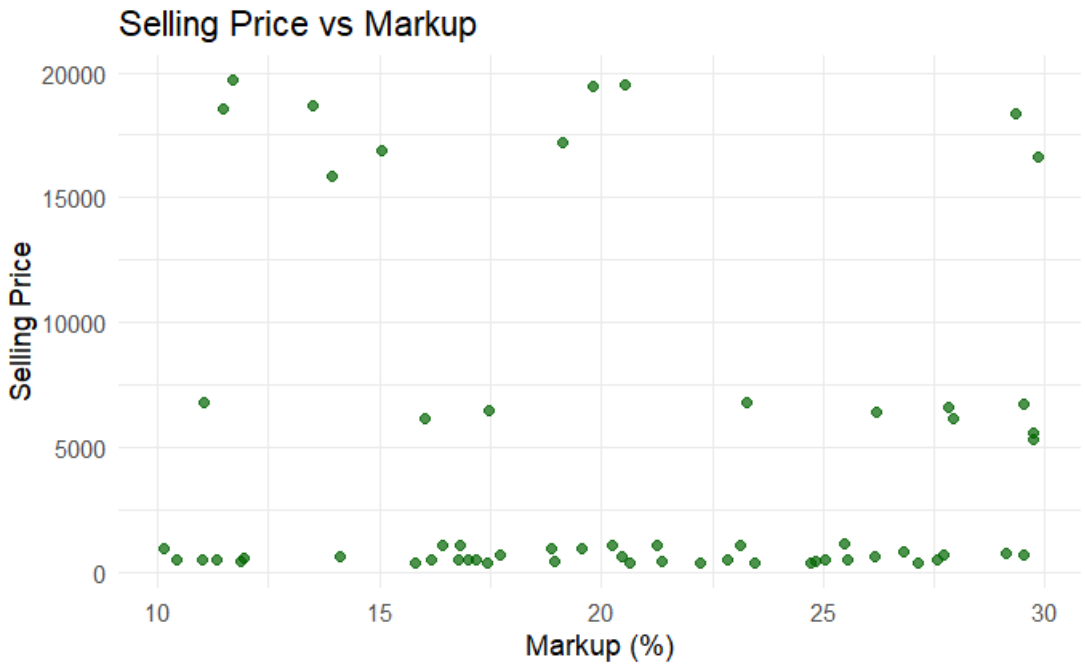


Figure 19 Scatterplot of selling price vs markup

The scatterplot shows that there is no linear relationship between the selling price and mark-up percentage. Products with higher prices (above 10000) appear across a wide range of markup percentages, while lower priced items are clustered across the full markup range (10%-30%). This suggests that the markup decisions are likely to not be tied to product price as premium products do not always carry higher markups and lower priced items can sometimes have relatively high margins.

Exploring Relationships

Figure 20 Bar chart of total sales quantity by city

The bar chart shows variation in sales quantities across different cities. Los Angeles and San Francisco stand out with the highest total sales, followed by New York and Seattle. Meanwhile, Miami and Chicago report comparatively lower sales volumes. This indicates that certain cities (particularly on the West Coast) are driving a larger share of sales, which may reflect larger customer bases, higher demand, or stronger company presence in those regions.

The bar chart shows variation in sales quantities across different cities. Los Angeles and San Francisco stand out with the highest total sales, followed by New York and Seattle with Miami and Chicago having comparatively lower sales volumes. This distribution highlights that two cities (interestingly, both on the west coast) drive a large share of the sales, indicating a larger customer base, higher demand or a stronger company presence in those regions.

Data correction (From section 4.3)

Following the identification of inconsistencies between the product and head office datasets, the company provided updated instructions to correct the errors. According to their correspondence, only the first ten entries for each product type were accurate, while the remaining entries required updating. Specifically, the first 10 values of *SellingPrice* and *Markup* per product type were repeated as instructed, and all *NA* prefixes were replaced with the correct product type (CLO, KEY, LAP, MON, MOU, SOF). The updated files *products_data2025.csv* and *products_Headoffice2025.csv*, were then used for reanalysis to evaluate the impact of these corrections on total sales and overall data consistency. Categories like Software (SOF) and Laptops (LAP) show slightly higher sales totals

where head office data had previously underestimated prices. Overall, the updated datasets allow for consistent descriptive statistics, reliable visualizations, and correct calculation of total sales, removing ambiguity and improving the quality of insights for business decision-making.

The figure below reflects the total sales value for 2023 with updated prices compared to the original dataset.

ProductType <chr>	TotalQuantity_Uncorrected <dbl>	TotalSalesValue_Uncorrect... <dbl>	TotalQuantity_Corrected <dbl>	TotalSalesValue_Corrected <dbl>
LAP	64414	0	64414	1163889479
MON	91782	0	91782	578385570
CLO	96691	0	96691	98715482
KEY	114357	0	114357	73499067
SOF	131349	61821700	131349	66468485
MOU	129613	0	129613	51219577

Figure 20 Comparison of total 2023 sales before and after data correction

In the original data analysis, with the uncorrected head office data, total sales values for most product types were severely underestimated, with only Software (SOF) showing a non-zero total value. Categories like Laptops (LAP), Monitors (MON), Keyboards (KEY), Mice (MOU), and Cloud Subscriptions (CLO) all recorded zero sales value because their ProductIDs did not match the prices in the original dataset.

After correcting the product data, repeating the first 10 prices and markups for all items, the total sales values were accurately reflected. Laptops (LAP) and Monitors (MON) were revealed to be the dominant contributors, together accounting for most of the revenue, followed by Keyboards, Software, Cloud Subscriptions, and Mice. This clearly illustrates how data inconsistencies can distort sales analysis and emphasises the importance of maintaining correct and complete product information for revenue reporting. The following visualisations further highlight this point, the first three are repeats of the initial plots that have been adapted to the new data.

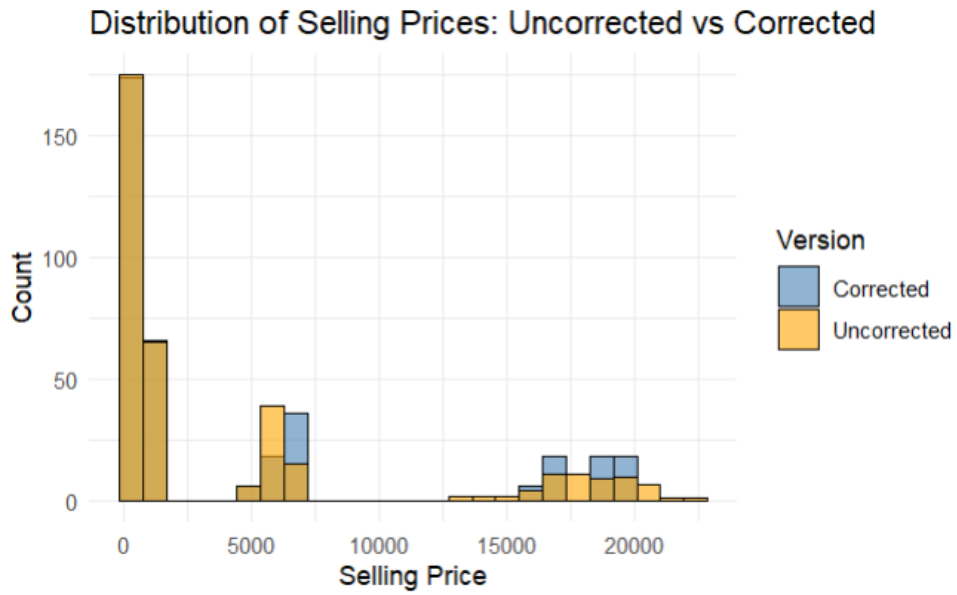


Figure 21 Histogram of selling price distribution of corrected vs uncorrected data

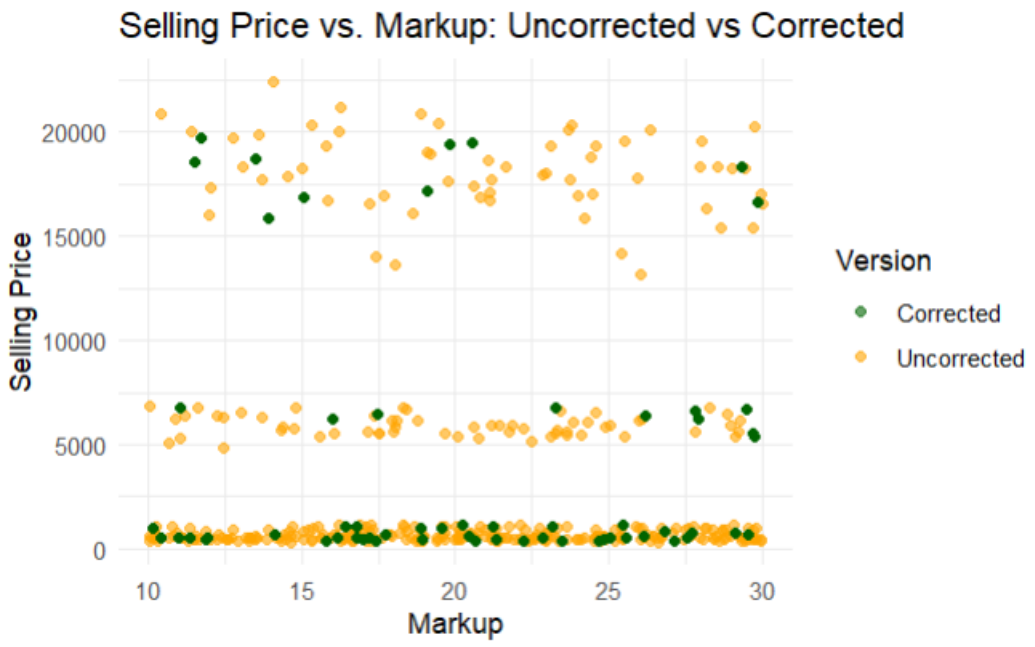


Figure 22 Scatter plot of selling price vs markup of the corrected vs uncorrected data

