2025

STELLENBOSCH UNIVERSITY
INDUSTRIAL ENGINEERING QA344
ECSA REPORT

27068641 - LJ Joubert

TABLE OF CONTENTS

TABLE OF GRAPHS

TABLE OF FIGURES

TABLE OF TABLES

1. INTRODUCTION

This report is completed as part of the Engineering Council of South Africa (ECSA) Graduate Attribute 4 (GA4) requirements for the Stellenbosch University Industrial Engineering program. The ability to apply quantitative engineering principles to solve complex, real-world problems is demostrated. The main goal of this report is to analyse operational datasets covering customer sales, product logistics, and service times, using statistical and optimisation techniques to improve data quality, control processes, and maximise profit.

The analysis is structured into seven key sections. It begins with a Basic Data Analysis using descriptive statistics to understand the structure of the data and identify any quality issues. This is followed by Statistical Process Control (SPC) to establish control limits and evaluate process capability for delivery times. The Risk and Data Correction section addresses inconsistencies in the data, calculates Type I and Type II errors, and revisits the descriptive statistics using the corrected data. Optimization techniques are then applied to determine optimal staffing levels for two coffee shops and a car rental company. Operational costs are balanced, reliability is considered, and revenue is maximised. Design of Experiments (DOE) and Analysis of Variance (ANOVA) are also used to test for significant differences in process performance.

This project was completed throughout the semester, serving as a practical study guide to reinforce statistical and optimisation concepts learned in class. It provided continuous hands-on experience, allowing the development of both analytical and problem-solving skills in a real-world context.

2. BASIC DATA ANALYSIS

2.1. DATA LOADING AND INSPECTION

To prepare for the visual inspection of data a general inspection is preformed on the provided data sets.

2.1.1. Customer

Table 1: *Customer Data*

CustomerID <chr>	Gender <chr>	Age <dbl>	Income <dbl>	City <chr>
CUST001	Male	16	65000	New York
CUST002	Female	31	20000	Houston
CUST003	Male	29	10000	Chicago

Table 1 depicts the customer data provided. It was confirmed that there are no missing values in any of the 5,000 rows of data. The CustomerID column can be linked to the sales datasets. The Gender column includes entries of “Male,” “Female,” and “Other.” The Age and Income columns both contain a range of numeric entries, while the City column lists the names of cities.

2.1.2. Products

Table 2: *Product Data*

ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
SOF001	Software	coral matt	511.53	25.05
SOF002	Cloud Subscription	cyan silk	505.26	10.43
SOF003	Laptop	burlywood marble	493.69	16.18

Table 2 depicts the product data. It was confirmed that there are no missing values in any of the 60 rows of data. The ProductID column can be linked to the sales datasets. The Category and Description columns contain various text entries, while the SellingPrice and Markup columns include a range of numeric entries.

2.1.3. Head Office

Table 3: *head Office Data*

ProductID <chr>	Category <chr>	Description <chr>	SellingPrice <dbl>	Markup <dbl>
SOF001	Software	coral matt	511.53	25.05
SOF002	Software	cyan silk	505.26	10.43
SOF003	Software	burlywood marble	493.69	16.18

Table 3 depicts the head office data. It was confirmed that there are no missing values in any of the 360 rows of data. The dataset is identical to the product data but noticeably larger. These two datasets will need to be compared.

2.1.4. Sales

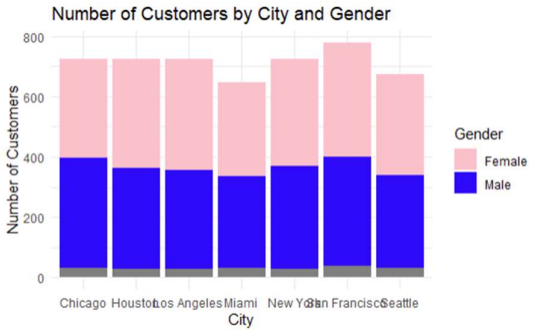
Table 4: *Sales 2022/2023 Data*

CustomerID <chr>	ProductID <chr>	Quantity <dbl>	orderTime <dbl>	orderDay <dbl>	orderMonth <dbl>	orderYear <dbl>	pickingHours <dbl>	deliveryHours <dbl>
CUST1791	CLO011	16	13	11	11	2022	17.72167	24.544
CUST3172	LAP026	17	17	14	7	2023	38.39083	31.546
CUST1022	KEY046	11	16	23	5	2022	14.72167	21.544

Table 4 depicts the sales data from 2022 to 2023. It was confirmed that there are no missing values in any of the 100,000 rows of data. The *CustomerID* and *ProductID* columns make this dataset the most connected to the other datasets. All columns, except for the two previously mentioned, contain numerical entries.

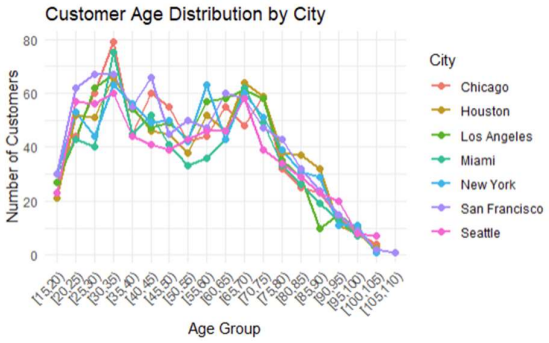
2.2. DATA VISUALIZATION

2.2.1. Customer Data



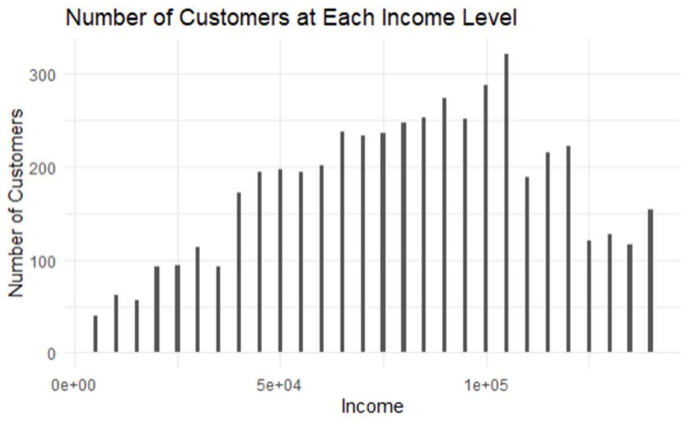
Graph 1: *Customer Gender*

Graph 1 displays the total number of customers and the gender distribution across each city. There is a relatively equal distribution between male and female customers across all cities. San Francisco is indicated to have the highest number of customers, while Miami has the lowest. However, it seems the number of customers never varies by more than 100 between cities.



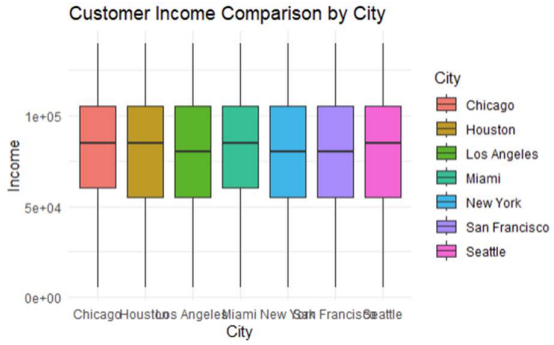
Graph 2: *Customer Age*

Graph 2 depicts the distribution of age among customers per city. Overall, the age distribution is very similar, with 2 significant peaks at ages 30 to 35 and 65 to 70 making it a bimodal distribution. The number of customers understandably tapers off after age 70. Miami has a notable dip from ages 50 to 65, with another notable dip at ages 85 to 90 for Los Angeles.

