ECSA Project Final Report

Kaitlyn Venter - 27005879

Contents

List of figures

List of Tables

Section 1

1.1 Data Loading and Inspection

Products_data:

```
## 'data.frame':    60 obs. of  5 variables:
##  $ ProductID   : chr  "SOF001" "SOF002" "SOF003" "SOF004" ...
##  $ Category    : chr  "Software" "Cloud Subscription" "Laptop" "Monitor" ...
##  $ Description : chr  "coral matt" "cyan silk" "burlywood marble" "blue silk" ...
##  $ SellingPrice: num  512 505 494 543 516 ...
##  $ Markup      : num  25.1 10.4 16.2 17.2 11 ...
```

ProductID	Category	Description	SellingPrice	Markup
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
SOF004	Monitor	blue silk	542.56	17.19
SOF005	Keyboard	aliceblue wood	516.15	11.01
SOF006	Mouse	black silk	478.93	16.99

Figure 1: Products_data structure

Products_headoffice:

```
## 'data.frame':    360 obs. of  5 variables:
##  $ ProductID   : chr  "SOF001" "SOF002" "SOF003" "SOF004" ...
##  $ Category    : chr  "Software" "Software" "Software" "Software" ...
##  $ Description : chr  "coral silk" "black silk" "burlywood marble" "black marble" ...
##  $ SellingPrice: num  522 467 496 389 483 ...
##  $ Markup      : num  15.6 28.4 20.1 17.2 17.6 ...
```

ProductID	Category	Description	SellingPrice	Markup
SOF001	Software	coral silk	521.72	15.65
SOF002	Software	black silk	466.95	28.42
SOF003	Software	burlywood marble	496.43	20.07
SOF004	Software	black marble	389.33	17.25
SOF005	Software	chartreuse sandpaper	482.64	17.60

Figure 2: Products_headoffice structure

Customer_data:

```
## 'data.frame':    5000 obs. of  5 variables:
##  $ CustomerID: chr  "CUST001" "CUST002" "CUST003" "CUST004" ...
##  $ Gender    : chr  "Male" "Female" "Male" "Male" ...
##  $ Age       : int  16 31 29 33 21 32 31 27 26 28 ...
##  $ Income    : num  65000 20000 10000 30000 50000 80000 100000 90000 35000 105000 ...
##  $ City      : chr  "New York" "Houston" "Chicago" "San Francisco" ...
```

CustomerID	Gender	Age	Income	City
CUST001	Male	16	65000	New York
CUST002	Female	31	20000	Houston
CUST003	Male	29	10000	Chicago
CUST004	Male	33	30000	San Francisco
CUST005	Female	21	50000	San Francisco
CUST006	Male	32	80000	Miami

Figure 3: Customer_data structure

Sales2022and2023

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

CustomerID	ProductID	Quantity	orderTime	orderDay	orderMonth	orderYear	pickingHours	deliveryHours
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
CUST3721	LAP024	31	12	18	7	2023	41.39083	24.546
CUST4605	CLO012	20	14	7	2	2022	15.72167	24.044
CUST2766	MON035	32	21	24	12	2022	21.05500	24.044

Figure 4: Sales2022and2023 structure

After inspection of the data, the following is evident:

- The products_data dataset consists of 60 instances with 5 columns - "ProductID", "Category", "Description", "SellingPrice" and "Markup". The first 3 of these are character type variables, while the last 3 are numerical.
- The products_headoffice is comprised of 360 observations with the same 5 columns as the products_data dataset.
- The customer_data dataset is composed of 5000 instances with 5 columns - "CustomerID", "Gender", "Age", "Income" and "City". The CustomerID, Gender and City are classified as character types, Income is numerical, and Age is an integer variable.
- The sales2022and2023 dataset consists of 100000 observations with 9 columns - “CustomerID", "ProductID", "Quantity", "orderTime", "orderDay", "orderMonth", "orderYear", "pickingHours", "deliveryHours". ProductID and CustomerID are classified as characters; Quantity, orderTime, orderDay, orderMonth, orderYear are integer variables; and pickingHours and deliveryHours are classified as numerical.

When comparing the samples of the products_data and products_headoffice datasets, it is evident that there are discrepancies between the category variables and description variables and their corresponding productID. For example, in products_data, the product ID of SOF004 corresponds to a blue silk monitor whereas in products_headoffice, the product corresponding to this ID is black marble software. This inconsistency in data compromises its integrity, making it less reliable and likely to result in issues when trying to analyze the data further. This mismatched product information may be due to data entry errors, misaligned coding standards for product categories, or inadequate version control for product descriptions.

1.2 Summary Statistics

Table 1: Products_data summary statistics

Variable	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
ProductID*	1	60	30.50000	17.464249	30.500	30.50000	22.239000	1.00	60.00	59.00	0.0000000	-1.2601448	2.2546249
Category*	2	60	3.50000	1.722237	3.500	3.50000	2.223900	1.00	6.00	5.00	0.0000000	-1.3258048	0.2223399
Description*	3	60	16.40000	10.078001	16.000	16.20833	13.343400	1.00	35.00	34.00	0.1029599	-1.2935763	1.3010643
SellingPrice	4	60	4493.59283	6503.770150	794.185	3189.25479	525.722547	350.45	19725.18	19374.73	1.4261752	0.4338057	839.6331159
Markup	5	60	20.46167	6.072598	20.335	20.51187	7.309218	10.13	29.84	19.71	-0.0367077	-1.2380989	0.7839690

Variable	Q1.25%	Q3.75%	IQR
SellingPrice	512.1825	6416.6600	5904.4775
Markup	16.1400	25.7075	9.5675

The summary statistics of products_data show that SellingPrice has a high standard deviation of 6503.77 which indicates a wide range of product prices. The mean of the SellingPrice is 4493.59 with the 3rd quartile equal to 6416.6600, however, the maximum selling price is 19725.18, which indicates that the distribution of SellingPrce is skewed towards lower prices. This can also be inferred by the skewness value of 1.43. On the other hand, Markup has a standard deviation of 6.07 and somewhat large range of 19.71. Considering that the minimum value is 10.13, the maximum value is 29.84 and the mean value is 20.46, it could be deduced that Markup has a normal distribution, which is further supported by the low skewness value of -0.04. However, the kurtosis value of -1.24 may suggest that the distribution is relatively flat and may be indicative of a uniformly distributed variable rather than a normally distributed one.

Table 2: Products_headoffice summary statistics

Variable	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
ProductID*	1	360	69.38889	23.217847	72.000	71.88542	22.239000	1.00	110.00	109.00	-0.8665172	0.4912730	1.2236880
Category*	2	360	3.50000	1.710202	3.500	3.50000	2.223900	1.00	6.00	5.00	0.0000000	-1.2781771	0.0901356
Description*	3	360	30.68611	17.319505	29.500	30.76736	22.980300	1.00	60.00	59.00	-0.0277818	-1.3900365	0.9128181
SellingPrice	4	360	4410.96186	6463.822788	797.215	3054.22903	515.752062	290.52	22420.14	22129.62	1.5279096	0.7789339	340.6733733
Markup	5	360	20.38550	5.665949	20.580	20.42868	6.664287	10.06	30.00	19.94	-0.0477692	-1.0739041	0.2986217

Variable	Q1.25%	Q3.75%	IQR
SellingPrice	495.9375	5843.333	5347.395
Markup	15.8400	24.845	9.005

The summary statistics of the products_headoffice data presents a similar standard deviation value (6463.82) for SellingPrice to that of the SellingPrice in the products_data dataset. The mean of the SellingPrice for this dataset is 4410.96 with a 3rd quartile value of 5843.332 and maximum value of 22420.14. This suggests that the distribution is similar to that of the SellingPrice distribution in the products_data dataset but with a slightly greater skewness towards lower selling prices which is further reinforced by a skewness value of 1.53. The standard deviation of the Markup (5.67) for this dataset is relatively smaller in comparison to the Markup in the products_data data but shows a similar range of 19.94. With a minimum value of 10.06, maximum value of 30.00 and mean of 20.39, the distribution of this Markup variables from this dataset and the products_data dataset likely share a similar distribution to one another which may be an indication of redundancy.