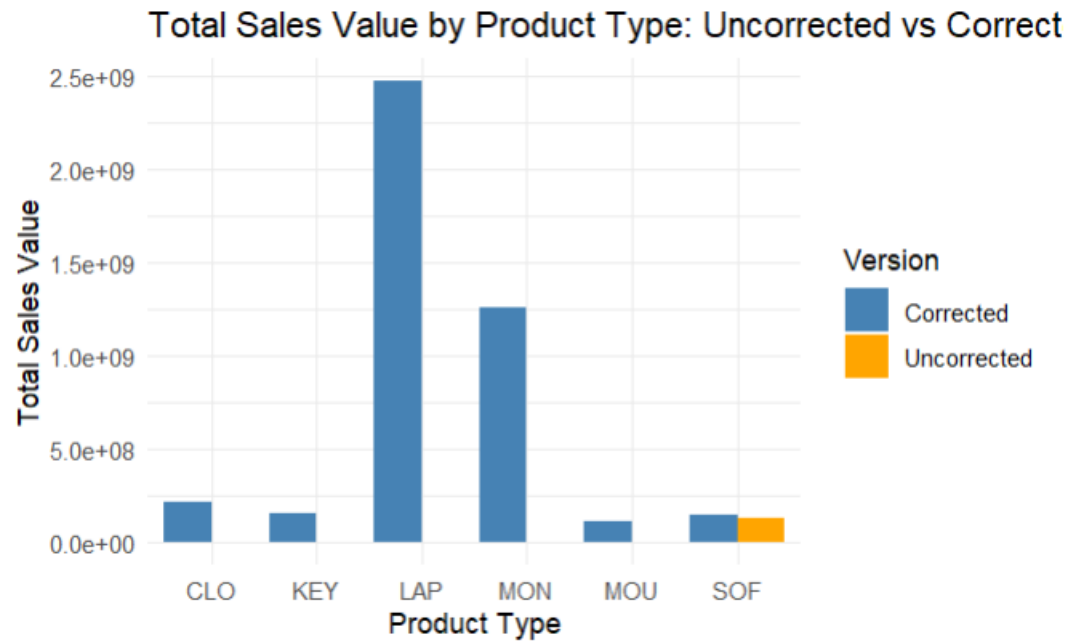


Figure 23 Histogram of Total Sales Value by Product Type of corrected vs uncorrected data

The histogram of the distribution of selling prices (figure 21) reveals a heavy skew towards low selling prices for the uncorrected data (yellow), which is to be expected due to the errors in the data causing underreporting. After correcting the data, we see a more balanced distribution with moderate counts across the various selling prices, this suggests successful normalisation that better reflects the true price variability and reduces bias.

The scatter plot of Selling Price vs. Markup (figure 22) shows the uncorrected data (orange points) to cluster tightly around low markups (10 to 15%) with limited price dispersion, implying markup inconsistencies or outliers that distort pricing relationships. The corrected data (green points) reveals a clearer positive correlation, with higher dispersion and a more upward trend, revealing higher

markups with higher prices with little overlap. For this reason, it is confirmed that the updated dataset has improved data quality for deriving reliable relationship between pricing and markups, as well as actional insights into pricing strategies.

The total sales value by product type in figure 23 is expected upon review of the comparison table in figure 20, where it was highlighted that the dataset errors lead to there only being recorded sales value for SOF (software). Now that the data is corrected (blue), we can see an accurate representation of the spread of sales across the product types. We see significantly elevated sales for laptops and moderate sales for monitors, indicating the dominant sales drivers of the product types. This data correction reveals the true diversity of sales which prevents the possibility of an over-reliance on the software metrics for business strategy.

Final recommendations

It is advised to prioritise the corrected dataset for all analytical and strategic planning, as it reveals the true, diversified revenue stream and will prevent important decisions to be made on incorrect information. It is recommended to audit raw data pipelines immediately for misclassification issues, as well as to invest in embedding validation points into the data gathering process to avoid large data errors in future. Moreover, the results of the data analysis reveal that Laptop and Monitor sales should be prioritised, due to their significantly higher sales. Markup dynamics were clarified, showing healthy markup relationships. The insights generated here can be leveraged for targeted actions, such as optimising laptop and monitor pricing to sustain high markups, reallocating inventory towards top performers, and accurately forecasting growth to reach untapped revenue levels.

Part 3: Statistical Process Control

The central limit theory (CTL) strengthens Statistical Process Control (SPC) as for any population distribution, large random samples (with n greater than 20), yield approximate normal sample means. This enables "3-sigma" limits and rules despite the unknown individual shapes of the original data. This process normalises subgroups, allowing one to distinguish random variation from special causes as it reduces the spread of the data and minimises type I and II errors. The figures below are used to illustrate this:

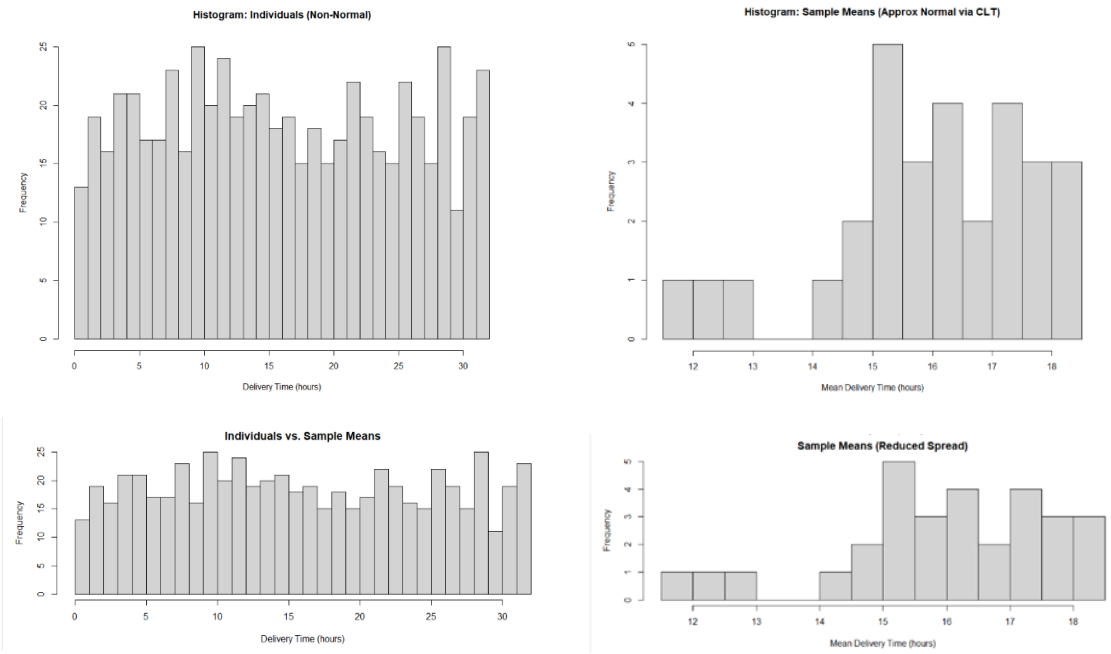


Figure 24 Demonstration of the effect of the Central Limit Theory

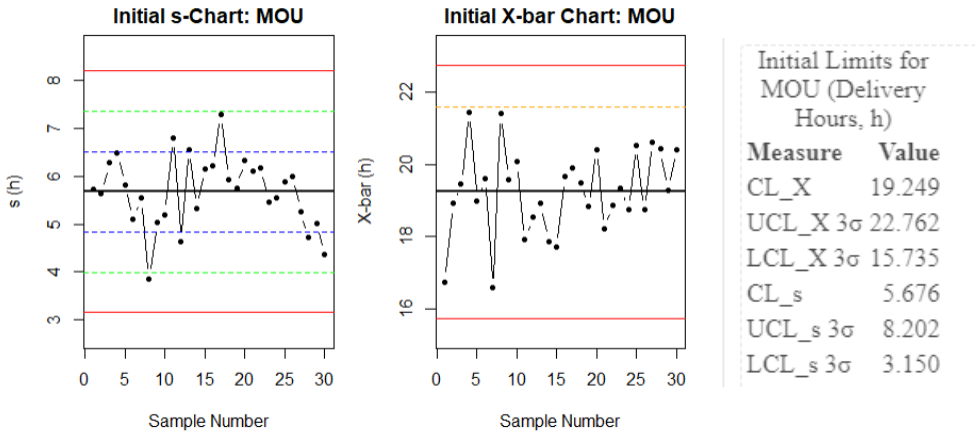
The histogram titled "Individuals (Non-Normal)" (top left), shows a rectangular distribution with frequencies ranging between about 15 to 25 per bin across 0 to 35 hours. This conforms the presence of uniformity, highlighting the equal probability of each interval. We see that there is no bell shape, or skew, further enforcing that there are no trends present. This unknow shape would validate the direct normality cased limits. Because random variation dominates the dataset, SPC rules such as out-of-control signals can't be applied reliably without CLT.