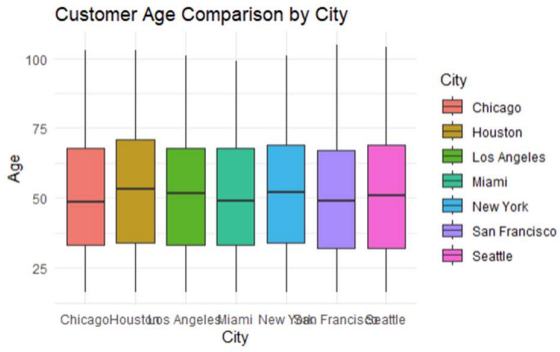


Graph 3: *Customer Income*

Graph 3 displays the total number of customers at each income level. There is a notable peak at an income level of 100,000. The majority of customers fall within an income range of 50,000 to 100,000. There is a steady increase up to 100,000, after which there is a significant drop in the number of customers.



Graph 5: *Customer Income per City*



Graph 4: *Customer Age per City*

Figure 5 and 4 are intended to show whether there is a significant distinction in customer age and income level between cities. These figures confirm a relatively equal distribution of age and income across the cities, indicating that the customer base is highly uniform and likely does not contain any significant patterns.

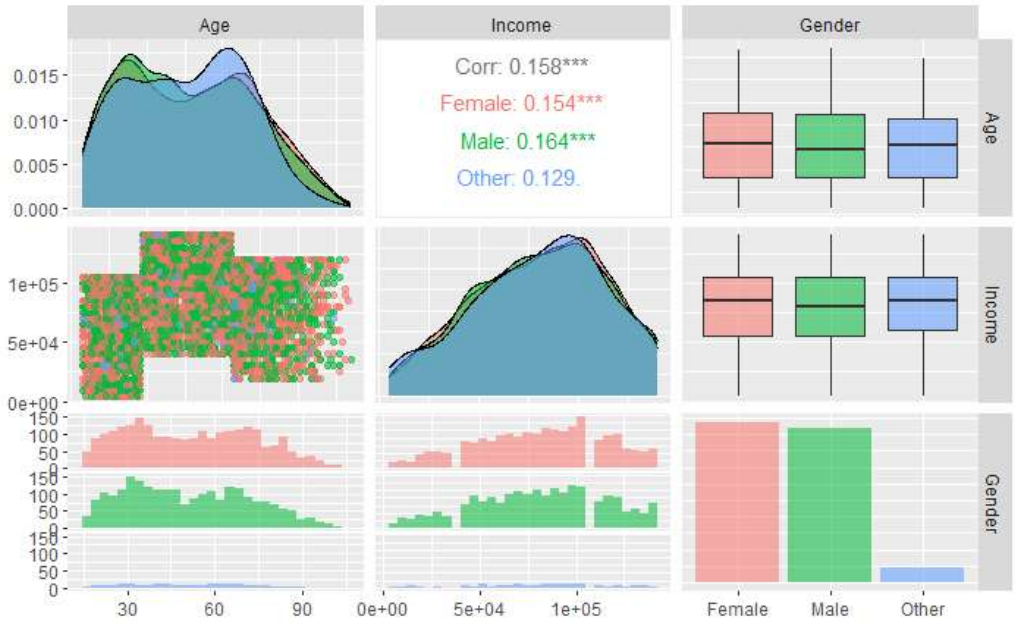
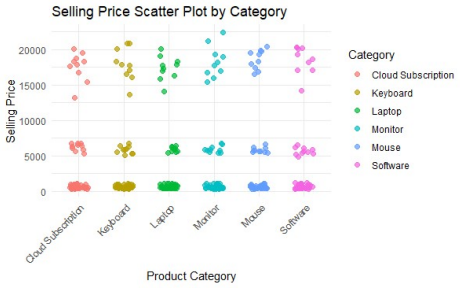


Figure 1: *Customer Correlation*

Figure 1 was created to identify any correlation within the customer dataset. A correlation of less than 0.1 is considered negligible. The figure compares a variety of graphs that compare the income level, age, and gender of each customer to identify potential correlations between these characteristics. Given the low correlation values, it is clear that the previous assumption was correct: no noteworthy pattern can be derived from the customer data alone.

2.2.2. Head Office and Product Data

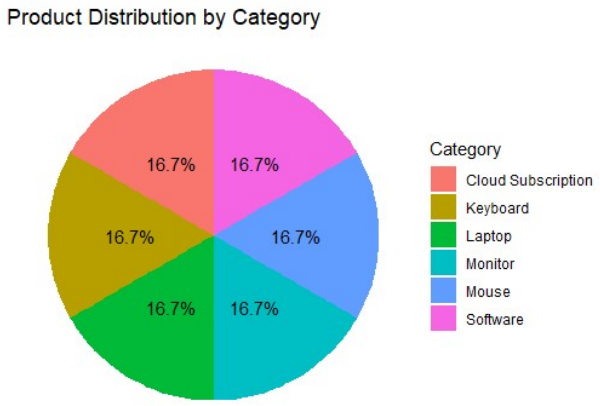
Given that the Head Office and Products data sets are structure similarly they were considered together.



Graph 6: *Selling Price Head Office*

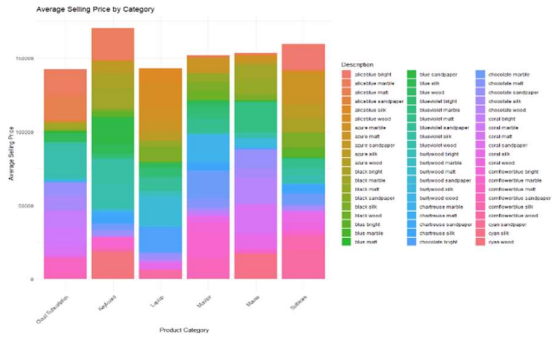
Graph 7: *Selling Price Products*

Graph 6 and 7 depict the selling price of products per category in a scatter plot. There are three price ranges consistent across all categories, with the majority of products in the lower price range (under 1,250) and the fewest products in the higher price range (over 15,000). There is no clear distinction in price range between categories in this graph.

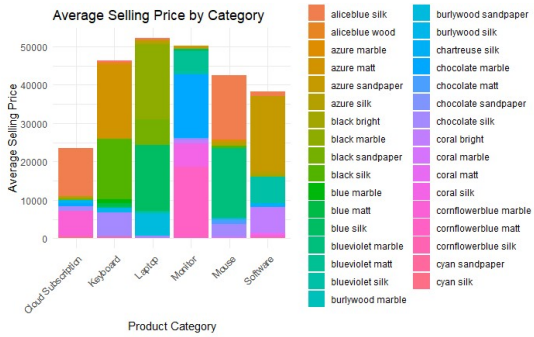


Graph 8: *Product Categories*

Graph 8 was used to confirm the product distribution between categories. It was found that, for both data sets, the distribution was exactly equal between the six categories.