Table 3: Customer_data summary statistics

Variable	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
CustomerID*	1	5000	2500.5000	1443.520003	2500.5	2500.50000	1853.2500	1	5000	4999	0.0000000	-1.2007200	20.4144557
Gender*	2	5000	1.5572	0.577923	2.0	1.51700	1.4826	1	3	2	0.4538869	-0.7240466	0.0081731
Age	3	5000	51.5538	21.216096	51.0	50.88275	26.6868	16	105	89	0.2041739	-0.9874439	0.3000409
Income	4	5000	80797.0000	33150.106741	85000.0	81665.00000	37065.0000	5000	140000	135000	-0.2135307	-0.7456542	468.8133055
City*	5	5000	3.9918	2.002232	4.0	3.98975	2.9652	1	7	6	-0.0108635	-1.2745838	0.0283158

Variable	Q1.25%	Q3.75%	IQR
Age	33	68	35
Income	55000	105000	50000

Analysis of the summary statistics for the customers_data reveals a standard variation of 21.22 for Age and a mean of 51.55. The maximum value for Age is 105 which cannot be considered feasible and may be indicative of an error in the data. The 1st quartile for Age is 33 while the 3rd is 68. Additionally, the values for skewness (0.20) and the kurtosis (-0.99) suggest that the distribution of the values for Age are likely close to being uniformly distributed and slightly right-skewed. On the other hand, the summary statistics for Income display a large standard deviation of 33150.11 and range of 135000 which indicates that this variable has a wide distribution of values with the minimum being 5000 and maximum being 140000. The mean value for this variable is 80797.00, the 1st quartile is equal to 55000 and the 3rd to equal to 105000. As the value for skewness is -0.21 and the value for kurtosis is 0.75, this implies that the distribution for Income is slightly left-skewed and close to being normally distributed.

Table 4: sales2022and2023 summary statistics

Variable	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
CustomerID*	1	1e+05	2492.33848	1444.5778106	2503.000	2491.191987	1862.1456	1.0000000	5000.0000	4999.00000	0.0020964	-1.2107656	4.5681561
ProductID*	2	1e+05	32.43610	18.0302099	35.000	32.819100	23.7216	1.0000000	60.0000	59.00000	-0.1603407	-1.3178154	0.0570165
Quantity	3	1e+05	13.50347	13.7601316	6.000	11.458100	5.9304	1.0000000	50.0000	49.00000	1.0443411	-0.2185180	0.0435134
orderTime	4	1e+05	12.93230	5.4951268	13.000	13.117888	5.9304	1.0000000	23.0000	22.00000	-0.2271685	-0.7101693	0.0173771
orderDay	5	1e+05	15.49683	8.6465055	15.000	15.495088	10.3782	1.0000000	30.0000	29.00000	0.0027726	-1.2007412	0.0273427
orderMonth	6	1e+05	6.44813	3.2834460	6.000	6.445538	4.4478	1.0000000	12.0000	11.00000	0.0069282	-1.1764404	0.0103832
orderYear	7	1e+05	2022.46273	0.4986115	2022.000	2022.453413	0.0000	2022.0000000	2023.0000	1.00000	0.1494937	-1.9776714	0.0015767
pickingHours	8	1e+05	14.69547	10.3873345	14.055	13.543098	6.9188	0.4258889	45.0575	44.63161	0.7357093	0.4143469	0.0328476
deliveryHours	9	1e+05	17.47646	9.9999440	19.546	17.775077	8.8956	0.2772000	38.0460	37.76880	-0.4704880	-0.8716457	0.0316226

Variable	Q1.25%	Q3.75%	IQR
Quantity	3.000000	23.00000	20.000000
orderTime	9.000000	17.00000	8.000000
orderDay	8.000000	23.00000	15.000000
orderMonth	4.000000	9.00000	5.000000
orderYear	2022.000000	2023.00000	1.000000
pickingHours	9.390833	18.72167	9.330833
deliveryHours	11.546000	25.04400	13.498000

The summary statistics for sales reveal that Quantity has a mean value of 13.50347 with the 1st quartile value being 3.0 and the 3rd quartile value being 23.0. Considering this variable has a large range of 49.0, a skewness value of 1.0443 and a kurtosis of -0.2185, this variable likely has right-skewed distribution with potential outliers towards larger values. The mean value of orderTime is approximately 12.93 and standard deviation of about 5.5 showing relatively large variation in order timing. The skewness value of -0.2272 and kurtosis of -0.7102 indicates that the distribution is slightly left-skewed and close to being normally distributed but is slightly flat. Majority of orders are likely to take place between 09:00 and 17:00 as indicated by the 1st quartile being 9 and the 3rd quartile being 17. It can be deduced that the pickingHours have a relatively flat and slightly right-skewed distribution from the skewness value of 0.7357 as well as the kurtosis value of 0.4143. The pickingHours also displays a mean value of 14. 6958, standard deviation of 10.3873 and range of 44.6313, indicating a wide variation in the time spent on picking activities. In comparison, the deliveryHours, variable can be expected to have a different distribution, as its summary statistics present a skewness value of -0.4705 and kurtosis of -0.8716, suggesting a left-skewed distribution with a greater peak than that of pickingHours. However, the delivery hours show a similar standard deviation of 9.9999, a slightly lower range of 37.7688 and slightly higher mean value of 17.4765.

1.3 Handling Missing Values

There are no missing values (which was tested using the sum(is.na()) function) and no duplicated instances within any of the datasets.

1.4 Data Filtering and Subsetting

The products_data dataset was filtered to exclude customers below 18 (below legal adult age) and customers above the age of 90, which is above the typical consumer age, so as to focus the analysis on a more likely active customer base.

A separate products_data dataset was created with products that had a markup value greater than 20 to focus on products that have meaningful products margin for meaningful sales analysis and business insights.

The sales data was also filtered to only include order times ranging from 7am to 8pm to focus on orders places during more typical business hours for operational analysis.

1.5 Data Visualization

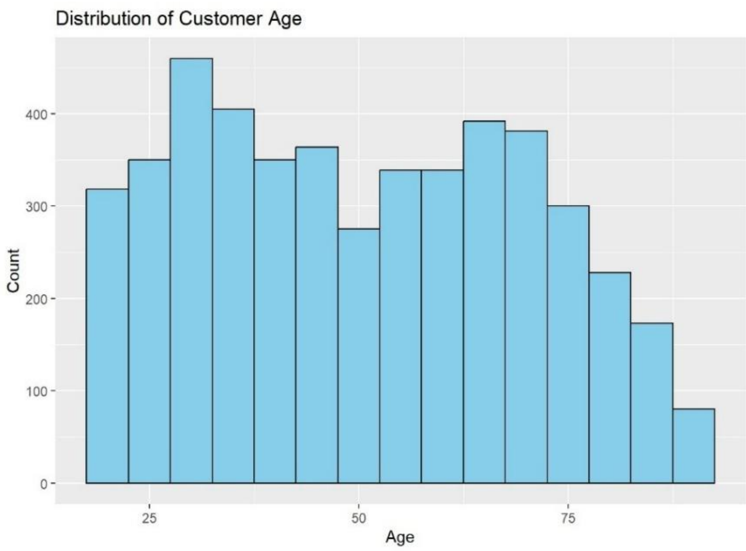

Figure 5: Histogram of Customer Age

The distribution of the customer ages is somewhat bimodal with a large majority of customers tending towards the age of approximately 30, and a secondary group tends to being approximately 65, with a large decline after the age of 70. In consideration of this, marketing and product targeting should reflect a wide customer base and should consider age-stratified analyses (such as age grouping) when planning targeted campaigns.