The histogram "Individuals vs. Sample Means" (bottom left) shows a wide uniform spread, whereas the histogram "Sample Means (Reduced Spread)" on the bottom right of the figure, shows a distribution with some peaks and a reduced spread. We see that the means cluster around 15 to16 hours, illustrating the "averaging" effect of CLT, where random errors are cancelled. No correlations are evident (independent samples). This histogram highlights the high variability in individual data (VOP), and how subgroups (VOSp) enable the detection of trends.

Lastly, the histogram titled "Sample Means (Approx Normal via CLT)" (top right), transforms to a bell-shaped normal distribution, peaking at a frequency of 5 at 15 to 16 hours. This plot mirrors a sample means or X-bar distribution seen in control charts, with a subgroup size of n=24 (greater than the typical threshold of 20), the normality is even more reliable, further supporting the use of 3-sigma upper and lower control limits (UCL/LCL).

3.1 Initialisation

The data was confirmed to be sorted in order of orderYear, orderMonth, orderDay, and orderTime to simulate sequential, real time order arrivals. Using a sample size of 24, the oldest 30 samples were used to initialise X-bar and s-charts. Here we can estimate centre lines, where CL_X represents the mean of subgroup means and CL_s represents the mean of subgroup standard deviations as well as control limits using the standard SPC formulas, with the constants A3=0.619, B3=0.555, and B4=1.445. This process allows for the Voice of the Sample Process (VOSp) to be established. Treating these control charts as an indication of delivery assembly line, helps us to spot if shipments for the individual product types are running smoothly or not, allowing inconsistencies to be caught quickly. Each product type will be analysed, starting with MOU below.

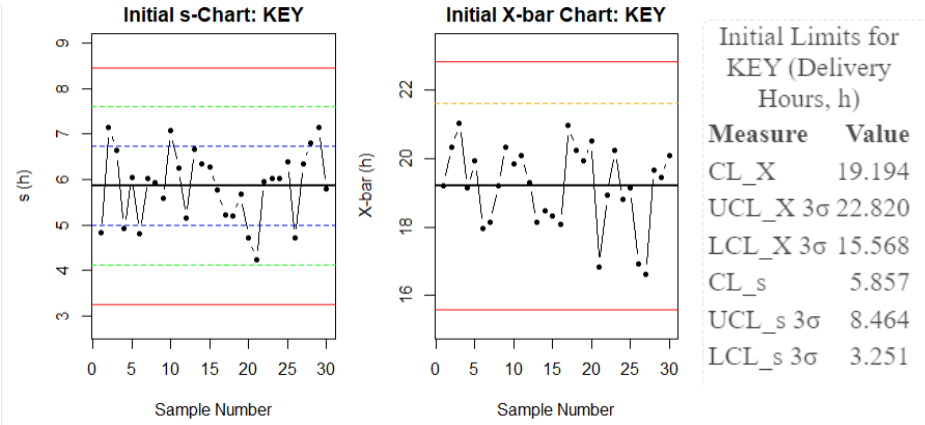


Initial Limits for MOU (Delivery Hours, h)

Measure	Value
CL_X	19.249
UCL_X 3σ	22.762
LCL_X 3σ	15.735
CL_s	5.676
UCL_s 3σ	8.202
LCL_s 3σ	3.150

Figure 25 Control charts for mouse (MOU)

The control charts and initial limits (baselines) for the first 720 MOU deliveries indicate an average delivery (CL_X = 19.2 hours), being over half a day and exceeding ideal thresholds, this has the potential to frustrate customers and signals systematic delays. The broad variability (CL_s = 5.7 hours) highlights inconsistent handling, this could be due to variable order volumes or routing issues. The s-chart (left) confirms a manageable spread of 3-7 hours per sample, staying close to the centre line with no breaches of the red 3-sigma limits at 8.2 and 3.2 hours, this indicated random fluctuations from everyday operations rather than breakdowns. The X-bar chart (right) shows means centred around 19.2 hours (16–22-hour range) all within the 3-sigma bounds at 15.7 and 22.8 hours. Overall, the process is stable but can be considered suboptimal. The process is in-control in that we are shipping reliably but inefficiently. Management should focus on reducing spread to minimise delays and avoiding type two risks of undetected late deliveries.

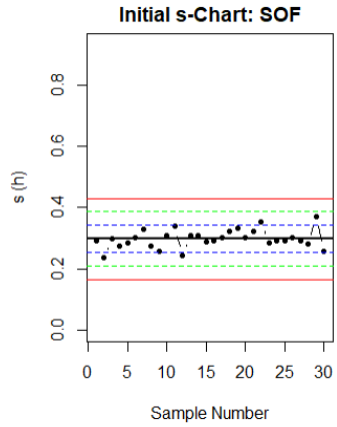


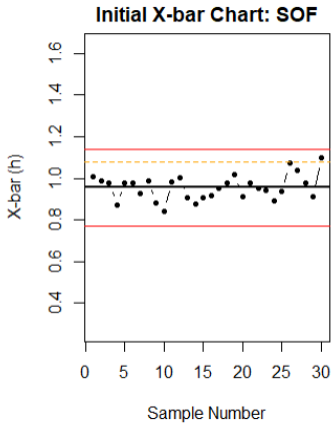
Initial Limits for KEY (Delivery Hours, h)

Measure	Value
CL_X	19.194
UCL_X 3σ	22.820
LCL_X 3σ	15.568
CL_s	5.857
UCL_s 3σ	8.464
LCL_s 3σ	3.251

Figure 26 Control charts for Keyboards (KEY)

Similarly, for keyboards (KEY), a centred average at CL_X = 19.2 hours indicates consistent timing, but the high variability of CL_s = 5.9 hours highlights potential handling inconsistencies, such as assembly delays that could push ordered over the target delivery line without that target being adjusted. Common-cause stability and random variation is further confirmed by the lack of patterns or runs beyond the 2-sigma limit. Looking at the s-Chart, sample spreads (3 to 8 hours) hover around the CL_s = 5.9 hours within the blue and green lines, these fluctuations are likely due routine factors, but the breadth indicates potential gains that can be made through better standardisation. The X-bar chart shows means of (16 to 22 hours), aligning tightly to the CL_X=19.2 hours. The points lie fully within the 3-sigma red lines (15.6 and 22.8 hours, this is not alarming however the slight upper clustering should be monitored.



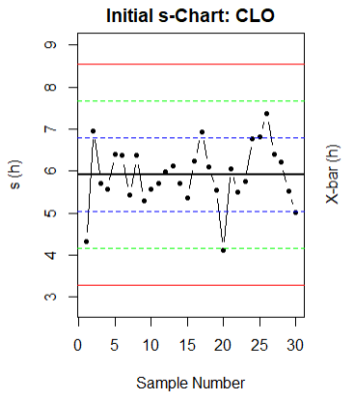


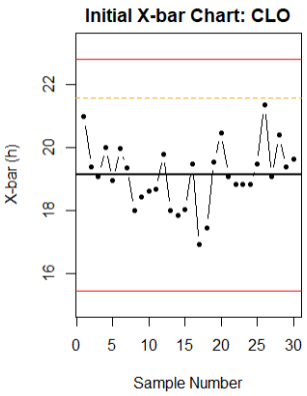
Initial Limits for SOF (Delivery Hours, h)

Measure	Value
CL_X	0.956
UCL_X 3σ	1.140
LCL_X 3σ	0.772
CL_s	0.297
UCL_s 3σ	0.430
LCL_s 3σ	0.165

Figure 27 Control charts for Software (SOF)

This charts for SOF show a lean operation, an average of CL_X=1 hour demonstrates efficiency but the modest variability with CL_s = 0.3 hours flags potential bottlenecks, likely in the upload or confirmation processes, risking minor over runs in high-volume spikes. However, this low CL_s value indicates tight control overall, with a low 3-sigma limit, minimising type one errors and non-conformance. Since there are no runs or outliers beyond 2-sigma, we can confirm a common cause reliability, making SOF a good benchmark for efficiency. The S-Chart clusters near CL_s=0.3 and spreads between 0.1 to 0.4 hours, fully inside the red 3-sigma limits at 0.43 and 0.17 hours, implying robust automation. The X-bar chart with means (0.8-1.1 hours) centred at 0.96 hours, enclosed by the red 3-sigma limits (0.77-1.14) does not have any alerts, peak loads should be monitored. SOF's in-control set up signals a high-performing process, needing some fine tuning in variability.