Graph 9: *Average Selling Price Head Office*



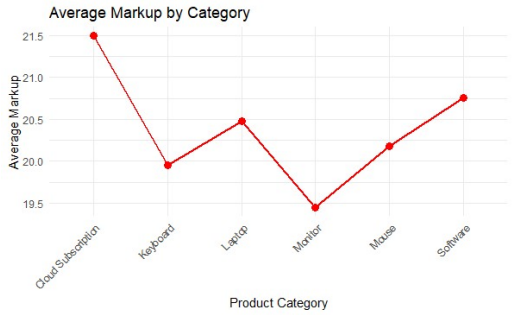
Graph 10: *Average Selling Price Products*

Graph 9 and 10 depict the average selling price per category, with a distinction between descriptions shown using a colour gradient. There is a clear inconsistency between the data from Products and Head Office. These values are expected to be equal however according to Head Office, the average selling price of keyboards is the highest, while according to Products, laptops have the highest average selling price. This discrepancy must be investigated.

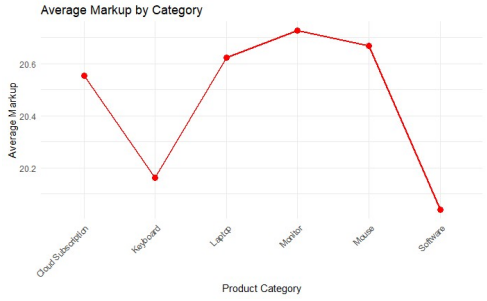
Graph 11: *Average Selling Price Head Office (line)*

Graph 12*: Average Selling Price Products (line)*

Graph 11 and 12 were generated to more clearly show the discrepancy between the average selling prices of each category in the two data sets. Graph 13 and 14 were also created to investigate whether this discrepancy extends to the average mark-up values. It was confirmed that there is a significant number of mismatched values between the data sets.



***Graph 13**: Average Selling Price Head Office (line)*

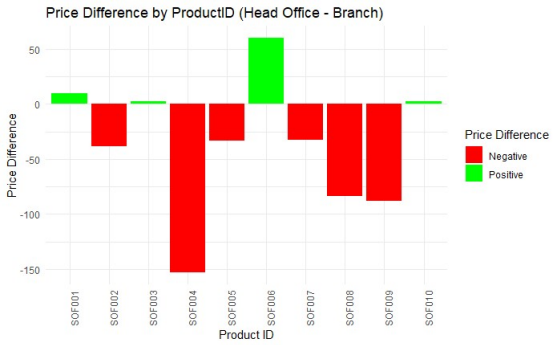


***Graph 14:** Average Mark-up Products (line)*

2.2.3. Investigating Mismatched Values

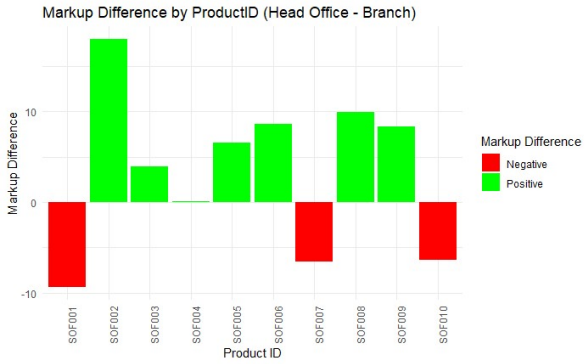
The ProductID's were matched between the two data sets. The remaining data was then compared between data sets.

To gauge the extent of the mismatched values, graph 15 and 16 were plotted to display the numeric difference between selling price and mark-up price between the two data sets for each ProductID.



***Graph 15:** Mismatched Selling Prices*

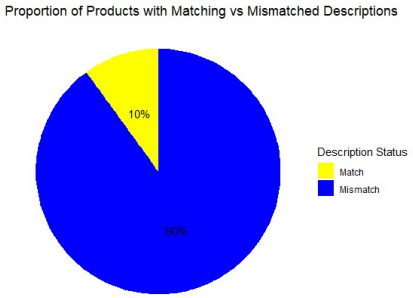
As seen on graph 15, the majority of the head office selling prices are higher than those listed by the products data set. There are quite significant differences displayed by products like SOF004.



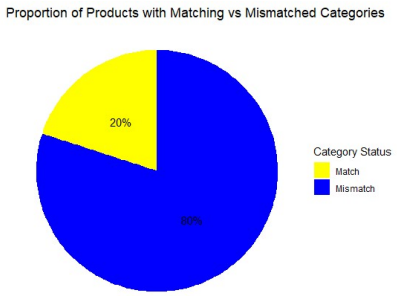
Graph 16: *Mismatched Mark-up's*

As seen on graph 16, the majority of the products mark-up values are higher than those listed by the head office data set. The discrepancies between the mark-up's and the selling prices do not mirror each other indicating that the error does not come from a miscalculation that includes these values. The error most likely stems from either the incorrect cataloguing of data or an independent calculation. These issues must be addresses before any reliable conclusions can be drawn from the data.

To further investigate the extent of the mismatched data the product category and descriptions assigned to each ProdutID entry was investigated in graph 17 and 18. It was found that 10% of the Descriptions did not match while 20% of the categories did not match across data sets. This indicated that the issue most likely lies in the communication and record keeping between the head office and the branch.

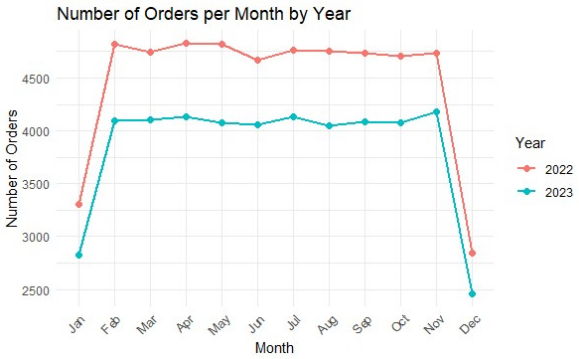


Graph 18: *Description Discrepancies*



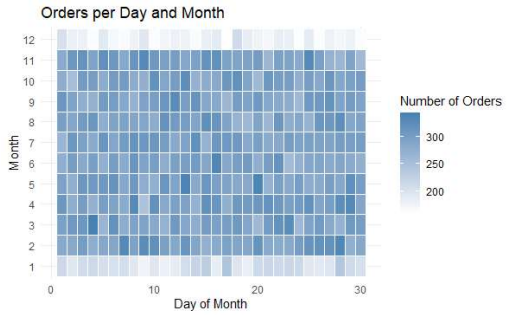
Graph 17: *Category Discrepancies*

2.2.4. Sales

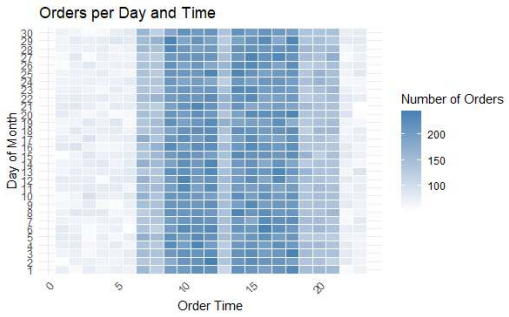


***Graph 19**: Oder Volume*

Graph 19 considers the total number of orders every month for 2022 and 2023. There was a significant decrease in sales from 2022 to 2023. December and January show the lowest sales numbers for both years. The overall trend of sales from month to month remained the same for both years however 2023 didn't have as big of a decrease in sales in June as 2022.