Figure 6: Boxplots of customer income per city

The boxplot distribution of income by city illustrates that the mean income does not vary significantly across cities. Seattle, Miami, Los Angeles, Houston and Chicago share an identical mean income of approximately 81 000, while San Francisco and New York exhibit slightly lower mean incomes. Miami and Chicago present identical interquartile ranges to each other, suggesting similar income dispersion within these cities. The remaining cities, however, display larger and identical interquartile ranges to each other, with the first quartile equal to approximately 55 000 and the third equal to approximately 105 000, reflecting a wider spread of income levels.

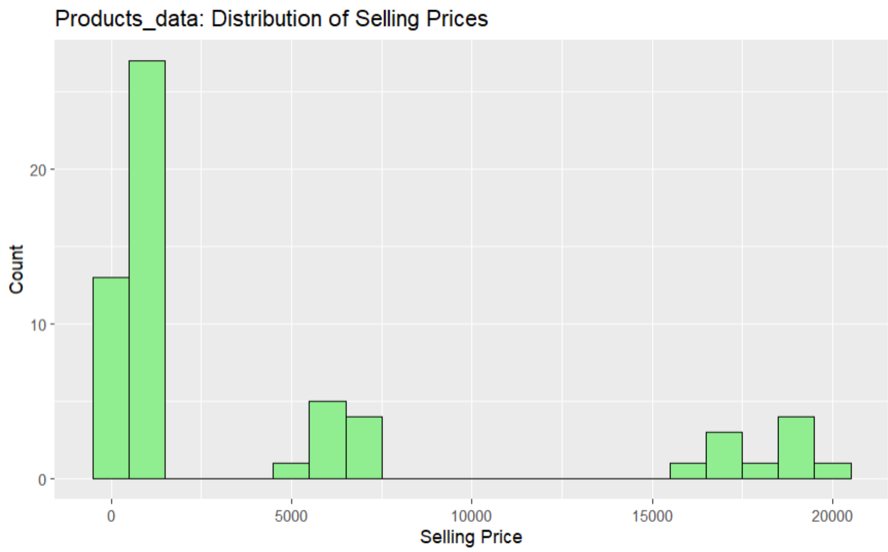

Figure 7: Histogram of selling prices for products_data

The selling prices for the products_data dataset is shown to have a right-skewed distribution with a majority of the products typically having prices far below 5000 and likely ranging between 350 and 2000. This distribution may be due to the company having a majority of products that are low to medium in value and a very small minority class of high value products that range from approximately 15500 to 20000.

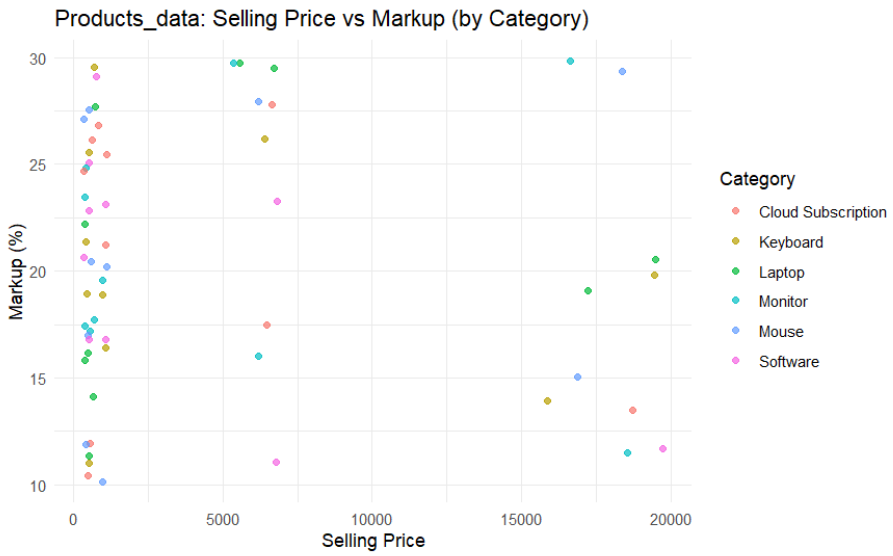

Figure 8: Scatter plot of markup (%) vs selling prices for products_data

The scatter plot of selling price versus markup illustrates how the markup percentage varies with the selling price for different categories in the products_data dataset. The data points are differentiated by colour according to the product category, thus helping to identify relationships between the selling prices and markup for certain types of products. A majority of data points are concentrated along lower selling prices (below 2000) with varied markups between 10% and 30%, despite the different products categories. For selling prices greater than 5000 to below 20 000, the markup per category decreases significantly with the exception of two outlying products, one classified as a monitor and the other as a mouse. These outlying products exhibit selling prices near 17 500 and yet both show a markup greater than 28% and thus may require further investigation.

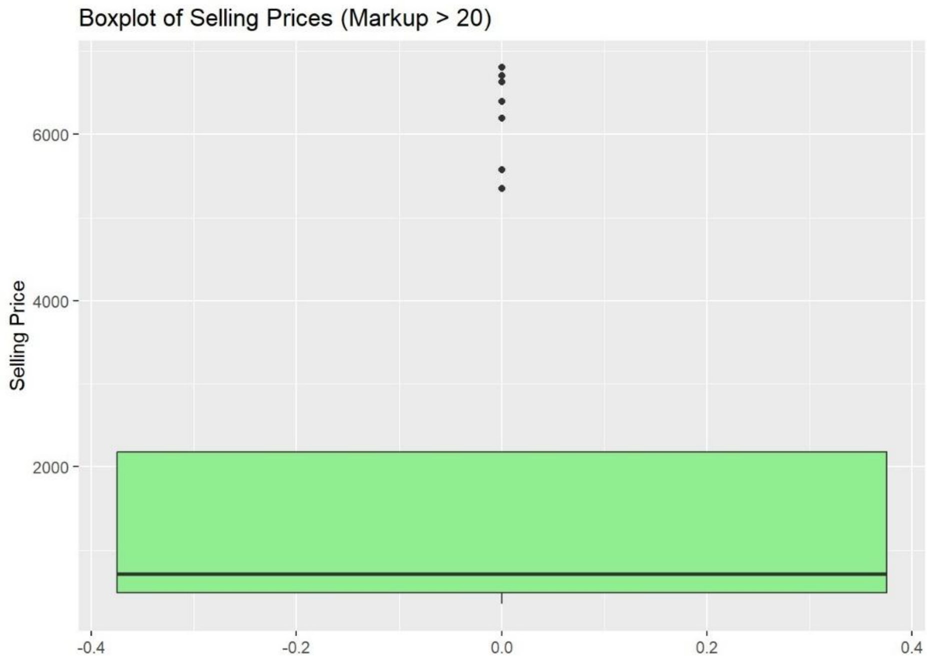

Figure 9: Boxplot of selling prices for products_data

The boxplot above displays the selling prices for the products_data after filtering to exclude products with a markup of less than 20%. The selling prices are now shown to have a mean of about 700 and a relatively high third quartile near 2200, suggesting considerable price dispersion despite the markup of greater than 20%. The boxplot also presents products with selling prices greater than 5000 as being outliers, indicating that there are products with unusually expensive products. In correspondence with the previous scatter plot displaying the relationship between selling prices and markup, these products with outlying selling prices should be investigated in order to identify any potential errors.

Figure 10: Histogram distribution of selling prices for products_headoffice

The selling prices for the products_headoffice dataset is shown to have a similar right-skewed distribution with a large portion of the products typically having prices far below

5000 and approximately ranging between 290 and 2000. Once again, this distribution is likely due to the company having a majority of products that are low to medium in value and a very small minority class of high value products that range from approximately 12500 to 18500. Although selling prices presented by the products_headoffice dataset and those presented in the products_data share very similar distributions, the range of values differ slightly between the two datasets, and it is thus apparent that the two datasets show discrepancies between their selling prices (and markup values) as stated previously.