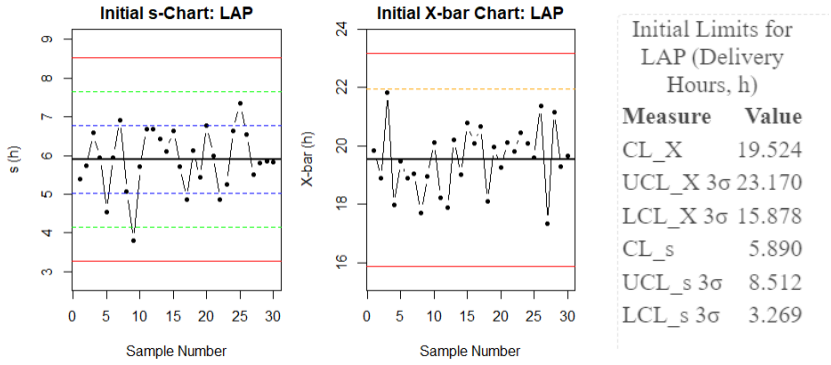


Initial Limits for CLO (Delivery Hours, h)

Measure	Value
CL_X	19.126
UCL_X 3σ	22.783
LCL_X 3σ	15.469
CL_s	5.908
UCL_s 3σ	8.537
LCL_s 3σ	3.279

Figure 28 Control charts for Cloud Subscription (CLO)

The charts for CLO have a centred mean at CL_X=19.1 hours reflecting a balanced throughput. A high variability is evident (CL_s= 5.9 hours), revealing potential inconsistencies. The s-chart indicates spreads oscillating around CL_s = 5.9 hours inside the blue and green zones, the red 3-sigma line is not crossed, showing routine flux. The breadth or spread indicates a need for improvements to be made to achieve better consistency. The X-bar chart has means anchored at CL_X=19.1 hours, bounded by red 3-sigma limits, not raising any flags, as the data is well centred. CLO's in-control profile signals reliability with some variability, that can be enhanced by minimising spread.
Another steady mean is observed at CL_X=19.5 with a variability of CL_s=5.9 hours. No runs pass 1- to 2-sigma limits, classifying the process as in control for this initial sample. The s-chart indicates everyday variance as the data is within the red 3-sigma lines. The width however reveals room for process hardening and better standardisation. The X-bar chart remains within the 3-sigma lines with no breaches, and the data points are well centred with a minor upper bias. This sample indicates an in-control process, signalling resilience to variability.

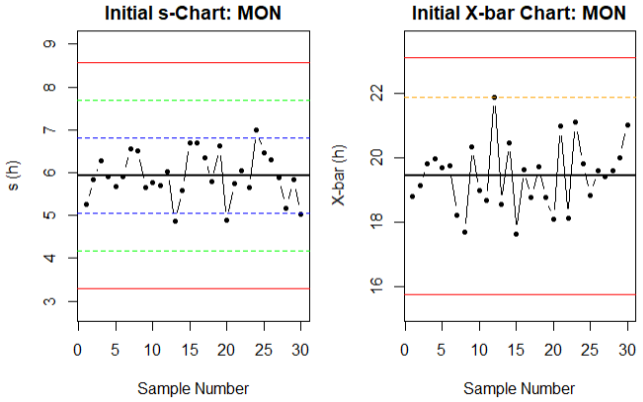


Initial Limits for LAP (Delivery Hours, h)

Measure	Value
CL_X	19.524
UCL_X 3σ	23.170
LCL_X 3σ	15.878
CL_s	5.890
UCL_s 3σ	8.512
LCL_s 3σ	3.269

Figure 29 Control charts for Laptops (LAP)

The charts for LAP have a reliable mean of CL_X=19.5 with evident variability with CL_s = 5.9 hours. The s_Chart spreads over 3 to 8 hours, surrounding the CL_s = 5.9 line and not passing the 3-sigma lines, indicating a fairly controlled process. The X-bar Chart has a strong alignment without exceeding the 3-sigma limits. While the process looks relatively controlled the variability should be investigated to fully optimise the process.



Initial Limits for MON (Delivery Hours, h)

Measure	Value
CL_X	19.426
UCL_X 3σ	23.092
LCL_X 3σ	15.760
CL_s	5.923
UCL_s 3σ	8.559
LCL_s 3σ	3.287

Figure 30 Control charts for Monitors (MON)

With a mean of CL_X = 19.4 and a variability of CL_s = 5.9 hours, MON's delivery process is very similar to that of the other product types. The process is in control with no breaches of the 3-sigma limits for both the s-chart and the X-bar chart. This in-control process signals resilience to variability.

3.2 Monitoring ongoing samples

In section 3.1, only the initial 30 samples were analysed to determine centre lines, out of control limits and the 1- and 2 -sigma control limits. Now the full data set will be monitored to assess the accuracy of the initial samples and to see the behaviour of the entire dataset filtered by each product type. The results are printed below.

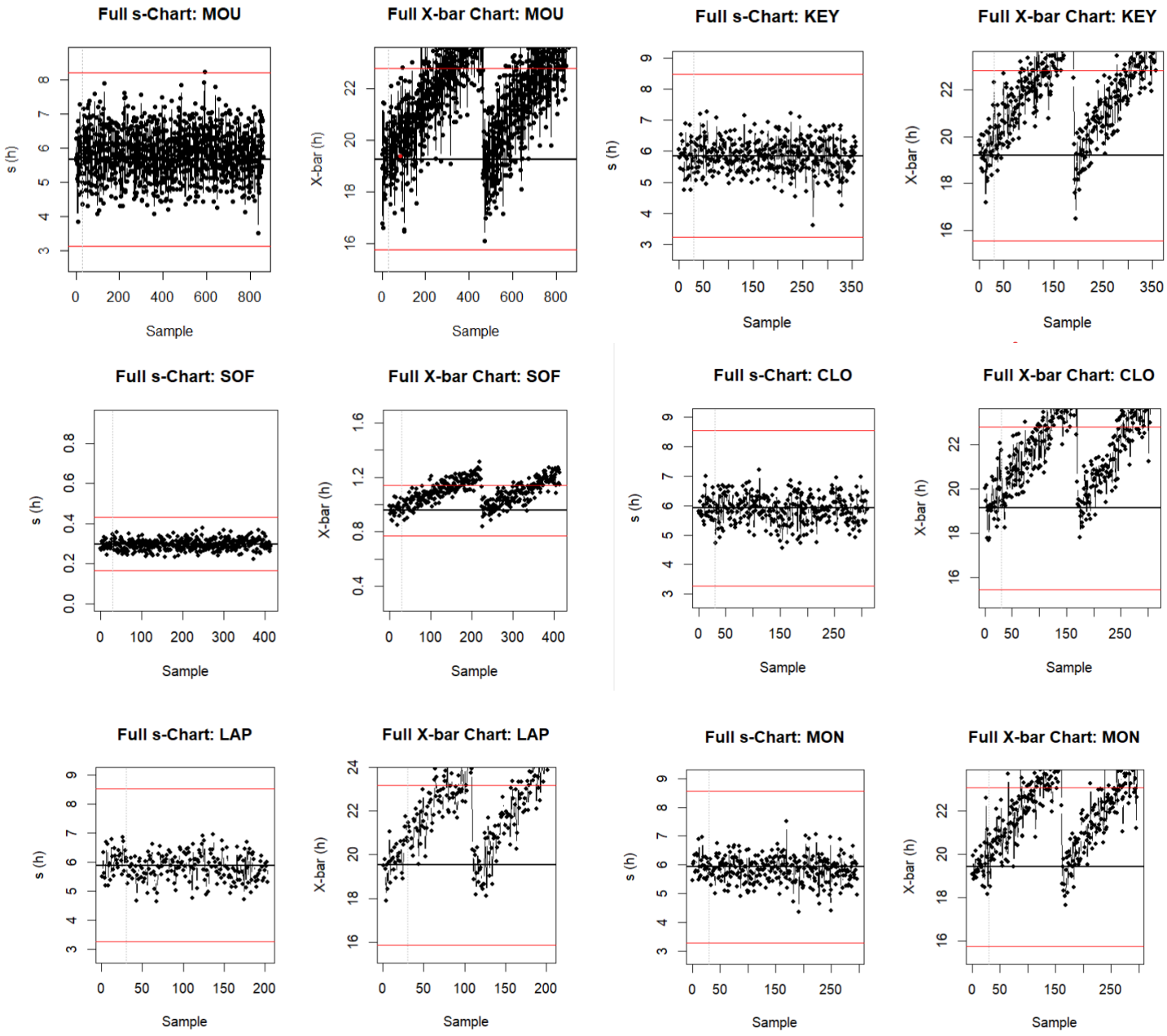


Figure 31 Control charts on the full dataset for each product type

Overall, there is a large discrepancy between the initial s- and x-bar charts created with the first sample of each product type in comparison to the charts created using the full dataset. While some product types remained relatively consistent with their s-charts, the X-bar charts for all product types display a large deviation from the controlled space. It is important to note that if the s-charts are not carefully checked for abnormally high spreads or distributions, the X-bar-charts cannot be properly evaluated. These results indicate a clear trend where performance declines over the course of the year and then returns to its initial level at the start of the new year, and repeats, suggesting reduced management and monitoring as the year progresses. The next section examines the accuracy of the initial control charts with the control charts on the full dataset in greater detail, exploring this observation further.