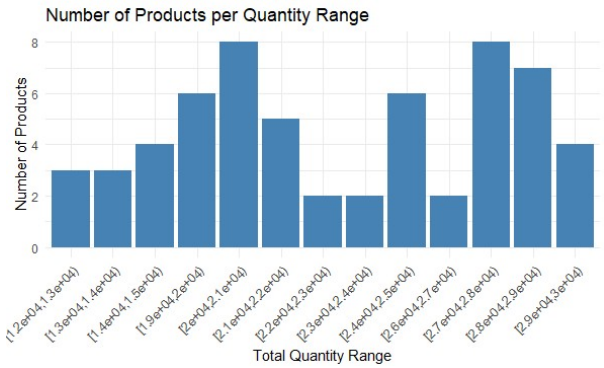
***Graph 20:** Day of Order*



***Graph 21:** Time of Order*

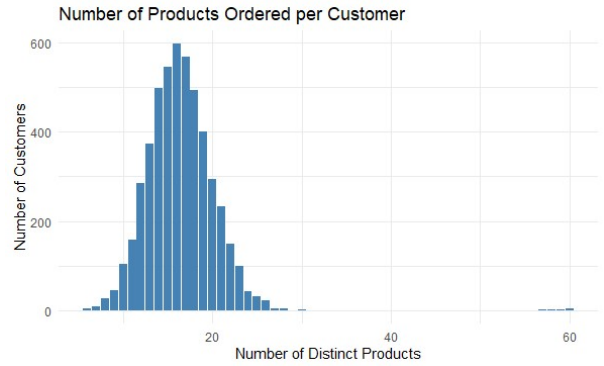
Graph 20 depicts the number of orders made each day of the month, while Graph 21 depicts the number of orders made in each hour of the day on each day. There seems to be no clear pattern in the days that orders are made; there are, however, a few patterns in the times that orders are made. There are very few sales from 23:00 to 06:00, which can be attributed to regular daylight hours. Orders begin to increase starting from 07:00, reaching a peak from 09:00 to 12:00. Orders then decline from 13:00 to 14:00 before peaking again from 14:00 to 18:00 and tapering off from 19:00

to 21:00. This creates a bimodal spread and aligns with the expected working hours in a day.



Graph 22: *Order Quantity*

Graph 22 depicts the quantity in which each product is ordered. There are two clear peaks. A large number of products are purchased in Quantity ranges of 20 000 to 21 000 and 27 000 to 28 000. All products are bought in quantities above 12 000 but below 30 000.



Graph 23: *Number of Products per Customer*

Graph 23 depicts the number of products purchased per customer. There is a clear right leaning normal distribution. This indicated that the largest number of customers order between 10 and 23 products. A very select view customers order between 55 and 60 different products.

3. STATISTICAL PROCESS CONTROL

3.1. X-BAR AND S CONTROL CARTS

New sales data was used that recorded the sales for 2026 to 2027. This data was briefly inspected and ordered chronologically by ProductID. This was done to simulate real-time data collection and can be seen on table 5.

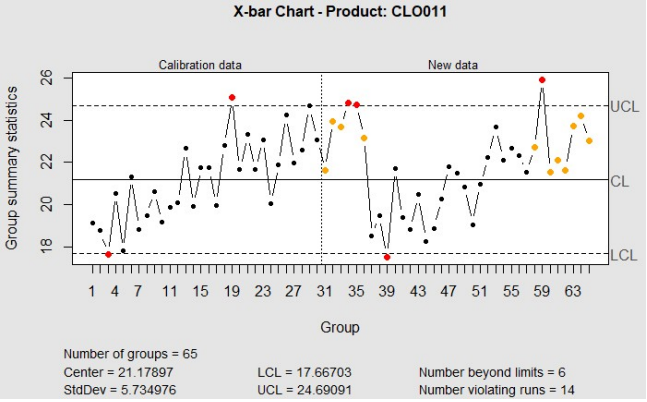
Table 5: *Sales Data (2026/2027)*

CustomerID <chr>	ProductID <chr>	Quantity <dbl>	orderTime <dbl>	orderDay <dbl>	orderMonth <dbl>	orderYear <dbl>	pickingHours <dbl>	deliveryHours <dbl>
CUST1791	CLO011	16	13	11	11	2022	17.72167	24.544
CUST3625	CLO011	1	13	9	8	2022	11.72167	21.044
CUST4239	CLO011	1	13	10	10	2022	16.38833	32.044
CUST1167	CLO011	5	16	14	6	2023	15.05750	31.046
CUST2645	CLO011	6	7	6	8	2023	12.72417	15.046
CUST1687	CLO011	37	11	10	9	2022	14.05500	15.544

6 rows

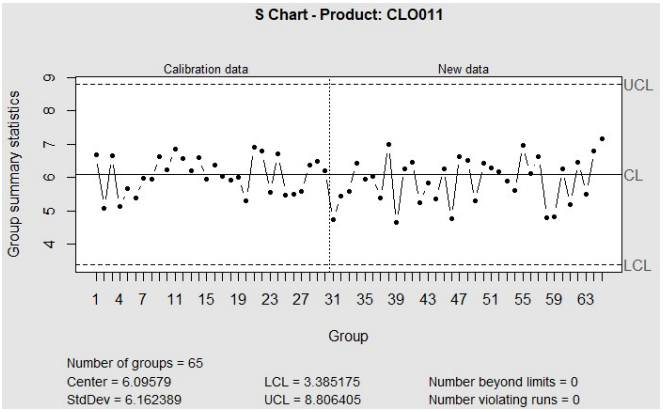
Defining the sampling is critical to the modelling process and sets the control threshold for further analyses. As specified, the data was divided into samples of 24 per product to create appropriate subgroup sizing for the SPC standard. X-bar and s control limits and centre lines were created with the first 30 samples in the data set. The SPC formulas are used to create x-bar and s charts for each product with the constants: 1, 2 & 3 sigma control limits and n=24.

The total number of graphs generated is 120. In the interest of a concise report the product CLO011 was chosen to discuss the graphs. The other graphs are provided, however to better view them it is suggested to refer instead to the provided R code.



***Graph 24**: X-Bar control chart for CLO011*

A X-bar chart shows variation in the process mean over time. This is used to help identify shifts in the average value of a measured characteristic. As seen on graph 24 many points are located close to the upper control limit. A view points are outside the upper control limit with one notably higher. There are very view points on or near the lower control limit. This implies a shift or special cause variation that affects the process mean.



***Graph 25**: S control chart for CLO011*

S charts show the variation in the standard deviation of a process over time. This helps identify changes in process consistency or variability. Graph 25 indicates that all of the standard deviations are within the control limits. This indicates a very stable process variability for the product.

Figure 2 and 3 display the X-bar and S charts for all the remaining products.