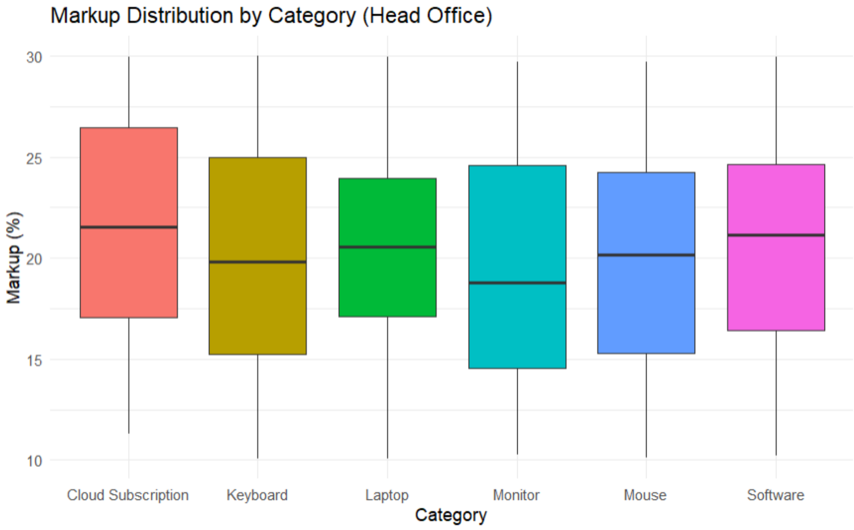

Figure 11: Boxplots of Markup Distribution by Category (products_headoffice)

The boxplots shown per category for products in the products_headoffice dataset suggest that the mean markup does not change significantly between each of the product categories, as the mean values only range approximately between 18.5% and 22%. Majority of the products per category have a 1^{st} quartile value for markup above 15% and a 3^{rd} quartile value below 25%. Laptops, in particular, exhibit the smallest variation in markup, with an interquartile range of only about 7%. Conversely, Cloud Subscriptions tend to have higher markup values, reflected in the highest mean of approximately 21.5% and third quartile (around 26.5%).

Figure 12: Boxplot of order times for sales2022and2023

Figure 14 displays a boxplot of the order times throughout a day with a mean value of around 13 to 14 which suggests that the average orders typically occur early-to-mid afternoon. The interquartile range is approximately 5hrs spanning from 11 to 16, indicating that the majority of orders take place from 11am to 4pm. The lack of skewness and outliers indicates a stable and predictable order behaviour during business hours.

Figure 13: Scatter plot of quantity sold vs selling price for sales2022and2023

The scatterplot of the quantity of products sold versus the selling price presents a high variation in quantity for low selling prices. As the selling prices increase, the quantity of products sold becomes slightly less dense for higher quantity values, indicating that customers tend to buy in lower quantities for higher value products.

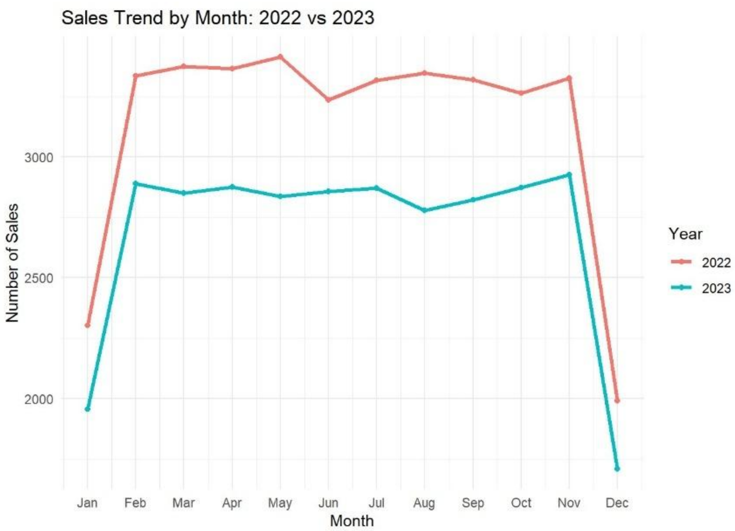

Figure 14: Number of sales per year trend by month

The sales trend compares the years 2022 and 2023 and shows that more sales were consistently made in 2022 compared to 2023. This trend may require further investigation into factors affecting sales, such as changes in consumer behaviour, economic conditions, or product offerings. The number of sales for both years display a similar shape with a significantly lower number of sales made in both January and December and a higher number of sales between these two months. The sales in 2022 display a steady increase in the number of sales from February through to May and then a more drastic decrease from approximately 3400 to 3250 between May and June. It then increases slightly to around 3350 in August and decreases back to 3250 in October. The number of sales in 2023 show little variance between the months of February and July, as they remain relatively close to 2850. These number of sales then decrease slightly towards August and then increase to approximately 2900 by November.