3.3 Process Capability Indices

For the company to be successful it is crucial that the voice of the process (VOP), such as our delivery time variability shaped by logistics and supply chains, is aligned with the voice of the customer (VOC) through clearly defined specification limits. The lower specification limit (LSL) represents the minimum acceptable threshold for individual items, in this case, it is 0 hours for delivery times, ensuring no negative values will be found. The upper specification limit (USL) defines the maximum tolerance, here it is 32 hours, to prevent excessive delays that could frustrate customers. Below, the first 1000 deliveries per product type are examined revealing clear differences between the product types.

Process Capability (First 1000 Deliveries; LSL=0h, USL=32h)

ProductType	Cp	Cpl	Cpu	Cpk	Capable
MOU	0.915	1.104	0.727	0.727	No
KEY	0.917	1.105	0.729	0.729	No
SOF	18.135	1.083	35.188	1.083	Yes
CLO	0.898	1.079	0.717	0.717	No
LAP	0.899	1.101	0.696	0.696	No
MON	0.889	1.079	0.700	0.700	No

Figure 32 Process capability (first 1000 deliveries)

The products Mouse, keyboard, cloud, laptop and monitor are not capable of consistently meeting the VOC, this is mainly due to the high variability in delivery times and a tendency for the longer deliveries to exceed the upper limit. Contrasting this, Software is the only product classified as capable (Cpk = 1.083), with very low variability and well centred delivery times, reflecting stable, well controlled processes. As a digital product, its delivery is handled electronically, such as through downloads, this reduces variability and ensures fast and consistent fulfilment. However, this specification for software is not an accurate representation of the capability of the delivery process as it is compared against the delivery of hardware components which allow up to 32 days to deliver. Software should rather be specified as an instantaneous delivery process with a much lower upper limit. This way, delivery times longer than what is expected for software components can be flagged and the process can be improved. Overall, the results show that while all products manage to avoid under-delivering, (Cpl values are all greater than 1), most of the hardware delivery lines require

improvements. These improvements may involve better routing or standardised packaging in order to reduce variability and increase Cpk, keeping deliveries reliably within customer expectations.

3.4 Identifying process control issues

Previously, we used X-bar- and s-Charts with n = 24 samples, and with limits based on the first 30 samples under the Central Limit Theorem. This was helpful to gain an idea of the mean and standard deviation of the process and to decide whether the process was in control. However, we cannot trust that this is a true reflection of the entire dataset or just a coincidental instance where the sample had consecutive points clustered neatly together. This is proven visually in the full X-bar- and s-plots in section 3.2, where we see the process deviate significantly away from the initial 3-sigma lines calculated in section 3.1.

This section applies sensitising rules to identify when normal, random variation (or common causes) turns into unusual variation (or special causes) that may need further investigation. By applying these rules, false alarms (type 1 errors) are reduced and missed detections (type 2 errors), that could result in late deliveries, are avoided.

Rule A: s > UCL_3σ (Post-30; First/Last 3)

ProductType	First3_A	Last3_A	Total_A
MOU	592	592	1
KEY	None	None	0
SOF	None	None	0
CLO	None	None	0
LAP	None	None	0
MON	None	None	0

Figure 33 Product types violating rule A

Rule A is triggered for any sample that occurs outside of the upper +3-sigma control limits for all product types. The table above summarises these instances. We see that only one s-chart sample exceeded the upper 3-sigma limit, and that was for the MOUSE line at sample number 592. This suggests a one-time disruption, perhaps due to a supplier issue or a demand spike. The rest of the products showed a stable variability within the control limits. It is recommended to perform a small root cause analysis on the affected mouse batch, but to be wary of overcorrecting the process as the process is stable overall.

Rule B: Longest Consecutive s in ±1σ (All Samples)

ProductType	Max_Consec_1sigma
MOU	16
KEY	15
SOF	21
CLO	35
LAP	19
MON	34

Figure 34 Product types violating rule B

Rule B is triggered when the largest number of consecutive samples between -1 and +1 sigma control limits is detected, for each product type, to identify good process control. The Max_Consec_1sigma values indicate how stable the process variability is relative to the 1-sigma limits. Longer consecutive runs within ±1-sigma suggest periods of unusually consistent subgroup variation. For example, CLO (35) and MON (34) show very long consecutive periods within ±1-sigma, which could indicate a process that is tightly controlled or possibly too uniform, whereas MOU (16) and KEY (15) show shorter runs, suggesting more variability. Overall, these runs help identify which products have the most or least stable dispersion over time.

Rule C: Starts of 4+ Consecutive X-bar > U2σ (Post-30; First/Last 3)

	ProductType	First3_C	Last3_C	Total_C
MOU	MOU	194, 235, 280	777, 811, 844	23
KEY	KEY	112, 172, 187	698, 721, 726	25
SOF	SOF	202, 237, 244	774, 803, 842	25
CLO	CLO	122, 179, 192	567, 604, 628	20
LAP	LAP	119, 130, 154	361, 374, 393	12
MON	MON	134, 179, 190	580, 610, 615	23

Figure 35 Product types violating rule C

Rule C is triggered when 4 or more consecutive X-bar samples outside of the upper, second control limits occur, for each product type. This rule identifies trends of sustained high averages, signalling potential shifts in the process mean. The Total_C values show how frequently these events occur, while the first and last starts highlight when they happen. For instance, KEY and SOF have the highest total events (25), indicating frequent upward shifts in X-bar beyond the 2σ threshold, which could signal a persistent drift in the process mean. LAP, with only 12 events, shows the least upward shift, suggesting better control of the mean. This analysis helps pinpoint which products are more prone to systematic increases in average delivery times.

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

4.1 Probability of making a Type I (Manufacture's) Error

The concept of a Type I Error is briefly described below:

Ho = Process is in control, centred on the CL from the first 30 samples.

Ha = Process is out of control.

Type I error (α) occurs if we reject Ho even though the process is in control.

The probability of one sample above the centreline is 0.5, because for a normal distribution, half the samples are expected to be above CL.
For Rule A, B and C, alpha can be estimated by considering the specific rule's trigger:

Rule	Trigger	Type 1 alpha
A	1 sample s > UCL_3σ	α =0.0027 (for 3σ)
B	8 consecutive s in ±1σ	α =0.5^8 =0.0039
C	4 consecutive X-bar > U2σ	α =0.0225^4 = 0.00025 (based on P(X > U2σ) =0.0225))

4.2 Probability of making a Type II (Customer's) Error

The concept of a Type II Error is explained below:

Ha is now true (because process mean shifted), but the chart fails to detect it.

Given:

- CL = 25.05 L, UCL = 25.089 L, LCL = 25.011 L
- Process shifts to μ = 25.028, σ = 0.017 (was 0.013)

The Type 2 Error probability (β) is the probability that the sample mean falls between LCL and UCL even though the mean has shifted:

$$\beta = P(LCL \leq \bar{X} \leq UCL \mid \mu = 25.028, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Step 1: Compute sample standard deviation

$$\sigma_{\bar{X}} = \frac{0.017}{\sqrt{n}}$$

Step 2: Standardize LCL and UCL (Z-scores)

$$Z_{LCL} = \frac{25.011 - 25.028}{\sigma_{\bar{X}}}, Z_{UCL} = \frac{25.089 - 25.028}{\sigma_{\bar{X}}}$$

Step 3: Compute β

$$\beta = P(Z_{LCL} \leq Z \leq Z_{UCL}) = pnorm(Z_{UCL}) - pnorm(Z_{LCL})$$

This calculation returns a beta value of 0.9999995

Therefore, there is approximately a 1% probability of failing to detect the shift. This outcome is sensible, as the control limits are very wide relative to the shift in the mean, which was very small. The subgroup standard deviation is still relatively large, so it is likely that most samples will still fall within the limits. Therefore, the type 2 error is likely to be very high, as it would be very easy not to detect the shift.

Part 5: Optimising for Maximum Profit

Coffee shop 1

After analysing the relationship between the number of baristas and the times taken to serve customers, the optimal number of baristas that each coffee shop requires on a given day can be determined, with the goal of maximising profit. To achieve this, we start by defining our objective function.

Max(Profit) = (Customers served x 30) - (Number of baristas x 1000)

Where R30 is the material profit made per customer including the fixed costs such as milks, syrups etc., and R1000 is the cost per day of appointing one barista.

This objective function is subject to the following constraints:

- A minimum number of 2 baristas
- A minimum service time threshold of 60 seconds

It is logical to assume that the service time must be less than or equal to some defined threshold in order to be considered reliable. This information has not been provided, therefore, a service time threshold has been set to 60 seconds, thus, a service time below this is considered reliable.

Assumptions

- Working days per year 365
- Profit per customer (after materials) R30
- Barista cost per day R1 000
- Reliable service time ≤ 60 seconds
- Average day length of 10 working hours (36,000 seconds/day)
- Each line in the data represents one customer served that year

The maximum profit can be calculated by following two methods:

1. Brute force method which calculates the profit for each feasible barista number as these are discrete values.
2. Analytically solve using the optimise() function in R to find the maximum.

To begin, we start by deriving metrics that will be useful to analyse the problem, such as our service time, reliability percentile, customers per year and per day as well as our profit for day. The results are summarised in the table below. Only the averages of these metrics per count of baristas were considered. The following visualisations provide more insight.