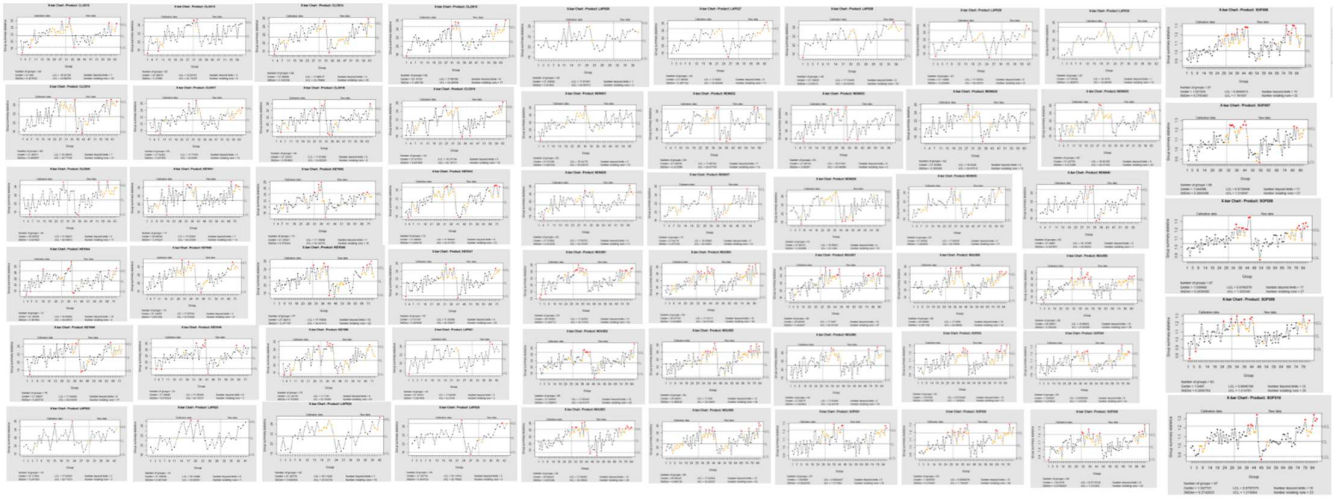


Figure 2: *X-bar Control charts for All Products*

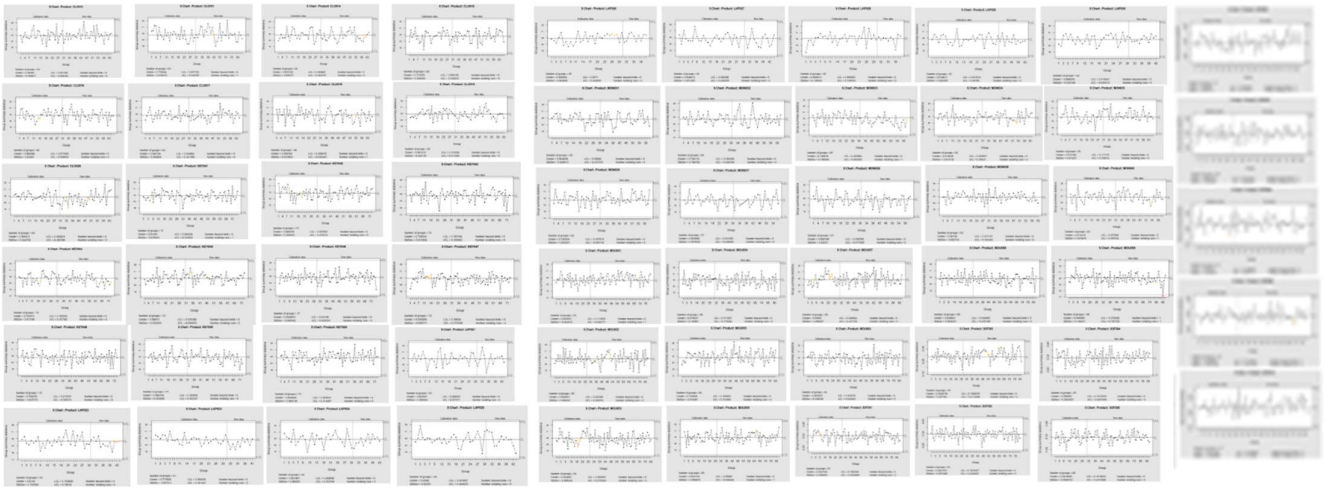


Figure 3*: S Control charts for All Products*

3.2. REAL-TIME PROCESS MONITORING

In practice, process data such as delivery times are recorded as events occur. This means that the control charts must be continuously updated and manually reviewed in real time as new information becomes available. When performing these checks, the process manager should review the S chart to determine whether the sample standard deviation falls outside the control limits. If this occurs, the process is considered unstable, and the corresponding X-bar chart will likely reflect this instability.

If the S chart remains in control, the X-bar chart must be reviewed to confirm that the sample mean is within the defined limits. Any deviation outside these limits would indicate a potential process shift or the presence of special cause variation.

The responsible manager must be notified to investigate the cause and implement corrective actions whenever a real-time control chart signals that the process is unstable or out of control. Table 6 presents a simulated real-time summary of samples that were flagged as out of control, providing early warning and guidance for timely intervention.

Table 6: *Control Charts Action*

	ProductID <chr>	SampleNumber <int>	S_Value <dbl>	Xbar_Value <dbl>	Action <chr>
1	CLO011	31	4.735772	21.62733	OK
2	CLO011	32	5.448967	23.96250	OK
3	CLO011	33	5.597457	23.66900	OK
4	CLO011	34	6.436008	24.83750	Check X-bar chart: potential shift
5	CLO011	35	5.963169	24.75417	Check X-bar chart: potential shift

3.3. PROCESS CAPABILITY

Process capability was calculated for each product type using delivery times. The lower specification limit (LSL) was assumed to be 0 hours, and the upper specification limit (USL) was defined as 32 hours, in line with the VOC requirements for acceptable

delivery times. The indices Cp, Cpu, Cpl, and Cpk were computed to evaluate both the overall capability and the centering of each process.

The Cp value represents the potential capability of the process, assuming it is perfectly centered between the specification limits. Cpu and Cpl measure the capability relative to the upper and lower specification limits, respectively. The Cpk index, being the minimum of Cpu and Cpl, provides an adjusted measure of process capability that accounts for any mean shift.

Table 7: *Process Capability*

ProductID <chr>	mean_delivery <dbl>	sd_delivery <dbl>	Cp <dbl>	Cpu <dbl>	Cpl <dbl>	Cpk <dbl>	Capable <chr>
CLO011	21.272088	6.2736472	0.8501169	0.5699987	1.130235	0.5699987	No
CLO012	21.686244	6.1681792	0.8646528	0.5573636	1.171942	0.5573636	No
CLO013	21.467364	6.2105911	0.8587481	0.5653051	1.152191	0.5653051	No
CLO014	21.335194	6.0771198	0.8776087	0.5849704	1.170247	0.5849704	No
CLO015	21.536648	6.0186104	0.8861403	0.5794999	1.192781	0.5794999	No

The calculated results are summarized in Table 7. Product types with a Cpk value greater than 1.00 are considered capable of consistently meeting VOC delivery time requirements, indicating that their process variation comfortably fits within the specification limits. Product types that are not capable require further investigation into potential sources of variability or bias.

This analysis provides a clear indication of which product types maintain a stable and capable delivery process, and which may require targeted process improvements to meet VOC performance standards. Table 8 displays the list of products that were found to be capable.

Table 8: *Confirmed Capability*

ProductID <chr>	Capable <chr>
SOF001	Yes
SOF002	Yes
SOF003	Yes
SOF004	Yes
SOF005	Yes
SOF006	Yes
SOF007	Yes
SOF008	Yes
SOF009	Yes
SOF010	Yes

10 rows

3.4. PROCESS CONTROL RULES

To evaluate the stability of the process, three standard SPC rules were applied across all product types, as seen in Table 9.

Rule A: Identifies samples where the standard deviation exceeds the 3-sigma control limits, signalling potential instability in the process.

Rule B: Highlights the longest consecutive runs of samples with S values within the ±1 sigma limits, indicating consistent process performance. The top three products with the longest consecutive runs are SOF008 with 8 samples, KEY049 with 6 samples, and MOU058 also with 6 samples.

Rule C: Identifies instances where four consecutive sample means exceed the ±2 sigma limits, signalling potential special cause variation.