1.6 Scatterplot Matrices

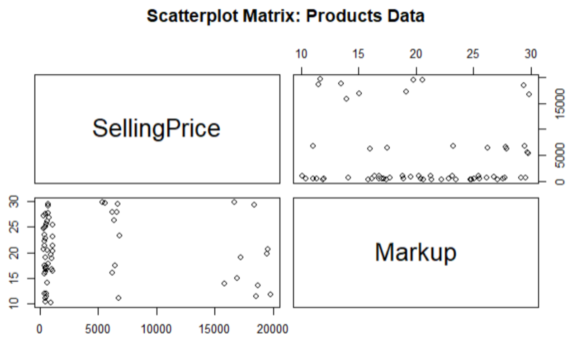

Figure 15: Scatterplot Matrix: Products_data

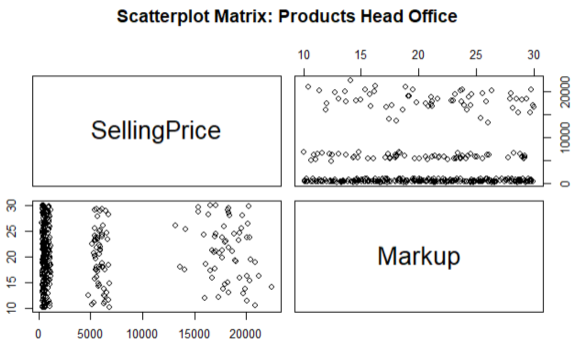

Figure 16: Scatterplot Matrix: Products_headoffice

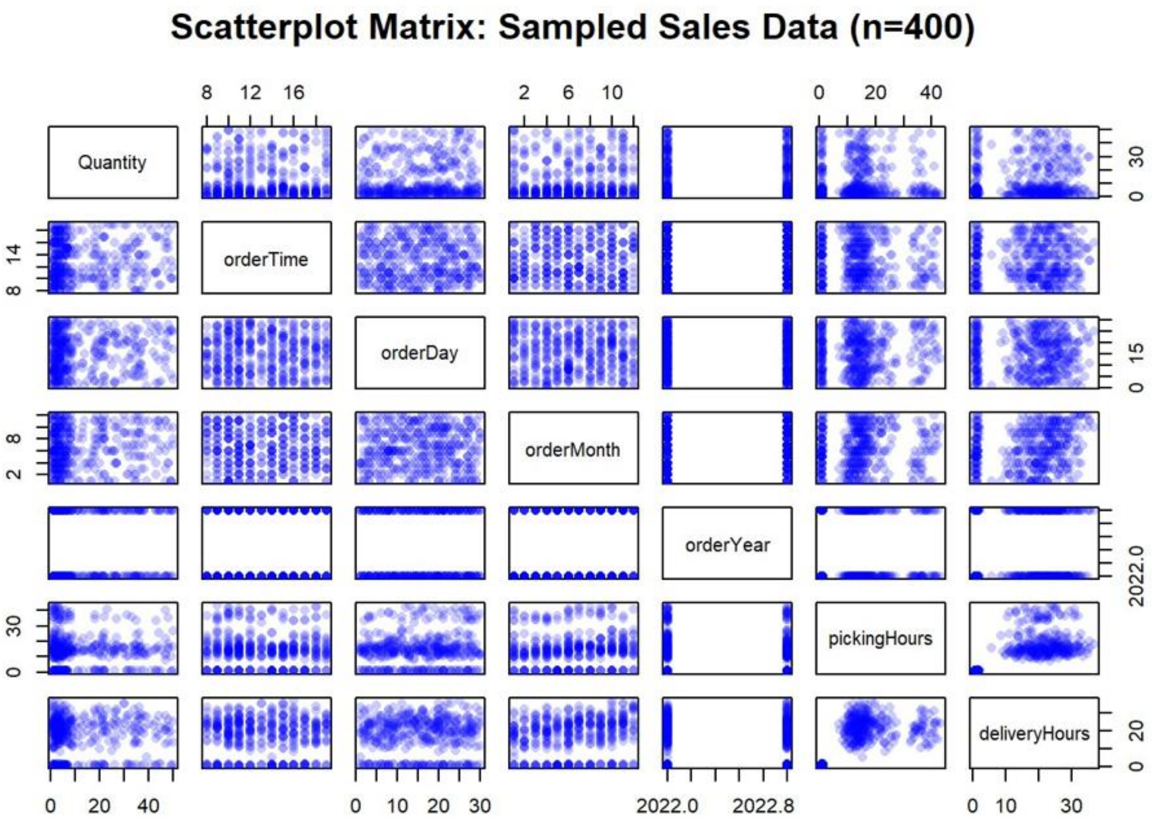

Figure 17: Scatterplot Matrix: Sampled sales data

Only 400 instances were randomly sampled from the sales data in order to achieve clear visibility of the correlations between the different variables. However, many of the features from this dataset do not share a clear correlation. orderTime, orderDay and orderMonth show discrete banded patterns, suggesting that the variables have limited possible values. The relationship between Quantity and the other features shows no strong correlations, suggesting that the quantity sold varies independently of other variables.

Section 3

3.1 Initialisation of X-charts and s-charts

The data for sales2026and2027Future was first rearranged to be in chronological order based on the year, month, day and order time in preparation of assessing the relative stability of the delivery process. This was done to simulate real-time data arrival and prepare for the SPC to be correctly implemented. Table 1 represents a sample of the sales data after this step was executed. The centre line, standard deviation and subsequently the control limits of sigma one, two and three for both the X-bar and s-charts were then calculated on the basis of using the first 30 samples per product with each of these samples containing 24 consecutive delivery times. Table 2 presents a sample of the means and standard deviations calculated from each of the first 30 samples.

*Note: despite the file being called sales2026and2027Future.csv, the order years display 2022 and 2023. I will assume it should be for the years 2022 and 2023 from this point.

Table 5: Arranged sales sample

	customer_id	product_id	quantity	order_time	order_day	order_month	order_year	picking_hours	delivery_hours	.timestamp
1	CUST3795	MOU059	4	1	1	1	2022	16.3883333	9.5440	2022-01-01 01:00:00
2	CUST2337	KEY049	7	1	1	1	2022	10.3883333	18.5440	2022-01-01 01:00:00
3	CUST3281	SOF009	5	1	1	1	2022	0.4258889	0.6772	2022-01-01 01:00:00
4	CUST3721	CLO019	47	1	1	1	2022	11.3883333	19.5440	2022-01-01 01:00:00
5	CUST4015	KEY045	1	1	1	1	2022	12.3883333	15.5440	2022-01-01 01:00:00
6	CUST3701	SOF010	2	2	1	1	2022	0.8925556	1.4272	2022-01-01 02:00:00
7	CUST1489	KEY046	39	3	1	1	2022	7.3883333	14.5440	2022-01-01 03:00:00
8	CUST2905	SOF009	1	3	1	1	2022	1.0925556	1.5272	2022-01-01 03:00:00
9	CUST597	CLO012	7	5	1	1	2022	6.3883333	20.5440	2022-01-01 05:00:00
10	CUST1246	KEY047	17	7	1	1	2022	15.3883333	24.5440	2022-01-01 07:00:00

Table 6: Sample means and sd summary sample

	sample	n	sample_mean	sample_sd	start_time	end_time
1	1	24	0.9792833	0.2684412	2022-01-01 02:00:00	2022-01-07 18:00:00
2	2	24	0.9980333	0.2918891	2022-01-08 01:00:00	2022-01-18 23:00:00
3	3	24	0.9480333	0.2937452	2022-01-19 09:00:00	2022-01-28 08:00:00
4	4	24	0.8938667	0.2919822	2022-01-28 14:00:00	2022-02-05 11:00:00
5	5	24	0.9022000	0.3290302	2022-02-05 20:00:00	2022-02-11 18:00:00
6	6	24	0.9126167	0.2722607	2022-02-11 22:00:00	2022-02-16 14:00:00
7	7	24	0.9980333	0.2812691	2022-02-16 15:00:00	2022-02-19 17:00:00
8	8	24	0.9834500	0.2486103	2022-02-19 20:00:00	2022-03-01 21:00:00
9	9	24	0.9532417	0.2574413	2022-03-02 03:00:00	2022-03-10 11:00:00
10	10	24	0.9688667	0.2846304	2022-03-10 16:00:00	2022-03-15 20:00:00

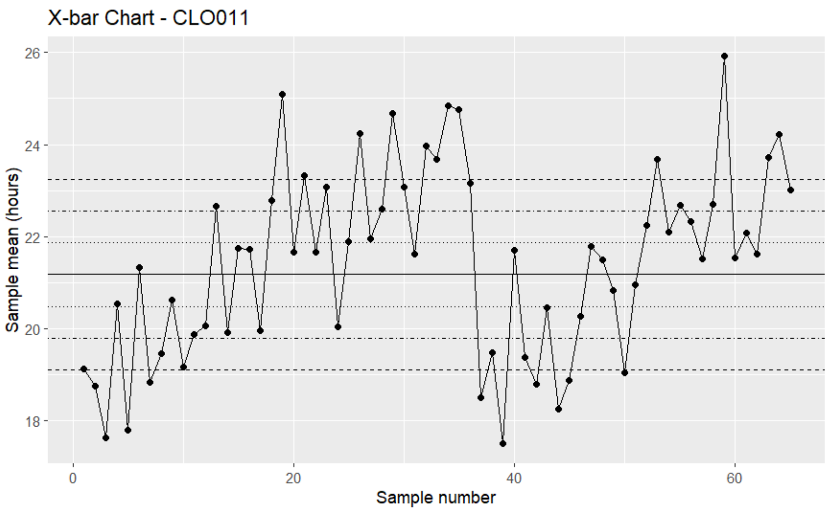

Figure 18: X-bar chart of product CLO011

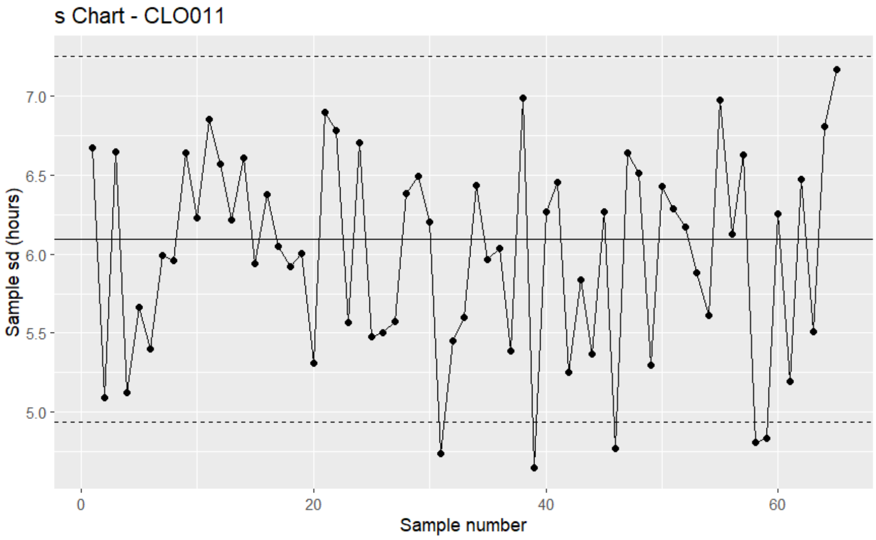

Figure 19: s-bar chart of product CLO011

The figures 18 and 19 represent the X-bar and s-chart for a particular product. The X-bar chart shows noticeable variation with approximately three points reaching beyond the sigma 3 (+/- 3 standard deviation) control limits. Many of the points dip near or below the lower control limits at first and then advance upwards and reach near or beyond the upper control limits. This pattern continues and potential occurrences in which the process mean delivery time is significantly higher or lower than expected, indicating periods of instability or potential special cases that affect the delivery times.