V. <int>	n_... <int>	avg_ti... <dbl>	reliable_pct <dbl>	customers_day <dbl>	profit_day <dbl>	customers_year <dbl>	coffees_per_barista <dbl>	extra_profit_vs_prev <dbl>
1	417	200.2	0.0	180	4396	65648.83	180.0	NA
2	35...	100.2	0.0	719	19563	262351.44	359.5	15167
3	12...	66.6	16.5	1621	45640	591787.54	540.3	26077
4	29...	50.0	97.2	2881	82434	1051612.68	720.2	36794
5	56...	40.0	100.0	4504	130129	1644068.65	900.8	47695
6	97...	33.4	100.0	6476	188270	2363617.82	1079.3	58141

Figure 36 Metric averages per barista count for coffee shop 1

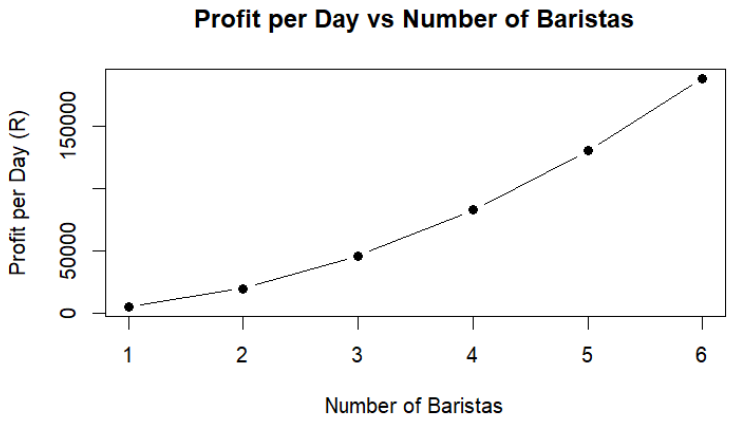


Figure 37 Profit per day vs number of baristas for coffee shop 1

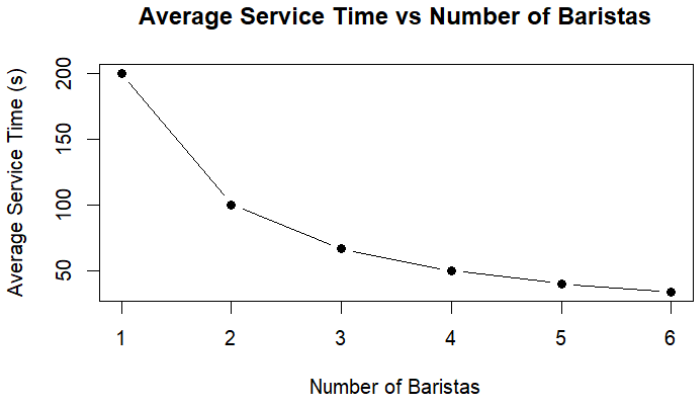


Figure 38 Average service time vs number of baristas for coffee shop 1

Using the observed service times (V2) for the observed barista counts (V1), we see that the average time to serve depends on the number of baristas. The relationship can be fitted with a simple inverse model, as baristas increase, service time decreases. However, one can see that this approach falls short. By only considering the observed baristas, the service time will always decrease, and the profit will continue to increase before the effect of rising staff costs is seen. The inverse model suggests that adding more baristas continuously decreases service time and increases profit, this is unrealistic beyond the observed staffing range.

To correct this, we introduced a constraint based on service reliability, the optimal staffing must achieve at least 95% of customers served within 60 seconds. With these adjustments, the optimisation identifies the true trade-off between staffing cost, throughput, and service quality. With this constraint, we see a different relationship in our graph, as our service reliability seems to plateau around 4 baristas (figure 39 below), suggesting that this may be the optimal number of baristas to hire.

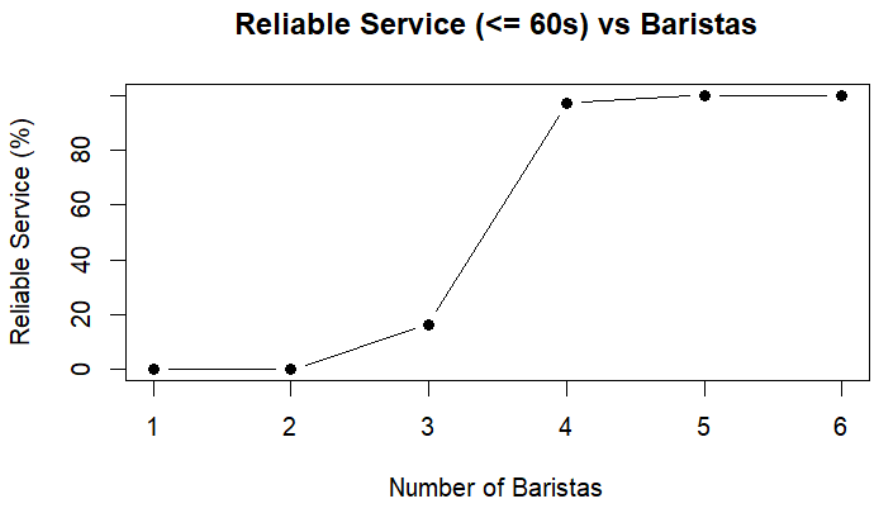


Figure 39 Reliable service vs number of baristas

Using the brute force method in R studio, a similar relationship occurs. Due to the straight-line distribution between the profit per day vs number of baristas plot, the most optimal number of baristas to achieve the maximum profit will clearly be the maximum number of baristas, 6. It therefore makes sense that the function outputs an optimal number of baristas of 6, with a maximum profit of R188 270 for the observed range. The *optimise()* function was also used to check this result. The function had a similar output of 5.77 baristas. After rounding up (for a discrete variable), we see that both approaches yield 6 baristas to be the optimal solution. It is important to keep in mind that the additional service reliability constraint was not accounted for in these functions, explaining why these functions did not output a value of 4. It is therefore important to observe the data critically and make reasonable assumptions to create constraints where necessary, instead of blindly trusting built in functions.

Coffee Shop 2

Similarly, Coffee Shop 2 has been analysed to determine the optimal number of baristas needed to achieve a maximum profit. The summary table of the derived metrics for the new coffee shop is shown below.

V1 <int>	n_obs <int>	avg_time <dbl>	reliable_pct <dbl>	customer... <dbl>	profit_day <dbl>	customers_year <dbl>	coffees_per_barista <dbl>	extra_profit_vs_prev <dbl>
1	2196	200.2	0	180	4395	65644.55	180.0	NA
2	8859	141.5	0	509	13263	185705.20	254.5	8868
3	19768	115.4	0	936	25066	341473.39	312.0	11803
4	35289	100.0	0	1440	39193	525519.73	360.0	14127
5	54958	89.4	0	2013	55378	734603.77	402.6	16185
6	78930	81.6	0	2646	73370	965670.90	441.0	17992

Figure 40 Metric averages per barista count for coffee shop 2

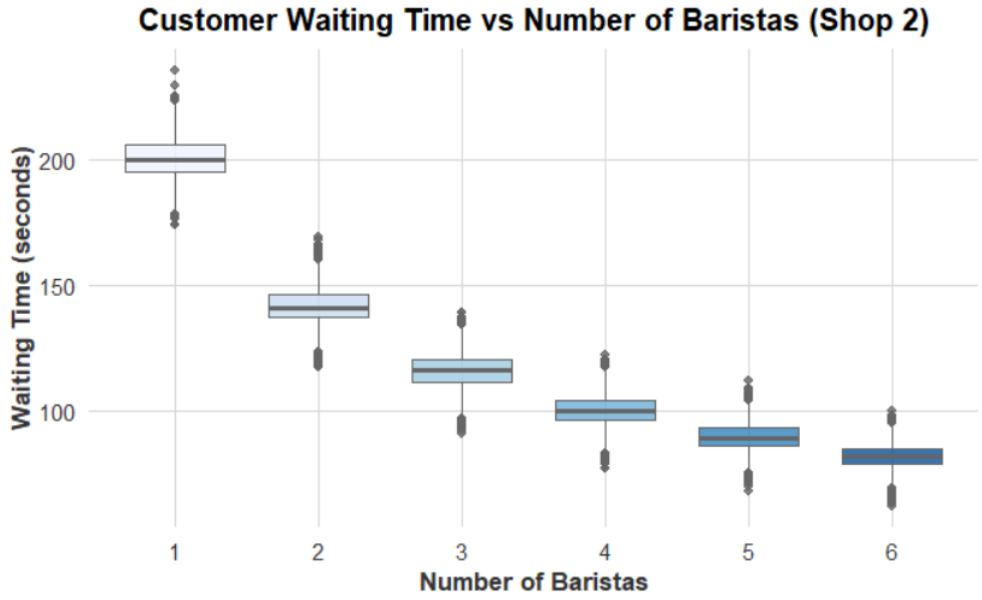


Figure 41 Box plot of customer waiting time vs number of baristas for shop 2

From this box plot, one can clearly see the waiting time decreases as the number of baristas increase, however this effect seems to reduce past 5 to 6 baristas as the line starts to flatten out or plateau. This could indicate that the added benefit or time saving achieved by having a 6th barista does not outweigh the cost associated with extra personnel. This is reinforced by the figure below.