Table 9: *SPC Rules*

ProductID <chr>	SampleNumber <int>	Xbar_Value <dbl>	S_Value <dbl>	mean_s <dbl>	sd_s <dbl>	mean_x <dbl>	sd_x <dbl>	RuleA_Flag <lgl>	RuleB_Flag <lgl>	RuleC_Flag <lgl>
CLO011	31	21.62733	4.735772	5.916347	0.7088607	21.68073	2.068587	FALSE	FALSE	FALSE
CLO011	32	23.96250	5.448967	5.916347	0.7088607	21.68073	2.068587	FALSE	TRUE	FALSE
CLO011	33	23.66900	5.597457	5.916347	0.7088607	21.68073	2.068587	FALSE	TRUE	FALSE
CLO011	34	24.83750	6.436008	5.916347	0.7088607	21.68073	2.068587	FALSE	TRUE	FALSE
CLO011	35	24.75417	5.963169	5.916347	0.7088607	21.68073	2.068587	FALSE	TRUE	FALSE
CLO011	36	23.16933	6.038266	5.916347	0.7088607	21.68073	2.068587	FALSE	TRUE	FALSE

Table 10 contains the a sumary of the three rules tested on the data. These analyses allow us to distinguish between random variations and instances where the process may be is affected by systematic issues.

Table 10: *SPC Rules Summary*

Rule <chr>	Total Issues <int>	First 3 ProductIDs <chr>	Last 3 ProductIDs <chr>
A	0	*NA*	*NA*
B	60	CLO011, CLO012, CLO013	SOF008, SOF009, SOF010
C	0	*NA*	*NA*

3 rows

4. DATA CORRECTION, OPTIMISING AND RISK

4.1. TYPE I ERROR

A Type I error occurs when a control chart wrongly signals that the process is out of control. In other words, the control chart issues false alarms. Assuming the process mean and variation are consistent (H_0), the theoretical probabilities of these false alarms can be estimated using statistical rules.

The probability of a false alarm for Rule A was calculated in R to be 0.27%. This was done by using the properties of the standard normal distribution to standardise any point beyond the upper 3-sigma limit with the Z-score formula.

A normal distribution is symmetric around the mean, this allows the probability of a single point lying above or below the centreline to be equal. Therefore, the probability of finding seven consecutive samples above the centreline is $0.5^7 \approx 0.0078$. This represents the false alarm probability for Rule B.

For Rule C, which flags four consecutive points beyond the upper 2-sigma limit, the probability of a false alarm is extremely low at 0.000027%. This value and assumes that the process is perfectly in control meaning any signal represents a rare statistical fluctuation rather than an actual process issue. Due to this assumption the value is theoretical.

4.2 TYPE II ERROR

A Type II error occurs when H_a is true meaning the process has shifted out of control, but the control chart has failed to detect the shift. The $\bar{X}$ chart has a target mean of 25.05 L, a lower control limits (LCL) of 25.011 L and an upper control limit (UCL) of 25.089 L for this process. Unbeknownst to us, the process mean has shifted to 25.028 L and the standard deviation has increased from 0.013 L to 0.017 L. The following

calculations are used to standardize the control limits to align with the new process mean and standard deviation values.

$$Z_{LCL} = \frac{25.011 - 25.028}{0.017} \approx -1$$

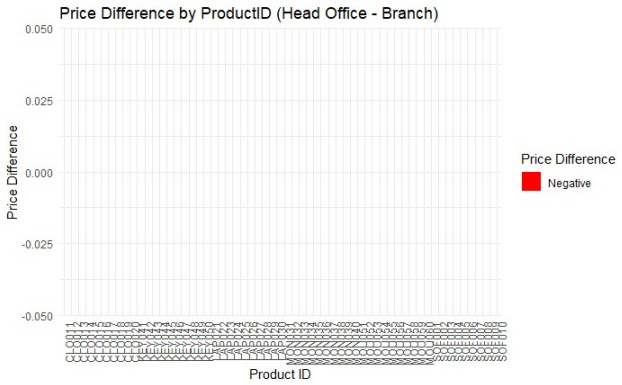
$$Z_{UCL} = \frac{25.089 - 25.028}{0.017} \approx 3.588,$$

$$\beta = \Phi(3.588) - \Phi(-1) \approx 0.9998 - 0.1587 = 0.8411$$

These calculations indicate that the likelihood of a Type II error occurring is 84.11%. (The calculation was also repeated in R yielding similar results.) This value indicates that the $\bar{X}$ chart is not very susceptible to small shifts in the process mean.

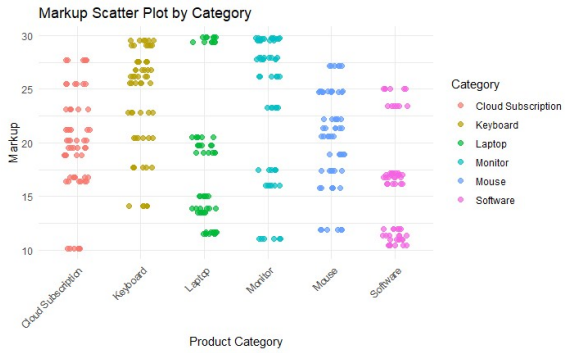
4.3. MISMATCHED DATA CORRECTION

The discrepancies noted between the products and head office data sets were addressed after receiving a response from a relevant party. The mismatched values identified were corrected and no mismatched values remain. When generating the graphs indicating the discrepancies after the correction this was confirmed. In the interest of being concise only one example of this will be displayed in graph 26.



Graph 26: *Corrected Data*

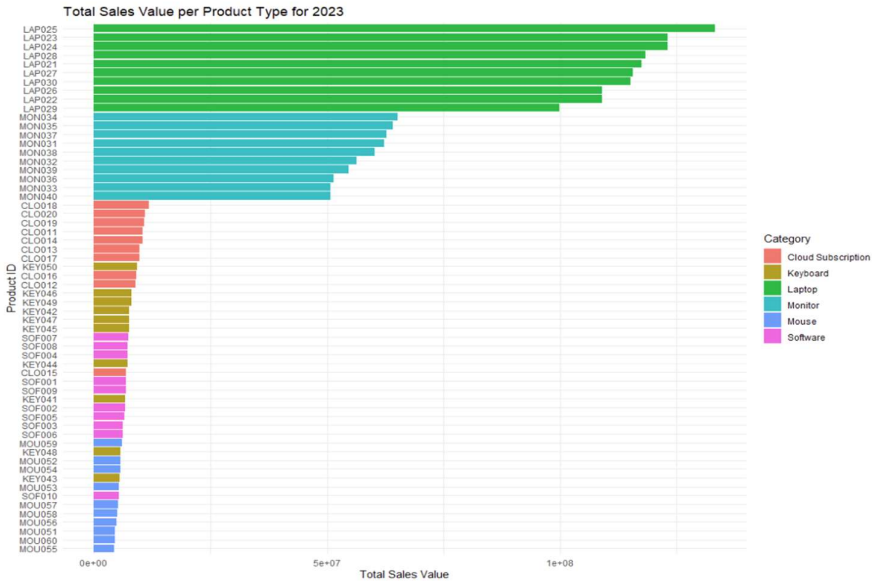
The corrected data sets were renaming as products_Headoffice2025.csv and products_data2025.csv. all previous graphs were regenerated and the noteworthy changes will now be discussed.