The s-chart for this product displays the sample standard deviation for each sample and shows that many of the points fluctuate significantly along the centre line. There are 5 points that fall below the lower control limit, indicating that the process variability was unusually low at those time periods which can be considered as an out-of-control signal. This should be investigated as true process control expects most points to be within the control limits.

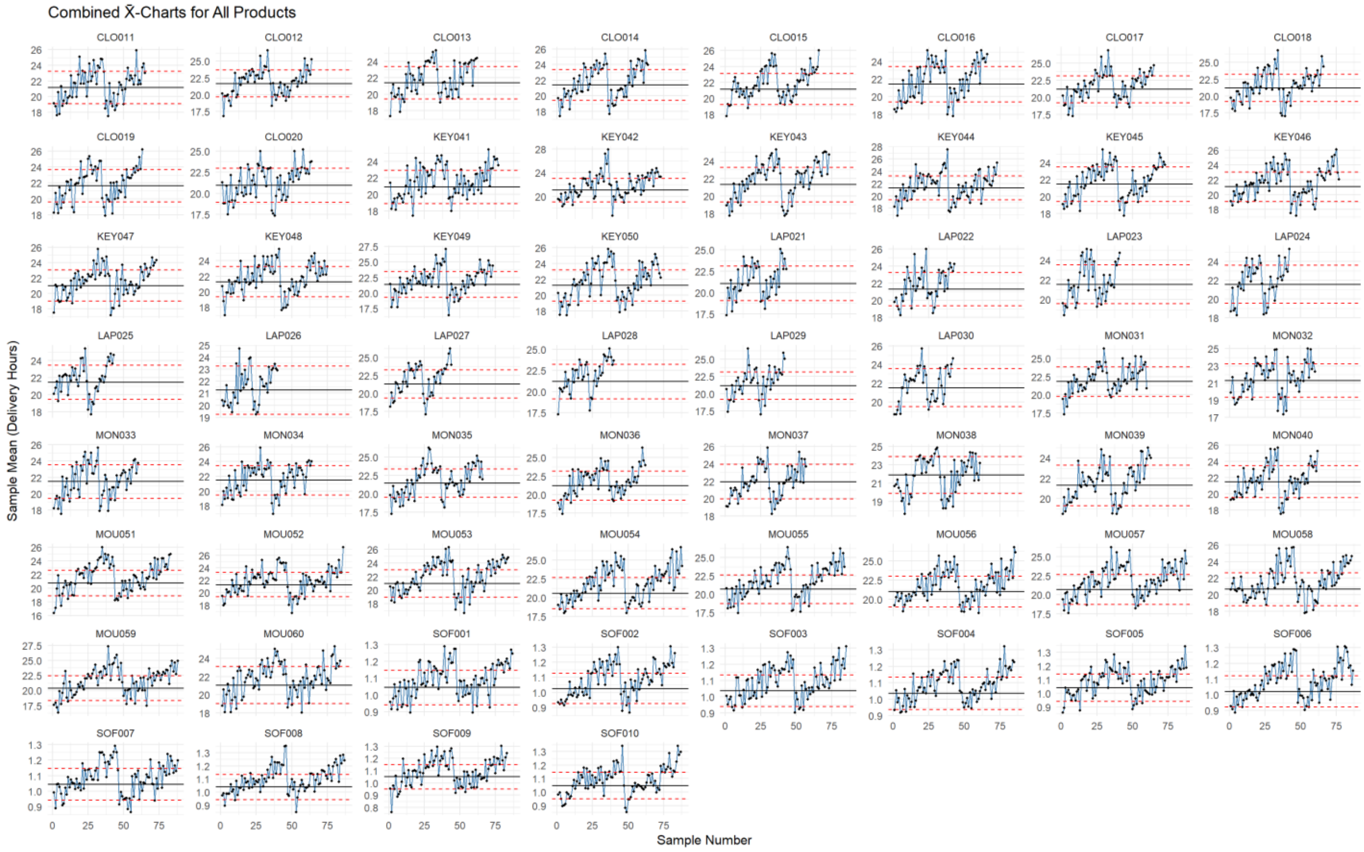

Figure 20: X-bar charts for all products

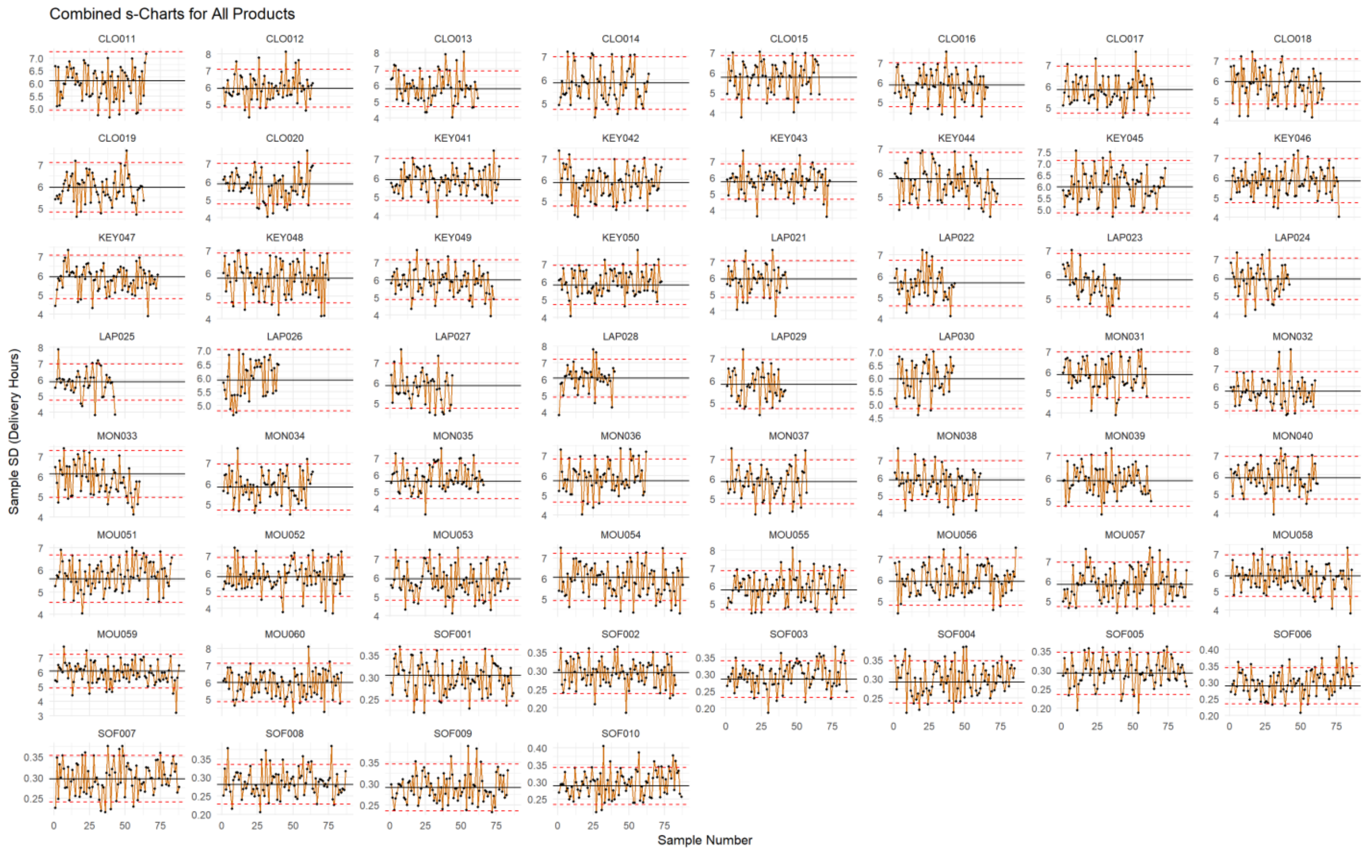

Figure 21: s-charts for all products

Figures 20 and 21 show the X-bar and s-charts of all the products.