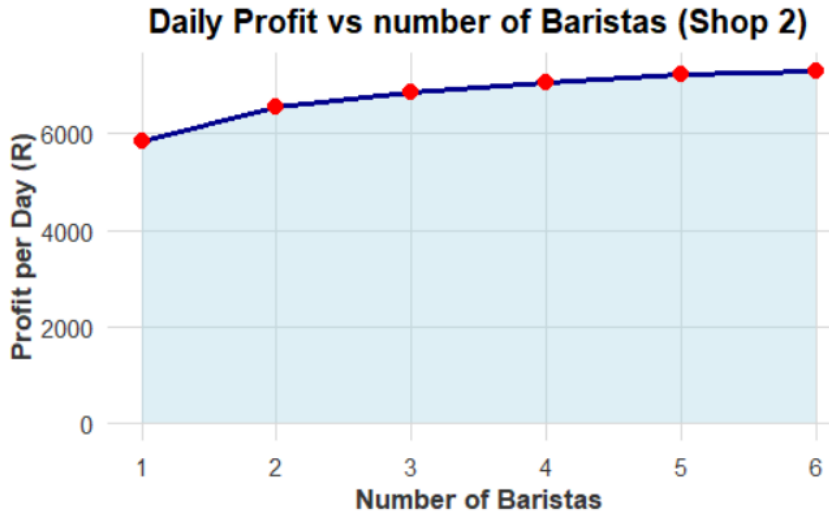


Figure 42 Line graph of Daily profit vs number of baristas for shop 2

In the plot of daily profit vs number of baristas, the plateau is clearly seen, as the trend seems to flatten out. After 5 baristas are hired, the increase in profit is insignificant as more baristas are hired. From this inspection one can deduce that the optimal number of baristas for coffee shop 2 is 5, as clearly indicated in figure 43 below.

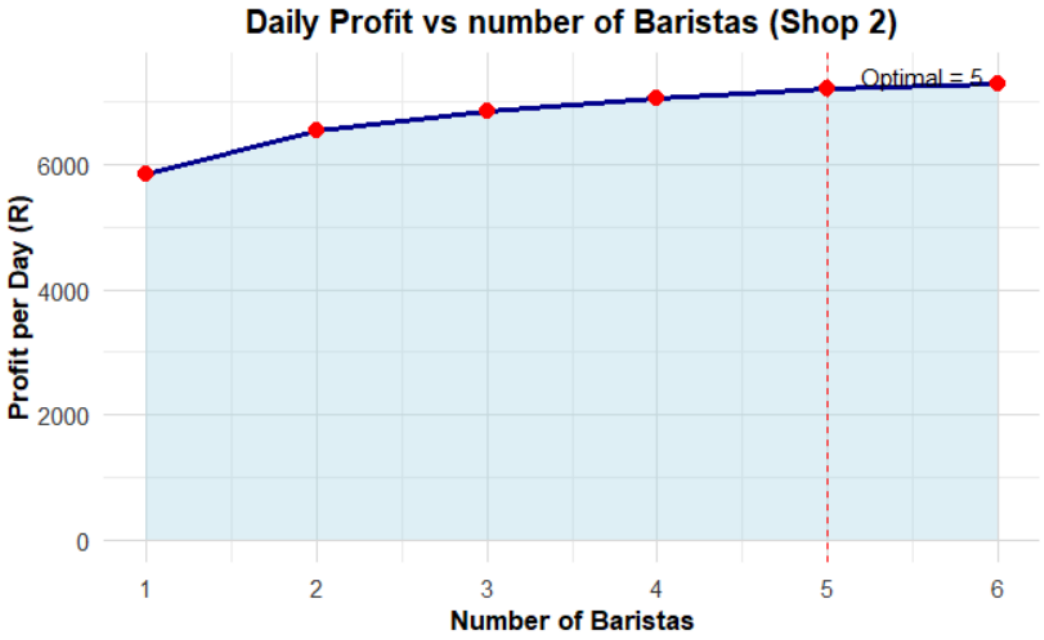


Figure 43 Line graph showing optimal number of baristas for maximum profit

This result is confirmed with the use of the optimise() function, indicating the optimal number of baristas to be 5, with a corresponding maximum daily profit of R7208.38.

Coffee shop 2's data is a better representation real-world dynamics compared to coffee shop 1. Unlike the coffee shop 1, which had a linear relationship between the number of baristas and service time, this dataset introduces diminishing returns, meaning each additional barista contributes less to reducing the average waiting time.

It is therefore recommended that coffee shop 2 hires 5 baristas to achieve a maximum daily profit. It is also important that management continues to monitor the daily profit and average service time graphs to ensure that they are always performing at optimum levels.

Part 6: DOE and ANOVA

Following the results of part 3, Statistical Process Control, our understanding of the data can be further enriched using the ANOVA process to test various hypotheses. In this section, we apply the principles of Design of Experiments (DOE) through a single factor ANOVA to investigate the structured variations of the outputs from our data in part 3, building on our statistical process control findings.

Using DOE, we are able to evaluate the impact of various input factors on response variables such as delivery times. Here, a single factor design is used where the factor is "year", and the two levels represent 2022 and 2023, and the response variable is DeliveryHours.

The SPC analysis revealed a stable but suboptimal initial control charts for delivery times across all product types. (CL_X was approximately 19 hours for MOU, KEY, CLO, LAP and MON, and much lower for SOF at approximately 1 hour). However, after comparing the initial control charts to the ones created on the entire dataset, large deviations in the X-bar charts were revealed, with performance degrading over the year (breaching 3-sigma limits) and only resetting at the beginning of the new year, suggesting special causes like reduced management oversight. Moreover, process capability indices further highlighted the incapability for hardware products due to high variability and exceeding the USL of 32 hours. These insights motivate testing temporal effects, such as whether delivery times differ significantly between the years 2022 and 2023, potentially indicating a decline in efficiency through the course of the year. The following hypotheses are created:

H_0 = There is no significant difference in mean DeliveryHours between 2022 and 2023 ($\mu_{2022} = \mu_{2023}$).

H_a: There is a significant difference in mean DeliveryHours between 2022 and 2023 ($\mu_{2022} \neq \mu_{2023}$).

Alpha is set to be 0.05 for Type 1 error control. The data is subset to all sales records (n approximately 1000) with DeliveryHours as the continuous response. There are no missing values (as confirmed in part one of this report) and normality is assumed via CLT from large n (more than 30 per group, from part 3). The analysis is performed in Rstudio.

ANOVA Results

Because n is unbalanced, ($n_{2022} = 53727$; $n_{2023} = 46273$), the function *aov()* in R is used to accommodate this naturally and to ensure a more robust and statistically efficient response.

	Df	Sum Square	Mean Sq F	Value	Pr(>F)
OrderYear	1	139	138.62	1.391	0.238
Residuals	99998	9967559	99.68		

F =1.39, p = 0.238 > 0.05 therefore we fail to reject H_0 (no significant effect).

Single Factor ANOVA with n balanced:

Source <chr>	Treatment <dbl>	Error <chr>	Total <chr>
SS	72.0500000	9226177.33	9226249.38
DoF	1.0000000	92544	92545
MS	72.0500000	99.7	-
F	0.7200000	-	-
p-value	0.3952487	-	-

Figure 44 ANOVA Table

Means: 2022 = 17.50 hours and 2023 = 17.43 hours (negligible difference).

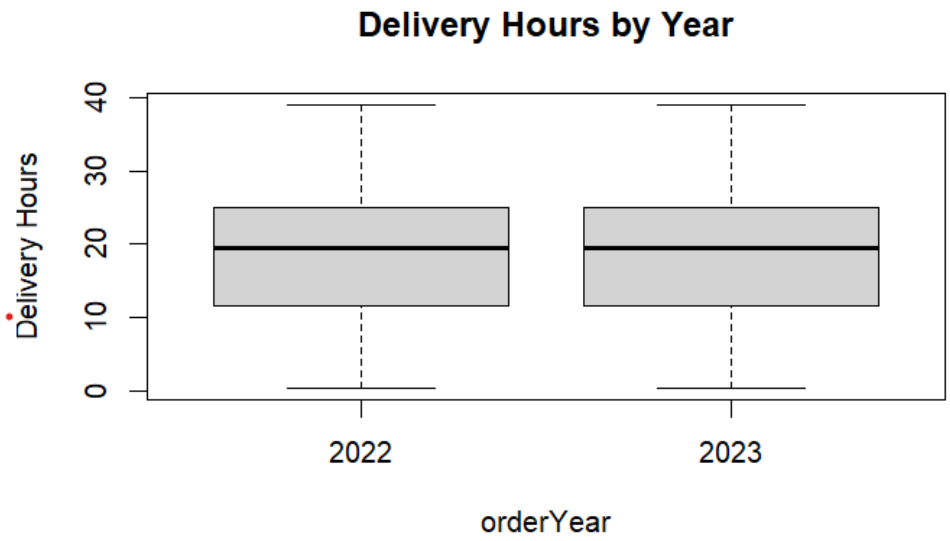


Figure 45 Box plot of delivery hours of 2022 vs. 2023

The results confirm a stable delivery process across the two years in the sales data (totalling about 100000 orders with 53727 from 2022 and 46273 from 2023). Both ANOVA methods tested if average delivery times differed between years, and the results show no meaningful change as delivery stayed consistent, averaging around 17.5 hours per order.