Graph 27: *Corrected Markup and Category*

Graph 28: *Corrected Selling Price and Category*

Graph 27 and 28 depict the mark-up and selling prices of the products in each category. Graph 28 now clearly displays a distinction between the product categories with laptops having the highest selling prices and monitors having second highest. Graph 27 shows more promising groupings of different mark-up's between categories. These groupings are likely linked to the product descriptions within these categories.



Graph 29: *Total Sales Value per Product*

As requested, graph 29 depicts the total sales value per product type. Each product is colour coded based on category. Laptops make up the overwhelming majority of sales while monitors take up the second greatest volume of sales value. This is most likely doe to the high selling prices of these products. As seen previously, these product categories have the highest selling prices.

5. MAXIMIZING PROFIT FOR COFFEE SHOPS

The provided data included the number of employees present on a day and the service time in seconds achieved on that day. This was summarised by calculating the mean service time for each number of employees per shop and including the minimum and maximum service times to assist in calculating the standard deviation of service time.

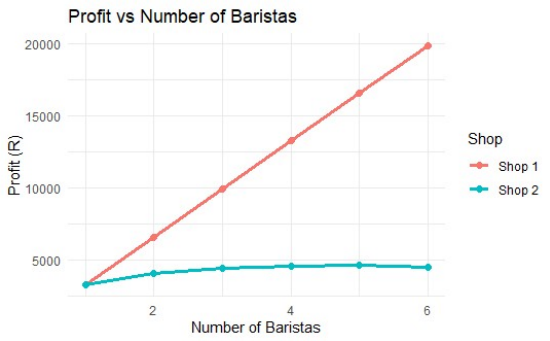
It was assumed that the shops operate 8 hours, therefor 28800 seconds in a day. The customers served per day was then calculated by dividing the total seconds in a day by the mean service time for each number of employees. Binomial calculations were performed to find the service time and the number of customers served with a 95% reliability. The total profit could then be calculated by subtracting the number of employees multiplied by a cost of R1000 from the revenue of R30 multiplied by the number of customers served.

All this information was summarised into a data frame. The data frame was filtered to find the number of employees that allow for the maximum profit for each shop. The result is provided in table 11.

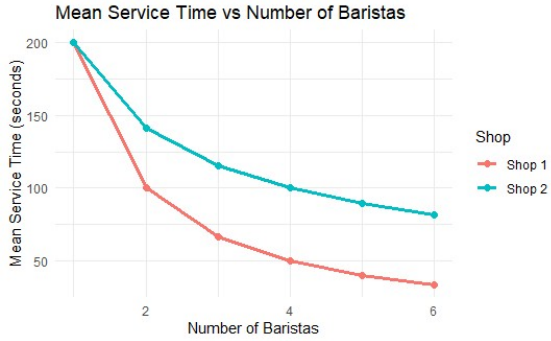
Table 11: *Maximum Pofit*

V1 <dbl>	mean_service_time <dbl>	min_service_time <dbl>	max_service_time <dbl>	customers_per_day <dbl>	revenue <dbl>	cost <dbl>	profit <dbl>	shop <chr>	service_time_95 <dbl>	customers_served_95 <dbl>
6	33.35565	13	53	863	25890	6000	19890	Shop 1	52.95565	543.8514
5	89.43597	68	112	322	9660	5000	4660	Shop 2	110.99597	259.4689

Gaph 31 and 30 were created to visualise the data and to confirm the suggested solution. As expected, the mean service time decreases as more baristas are employed. With a lower service time it allows for more customers to be served resulting in a higher profit. Interestingly for Shop 2 the total profit decreases from five employees to six while the profit for shop 1 continuously increases with the number of employees.



Graph 30: *Profit for number of Baristas*



Graph 31: *Service Time for number of Baristas*

The graphs confirm that the optimal number of employees to maximize profit is six for shop 1 and five for shop 2. This ensures a potential maximum profit of R19 890 for shop 1 and R4 660 for shop 2. The number of customers served at 95% reliability was not used to calculate the profit, instead a reliability of 100% was assumed, it is important to note that the profit will decrease depending on the reliability rate.

Comparing this to the Taguchi Loss Function the method used does not enable the company to continuously improve the service time beyond the 100% reliability threshold. The Taguchi approach focusses on the loss of quality and considers any deviation from the target value as a loss to society. This approach defines a lack of customer satisfaction and an unfavourable brand reputation as financial losses. This contest the method used that specifically promised minimising actual costs without considering the quality of the serve provided.

6. DOE AND MANOVA OR ANOVA

6.1. CONTENT OVERVIEW

Design of Experiments (DOE) refers to a systematic method to design experiments that allow for credible, objective conclusions to be drawn from their results efficiently. Changes in one or more independent variables (input) are studied to understand how the dependent variable (output) is affected. In this way the factors that have the most significant impact on the response are identified. The interaction between factors is investigated to find the most optimal combination of parameters to achieve the desired results.

Analysis of Variance (ANOVA) is used in DOE to determine whether the means of multiple samples are significantly different from each other. The null hypothesis (H_0) assumes that all sample means are equal, while the alternative hypothesis (H_1) suggests that at least one sample mean is different. When the p-value is below the chosen significance level (e.g., 0.05), H_0 is rejected, indicating that the differences in methods or conditions have a measurable effect on the mean response.

The Multivariate Analysis of Variance (MANOVA) is applied for cases involving multiple dependent variables, to assess whether the means across all samples differ simultaneously, capturing interactions between responses.

6.2. APPLIED TO 2026/2027 SALES DATA

The data was assessed and it was decided to compare the orderTime to the pickingHours in order to identify whether or not there is any correlation. The KEY049 product was chosen to perform the ANOVA analysis. The following hypothesises were made with the chosen factors:

The null hypothesis (H_0): There is no significant difference in mean order picking times for product KEY049 between different order times.

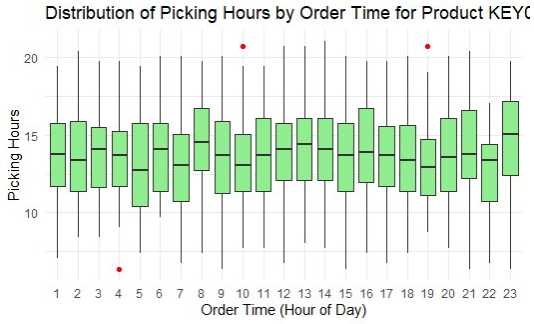
The alternative hypothesis (H_1): There is a significant difference in mean order picking times for product KEY049 between different order times.

Table 12: *ANOVA Data Preperation*

ProductID <chr>	n_hours <int>	N <int>	F_value <dbl>	p_value <dbl>
MOU059	23	2118	1.2686363	0.18044209
KEY049	23	1773	1.6574291	0.02823136
SOF009	23	2003	1.2730703	0.17729530

Table 12 displays the data preparation. Each product was assessed to find the F-value's that represent the ratio of between-group variance to within-group variance for pickingHours across different orderTime. The p-value's that refer to the probability of observing an F-value at least as extreme as the one calculated assuming the null hypothesis is true was also calculated.