3.2 Ongoing Process Monitoring

After calculating the control limits from the initial 30 samples, further samples of sample size 24 were sequentially extracted for each product to simulate ongoing process monitoring. The sample mean and standard deviation for each new sample was tested against the previously calculated control limits and violations, and in-control signals were tracked, in correspondence to what would be monitored in production.

3.3 Process Capability Analysis

The process capability of delivery times for each product was then assessed using the first 1000 deliveries and in alignment with the specification limits of LSL = 0 and USL = 32. The Capability indices Cp, Cpu, Cpl and Cpk were calculated per product in order to determine the extent to which particular processes consistently met the delivery requirements. Products with Cpk's greater than or equal to 1 were considered capable, whereas those with values lower than this were identified as potentially having issues with process performance. Table 3 displays a sample of products and their respective process capability indices. From this sample, the product with ID SOF001 has a Cpk 1.14973 and can thus be considered to have a delivery process capable of consistently meeting the delivery requirements, whereas the product with ID CLO001 has a Cpk of 0.57000 which fall below 1 and thus indicates that the delivery process for this product is incapable meeting the delivery requirements consistently due to high variation and an inability to frequently fall within the time specifications. After analysis of the entire table summarising the capability indices per product sample, it can be concluded that software products are the only 10 products (out of 60 products) that have a Cpk greater than 1. Products outside of this category exhibit Cpk values smaller than 1, indicating that 50 products are incapable of meeting the delivery requirements and further investigation into the delivery capabilities for these products is highly recommended.

Table 7: Summary of Capability Indices per product sample

product	n_values	n_samples	init_center	s_center	Cp	Cpu	Cpl	Cpk
CLO011	1581	65	21.178967	6.0957900	0.8501169	0.5699987	1.130235	0.5699987
CLO012	1546	64	21.645000	5.9449204	0.8646528	0.5573636	1.171942	0.5573636
CLO013	1499	62	21.386789	5.7762635	0.8587481	0.5653051	1.152191	0.5653051
CLO014	1543	64	21.388056	5.8721218	0.8776087	0.5849704	1.170247	0.5849704
CLO015	1606	66	21.151067	5.7733758	0.8861403	0.5794999	1.192781	0.5794999
SOF001	2089	87	1.043693	0.3046912	17.2014995	33.2532669	1.149732	1.1497321
SOF002	2014	83	1.025637	0.2939940	17.3031630	33.4549897	1.151336	1.1513362
SOF003	2059	85	1.037026	0.2846738	18.0499544	34.8934667	1.206442	1.2064420
SOF004	2046	85	1.036297	0.2919018	17.5268865	33.8818541	1.171919	1.1719188
SOF005	2115	88	1.041818	0.2916466	17.2951168	33.4247319	1.165502	1.1655017

3.4 Identification of Process Control Issues

Table 8: Violation summary per product

	product	total_s_violations	longest_s_in_1sigma_run	total_x_4seq	Cp	Cpu	Cpl	Cpk
1	CLO011	0	5	1	0.8501169	0.5699987	1.130235	0.5699987
2	CLO012	5	5	2	0.8646528	0.5573636	1.171942	0.5573636
3	CLO013	11	4	1	0.8587481	0.5653051	1.152191	0.5653051
4	CLO014	7	5	2	0.8776087	0.5849704	1.170247	0.5849704
5	CLO015	4	3	2	0.8861403	0.5794999	1.192781	0.5794999
6	CLO016	1	3	2	0.8562397	0.5602975	1.152182	0.5602975
7	CLO017	3	3	2	0.8782082	0.5803221	1.176094	0.5803221
8	CLO018	3	3	1	0.8464236	0.5725528	1.120294	0.5725528
9	CLO019	2	4	2	0.8694731	0.5684852	1.170461	0.5684852
10	CLO020	3	6	2	0.8951881	0.6213139	1.169062	0.6213139

A. A total of 263 samples for all the product types have s values that outside of the upper +3 sigma-control limits.

Table 9: First 3 and last 3 samples with s values outside +3 sigma-control limits

	product	sample	sample_sd	flag_s_above_UCL3
1	CLO012	10	7.5277265	TRUE
2	CLO012	26	7.7743855	TRUE
3	CLO012	45	8.1253484	TRUE
261	SOF010	78	0.3451462	TRUE
262	SOF010	81	0.3779174	TRUE
263	SOF010	83	0.3614333	TRUE

B. Product SOF008 has the most consecutive samples of s between the -1 and +1 sigma-control limits across all the product types with 9 consecutive samples of s staying between the specified control limits.

Table 10: Sample table of products with longest consecutive samples of s between -1 and +1 sigma

	product	longest_in_control_run	start_sample	end_sample
1	CLO011	5	15	19
2	CLO012	5	37	41
3	CLO013	4	54	57
4	CLO014	5	17	21
5	CLO015	3	18	20
55	SOF005	4	34	37
56	SOF006	6	16	21
57	SOF007	5	78	82
58	SOF008	9	68	76
59	SOF009	5	5	9

C. A total of 873 X-bar samples display 4 consecutive points outside of the upper, second control limits for all product types.

Table 11: X-bar samples with 4 consecutive violations of the upper, second sigma control limit

	product	sample	len
1	CLO011	32	5
2	CLO011	33	5
3	CLO011	34	5
871	SOF010	85	6
872	SOF010	86	6
873	SOF010	87	6

Section 4

4.1 Estimation of the likelihood of making a Type I error

Table 12: Type I error probabilities

Type I (Manufacturer's) Error probabilities for SPC Rules A–C

Rule	Probability
A: One sample outside ±3σ limits	0.0026998
B: One sample within ±1σ limits (good control)	0.6826895
C: Four consecutive samples beyond ±2σ limits	0.0000005

Table 12 presents the probabilities for Type I errors to occurs, where the process may signal an unstable condition despite it being stable in reality. These error probabilities control the frequency of unnecessary adjustments that may disrupt the production process and increase costs. The probabilities for A and C are small and indicate that the likelihood of a sample either being outside of the +-3 sigma limits or having four consecutive samples beyond the +-2 sigma limits is extremely low. Additionally, Rule B shows a relatively promising probability of samples staying within the +-1 sigma limits, which suggests good control of the process.

4.2 Estimation of likelihood of making a Type II error for Bottle Filling Process

Metric	Probability
Type II Error (Consumer's) Probability	0.8411783

Figure 22: Likelihood of Type II error for Bottle Filling Process

The calculated probability of making a Type II error is 0.8411783, suggesting that the current test would fail to detect actual process shifts approximately 84% of the time. This high probability of failure indicates that the existing sampling plan is not sensitive enough for detecting slight changes in the mean values and thus further research should be done to redesign the current control-chart rule. This can be done by either increasing sample size or tightening control limits to ensure timely detection of any process variations.