The Standard ANOVA generated with the unbalanced data revealed an F-statistic of 1.39 and p-value of 0.238 meaning that we can't rule out that the slight decrease from 17.5 hours in 2022 to about 17.4 hours in 2023 is just random noise and not a trend. Variability within the years (likely driven by outliers up to 40 hours) explains most differences, not the year itself.

The ANOVA generated by the balanced data set that was reduced to 46273 orders per year, revealed an F-statistic of 0.72 and a p-value of 0.395, with almost identical means (17.48 compared to 17.43 hours), reinforcing no year-based shift.

The boxplots confirm that both years show similar spreads. Deliveries cluster between 10 to 22 hours (median approximately 17), with extreme points reaching to 0 and over 40 hours due to rare delays. There is no obvious skew or separation between the boxes, highlighting steady performance but room

for better efficiency if outliers can be reduced. Overall, this supports our SPC findings, as the process is in control over the two years, but focussing on reducing variability will optimise and improve the process significantly.

Part 7: Reliability of Service

7.1 Estimation of the proportion of reliable service days per year

Upon review of the car rental agency, the following information was obtained. We see that reliable service days occur with 15 or more workers present. Days are described as "reliable" if no problems occur and operations run smoothly. From the provided graph we see that that there is one day where 12 workers are present, 5 days with 13 workers, 25 days with 14 workers, 96 days with 15 workers and 270 days with 16 workers.

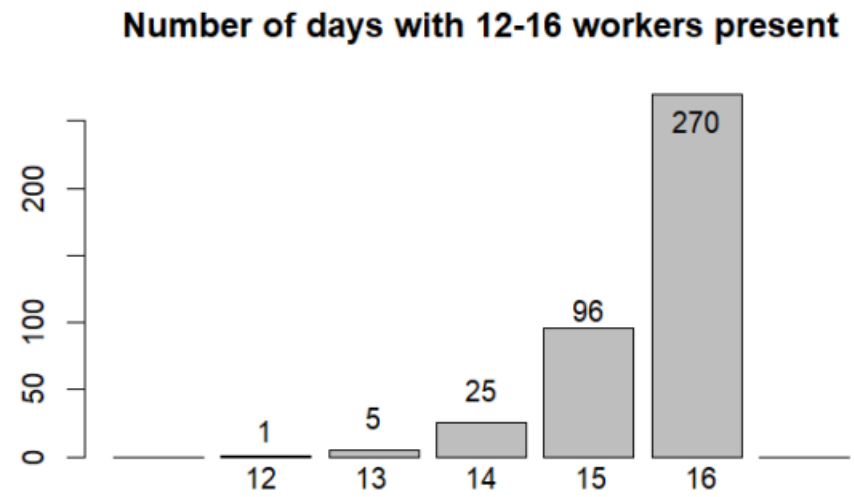


Figure 46 Histogram of number of days vs. number of workers present

Thus, the days where issues occurred due to being understaffed (less than 15 workers) would sum to 31 days (1 + 5 + 25 = 31), and the number of days with reliable service is 397 – 31 = 366 days. From this we can deduce the proportion of reliable days to be 366/ 397 = 0.9219. Assuming a standard year has 365 days, we can expect the number of reliable service days per year to be (366/397) x365 = 336.5 days or 337 days. We can treat this 337-day period as a representative sample, assuming the relationship between service reliability and number of workers present is proportional.

7.2 Optimising profit for the company

By modelling the data provided as a binomial problem the number of personnel assigned can be optimised. To do this, the following is accounted for:

- The company experiences problems is there are less than 15 people on duty.
- Each problematic day yields on average R20 000 less in sales for the day.
- More personnel can be appointed at a cost R25 000 per month per person.

To optimise profit, the daily number of workers present, X, is modelled to follow a binomial distribution $X \sim \text{Binomial}(M, p)$, where:

- M is the total personnel assigned (currently M = 26, based on the observed range of 12 to 16 workers).
- p is the probability that an individual worker is present, each worker is independent, and this value is estimated from the data.

Step 1: Estimating the probability p

Total worker-days observed: $12 \times 1 + 13 \times 5 + 14 \times 25 + 15 \times 96 + 16 \times 270 = 6187$. Total possible worker-days (assuming $M = 16$): $16 \times 397 = 6352$. Thus, $\hat{p} = \frac{6187}{6352} = 0.974$.

This fits the data reasonably well as the model's $P(X < 15) = 0.0636$ (expected approximately 25 problem days over 397 days), compared to the observed 31 days (approximately 7.8%).

Step 2: Profit Model

- Expected problem days per year: $365 \times P(X < 15 \mid M, p)$.
- Expected annual loss from problems: $[365 \times P(X < 15)] \times 20000$.
- Additional annual personnel cost: $(M - 16) \times 25000 \times 12$ for $M > 16$ (0 otherwise). Current costs for 16 workers are fixed and excluded from optimization.

If we want to maximise profit, we need to minimise the total extra annal cost, which comes from the expected loss and the additional personnel cost.

Step 3: Use brute force method to optimise
Since M is a small integer ($M \geq 16$), we can enumerate values for M and compute costs using the binomial cumulative distribution function. We can use the brute force method to do this as we are only using discrete values. The results are displayed below.

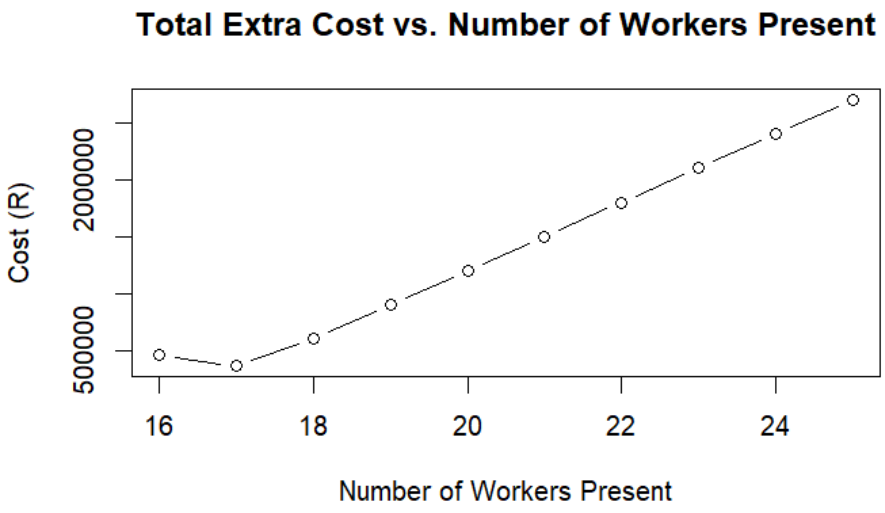


Figure 47 Total extra cost vs. number of workers present

This plot provides a clear visualisation of the relationship between the extra annual cost and the number of personnel assigned (M). The x-axis shows the values for M ranging from 16 (current staffing) to 24, and the y-axis scales form R0 to R2 000 000, capturing the expected loss from issues caused by understaffing and includes the effects of additional hiring costs, revealing the cost-trade off associated with optimisation problems.

We notice that the minimum cost point is at M=17, corresponding to an additional cost of R366386.30. This result makes logical sense as it reveals that hiring one additional person reduces the expected losses from R465259.80 at M = 16, by more than the added personal cost (R300000/year), yielding a net savings of R465259.80 - R366386.30 = R98873.50.

Beyond M = 17, we see a steep rise in cost, increasing linearly with the added hiring expenses. From this we see that the hiring expenses outpace the diminishing marginal reduction in understaffing risk (as $P(X < 15)$ approaches 0). On the left of the graph, a sharp decline is seen between M = 16 and M = 17, thus reflects the impact of one extra worker in a high probability (0.974) binomial setup, it highlights the "easier" fix of solving the problem of having fewer than 15 workers. Moreover, no local minima or irregularities are seen, indicating a convex function which is ideal for optimisation using simple methods such as the brute force method

Recommendations for the Car Rental Agency

The graph substantiates that that M = 17 represents the most optimal staffing level. If 17 workers are assigned, additional costs decrease by approximately 21%, and the net profit is predicted to be improved by R98000/ year. From a risk perspective, with 17 workers, the number of problematic days is expected to drop to 3 days, which is significantly lower than the 23 problematic days experienced with the current 16 workers, this will drastically improve the company's service reliability.

It is therefore recommended that the company hire one more full-time worker (annual cost of R300 000) to achieve this improved profit and savings on additional costs. With that, it is recommended that the p value is monitored quarterly. If the employee attendance starts to drop again, reassess the model and consider hiring an 18^{th} worker or implementing stricter employee attendance rules.

Conclusion

Throughout this report it has been confirmed that data corrections and statistical techniques can greatly improve the accuracy and reliability of insights. Statistical process control revealed stable yet variable delivery processes, while ANOVA verified consistency across years. Profit optimisation models balanced efficiency and cost by identifying the optimal staffing levels for both coffee shops and a car rental agency. Overall, the report emphasises the value of quality assurance in business data, highlighting the importance of accurate data management, continuous process monitoring, and informed decision making to maximise profit and service reliability.

References

Dirkse van Schalkwyk, T.G. 2025. *Quality Assurance 344 Lecture notes*. Stellenbosch University, Faculty of Engineering. Unpublished lecture notes.