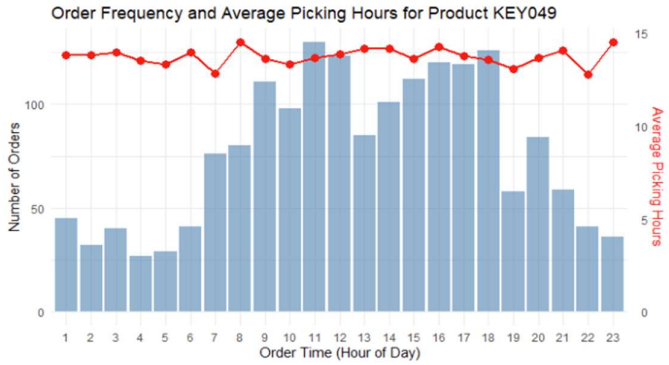
A one-way ANOVA analysis is performed with the picking hours as the responsive variable and the order time as the independent factor. A significance level of 0.5 is chosen or this analysis. Based on the p-values calculated for all the products.



Graph 32: *Picking Hours and order times KEY049*

Graph 32 depicts the box plots for the picking hours for product KEY049 at each order time. Based on the variation present on these plots it seems likely that the H_1 hypothesis will be proven true. Graph 33 is included to represent the number of orders

at each order time with an additional plot to indicate the average picking hours. This is done to possibly reveal another correlation between factors.



Graph 33: *Number of orders and order times KEY049*

To test the hypothesis' the F- and p-values of must be calculated to be compared to the chosen significance level.

Table 13: *Filtered ANOVA Data Preperation*

	ProductID <chr>	n_hours <int>	N <int>	F_value <dbl>	p_value <dbl>
2	KEY049	23	1773	1.657429	0.02823136
17	SOF001	23	2089	1.568557	0.04483152
41	SOF004	23	2046	1.658888	0.02785339

3 rows

Consulting table 12 that was described previously, the chosen product, KEY049, was considered. Table 13 displays the calculated F- and p-values. Comparing the p-value of 0.02823 to the chosen significance level of 0.5: the significance value is greater. This indicates that the H_0 hypothesis can be rejected. The H_1 is therefore assumed to be correct meaning that there is a statistically significant difference in mean picking times for product KEY049 depending on the order time.

7. RELIABILITY OF SERVICE

7.1. EXPECTED SERVICE RELIABILITY

As provided, figure 4 depicts the number of people on duty over 397 days.

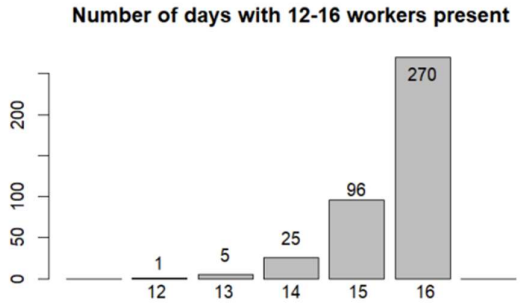


Figure 4: *Provided Data*

The minimum number of workers needed to prevent problems from occurring is 15. The current reliability rate was calculated to be 92,19%. This was calculated by summing the total number of days with a minimum of 15 workers and dividing it by the total number of days in the year.

This value is already satisfactory for most companies, given that it ensures customer satisfaction for 336 days (92,19%) of the year, however it can still be improved.

7.2. OPTIMISE PROFIT

Knowing that each day with problems negatively impacts sales by R20 000, but also that appointing additional personnel costs R25 000 per month per person a balance must be struck between the desired reliability and the cost associated with the change.

It was assumed that 16 workers were scheduled for every day and that the actual attendance was documented in the data provided. The average number of workers present each day (k-bar) is 15.5844. The average numbers of workers present each day was calculated into an estimated attendance probability (p) of 0.974 by dividing it by the number of scheduled workers.

The following was performed in R. A standard error (SE) was calculated for the binomial proportion using 16 workers as the initial number of staff employed and the attendance probability calculated previously (p = 0.974).

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

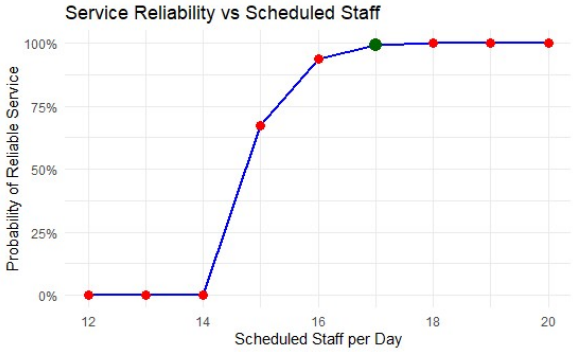
A 95% upper and lower confidence interval was then calculated with the provided formula. The values were found to be 0.896 for the lower limit and 1.052 for the upper limit. Both these values are rounded to 1 due to being applied to workers.

$$\boldsymbol{CI = p \pm 1.95 \times SE}$$

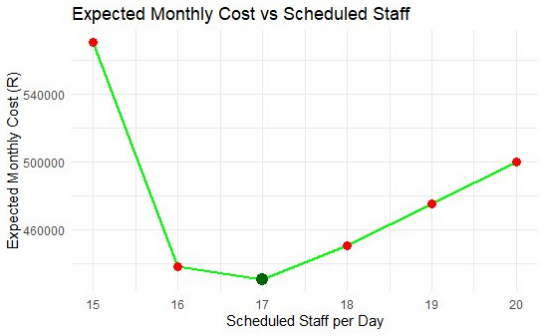
Given these CI values, the limits are 15 and 17 workers. The attendance probability (p = 92,19%) was used to calculate the probability of reliable service at each number of staff; he results are depicted in graph 34.

The expected monthly cost at each number of staff was calculated by comparing the loss in sales to the cost of an additional worker and is depicted in graph 35.

These two graphs were consulted and compared to find the optimal number of staff (S*) with the highest probability of reliable service and lowest expected monthly cost. The answer was found to be 17 staff members to offer 99.1% service reliability at a cost of R430 443per month.



Graph 34: *Service Reliability Optimization*



Graph 35: *Monthly Costs Optimization*

8. CONCLUSION

This report successfully applied core industrial engineering and quality assurance principles to a comprehensive business dataset. The initial analysis revealed several data quality issues, most notably mismatches between the pricing and markup values in the product dataset and the head office dataset. These discrepancies were systematically corrected, ensuring the integrity of all subsequent analyses.

The application of Statistical Process Control (SPC) showed that while process variability was generally stable across most products, certain products were flagged as out of control. This highlights the need for management to investigate and address the special causes of variation. Process capability analysis further indicated that non-physical products were highly capable, whereas the delivery processes for physical goods did not consistently meet the VOC standard of 32 hours.

Optimization models provided clear, actionable solutions. For the coffee shop, scheduling six baristas at location one and five at location two maximized profits, while for the car rental agency, 17 employees per day were required to maintain a service reliability of 99.1%. The staffing recommendation for the car rental agency was supported by a binomial model, demonstrating the practical value of predictive workforce planning.

Finally, ANOVA analysis confirmed that the difference in order times for product KEY049 is statistically significant with respect to mean picking hours. Overall, the integrated approach of data correction, process monitoring, risk assessment, and profit optimization provides a robust, actionable framework for improving operational quality, reducing financial loss, and supporting informed, data-driven management decisions.

The project's findings delivered clear insights into process stability, data governance, and strategic resource allocation, and provided practical guidance for continuous operational improvement.

9. REFERENCES

ECSA Project Brief. (2025). Preamble to the Engineering Counsel of South Africa (ECSA) report that proves graduate attribute 4 (ECSA GA4) in 2025 (ProjectECSA2025Final.pdf). Stellenbosch University, Department of Industrial Engineering.

Stellenbosch University, Department of Industrial Engineering. (2025). Statistical Methods in Quality Assurance Part 1 summary (Study Material: QA344 Statistics.pdf).