4.3.1 Fixing products_data and products_Headoffice

Table 13: Sample of fixed products_data

	ProductID	Category	Description	SellingPrice	Markup
1	SOF001	Software	coral matt	511.53	25.05
2	SOF002	Software	cyan silk	505.26	10.43
3	SOF003	Software	burlywood marble	493.69	16.18
4	SOF004	Software	blue silk	542.56	17.19
5	SOF005	Software	aliceblue wood	516.15	11.01
6	SOF006	Software	black silk	478.93	16.99
7	SOF007	Software	black bright	527.56	16.79
8	SOF008	Software	burlywood silk	549.02	11.95
9	SOF009	Software	azure sandpaper	540.41	11.34
10	SOF010	Software	chocolate sandpaper	396.72	23.47

The sample of the adjusted products_data dataset confirms that the previously mismatched products IDs and categories have been resolved.

Table 14: Sample of fixed products_Headoffice

	ProductID	Category	Description	SellingPrice	Markup
55	SOF055	Software	coral sandpaper	516.15	11.01
56	SOF056	Software	black marble	478.93	16.99
57	SOF057	Software	blueviolet marble	527.56	16.79
58	SOF058	Software	black marble	549.02	11.95
59	SOF059	Software	cornflowerblue marble	540.41	11.34
60	SOF060	Software	chocolate sandpaper	396.72	23.47
61	CLO001	Cloud Subscription	blue bright	1070.54	16.41
62	CLO002	Cloud Subscription	chocolate marble	963.14	10.13
63	CLO003	Cloud Subscription	blue sandpaper	1067.54	16.80
64	CLO004	Cloud Subscription	chocolate marble	1083.11	21.25
65	CLO005	Cloud Subscription	chocolate marble	728.26	27.70

The sample of the modified products_Headoffice displays the corrected selling prices and markup values corresponding to those in the products_data dataset with each selling price and markup repeating every 10 products per products category.

4.3.2 Re-analysis and results of fixing data

Summary Statistics

The errors within the products_data file were corrected by matching the product id to the corresponding category. However, this error did not have any initial impact on the selling price and markup values within the dataset and thus the summary statistics for products_data stayed the same as in the initial analysis.

Table 15: Revised Summary Statistics for products_Headoffice

Variable	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
ProductID*	1	360	180.50000	104.067286	180.500	180.50000	133.434000	1.00	360.00	359.00	0.0000000	-1.2100045	5.4848276
Category*	2	360	3.50000	1.710202	3.500	3.50000	2.223900	1.00	6.00	5.00	0.0000000	-1.2781771	0.0901356
Description*	3	360	30.68611	17.319505	29.500	30.76736	22.980300	1.00	60.00	59.00	-0.0277818	-1.3900365	0.9128181
SellingPrice	4	360	4493.59283	6458.320465	794.185	3189.25479	525.722547	350.45	19725.18	19374.73	1.4564972	0.5314909	340.3833755
Markup	5	360	20.46167	6.030161	20.335	20.51187	7.309218	10.13	29.84	19.71	-0.0374882	-1.1879762	0.3178174

Variable	Q1.25%	Q3.75%	IQR
SellingPrice	512.1825	6416.6600	5904.4775
Markup	16.1400	25.7075	9.5675

The adjustments made to the products_Headoffice data resulted in significant changes to the initial summary statistics for this dataset. For instance, the mean value for Selling Price shifted from 4410.96 to 4493.59 and the markup mean value changed from the original 20.386 to 20.46. The standard deviation values for selling prices and markup also changed from 6463.82 and 5.666 to 6458.32 and 6.030 respectively. These adjusted values also resulted in distribution shape changes and selling prices for this dataset can now be expected to have a flatter distribution (selling prices are more evenly distributed) and will be slightly less right-skewed. The markup distribution can be expected to now have a flatter peak with thinner tails. The interquartile ranges for selling price and markup have also changed from 5347.40 and 9.01 to 5904.48 and 9.57 respectively.

Table 16: Sample of 2023 Sales Combined with products_Headoffice

	CustomerID	ProductID	Quantity	orderTime	orderDay	orderMonth	orderYear	pickingHours	deliveryHours	Category	Description	SellingPrice	Markup	SalesValue
1	CUST3172	LAP026	17	17	14	7	2023	38.3908333	31.5460	Laptop	chocolate bright	18711.72	13.51	318099.24
2	CUST3721	LAP024	31	12	18	7	2023	41.3908333	24.5460	Laptop	burlywood sandpaper	18366.92	29.35	569374.52
3	CUST582	MON032	1	19	9	6	2023	17.0575000	22.0460	Monitor	blue silk	6634.13	27.80	6634.13
4	CUST3343	MON040	10	19	13	12	2023	24.0575000	24.0460	Monitor	cornflowerblue bright	5346.14	29.74	53461.40
5	CUST1628	CLO015	5	10	9	8	2023	13.7241667	14.0460	Cloud Subscription	azure silk	728.26	27.70	3641.30
6	CUST4713	KEY043	6	9	30	9	2023	15.0575000	30.5460	Keyboard	blue silk	516.41	22.83	3098.46
7	CUST3847	CLO015	1	15	28	7	2023	11.3908333	30.5460	Cloud Subscription	azure silk	728.26	27.70	728.26
8	CUST4460	MON038	6	12	16	9	2023	23.0575000	17.5460	Monitor	black matt	6478.10	17.46	38868.60
9	CUST1785	KEY046	6	14	5	8	2023	11.7241667	33.0460	Keyboard	black sandpaper	708.18	17.72	4249.08
10	CUST2641	SOF003	1	11	26	5	2023	0.8482778	0.9773	Software	burlywood marble	493.69	16.18	493.69

After adjusting the products_Headoffice dataset, it was then joined to the sales for 2023 only in order to investigate the total sales value by multiplying the quantity sold of a product sold with its corresponding selling price, resulting in the added last column "SalesValue". This was done in order to further investigate revenue generated by all the products. This process was repeated for the sales made in 2022.

Visualizations

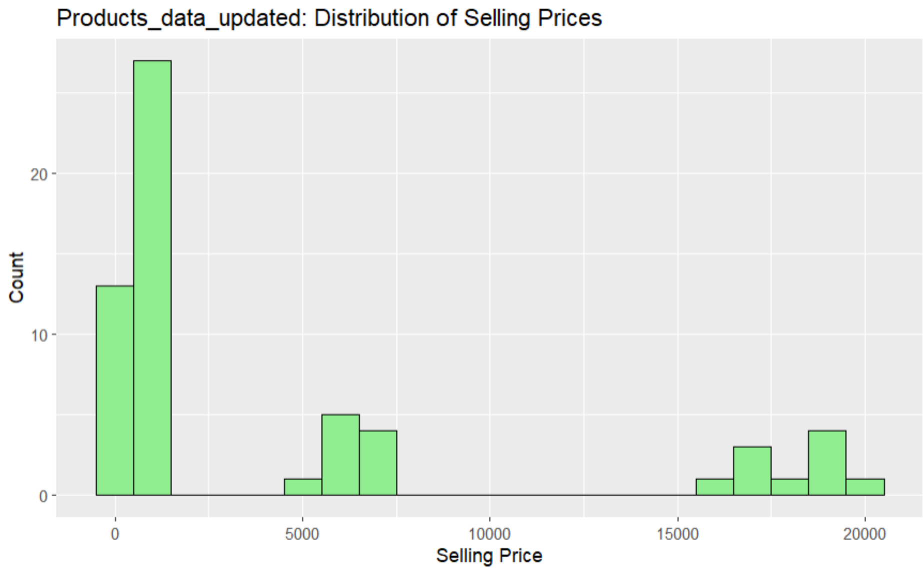

Figure 23: Histogram of Selling Price for Products_data_updated

The distribution of selling prices for the products_data dataset did not change as there were no changes to the selling price and markup values within the dataset. The only modifications made to this dataset was the correct matching of product category with the product ID.

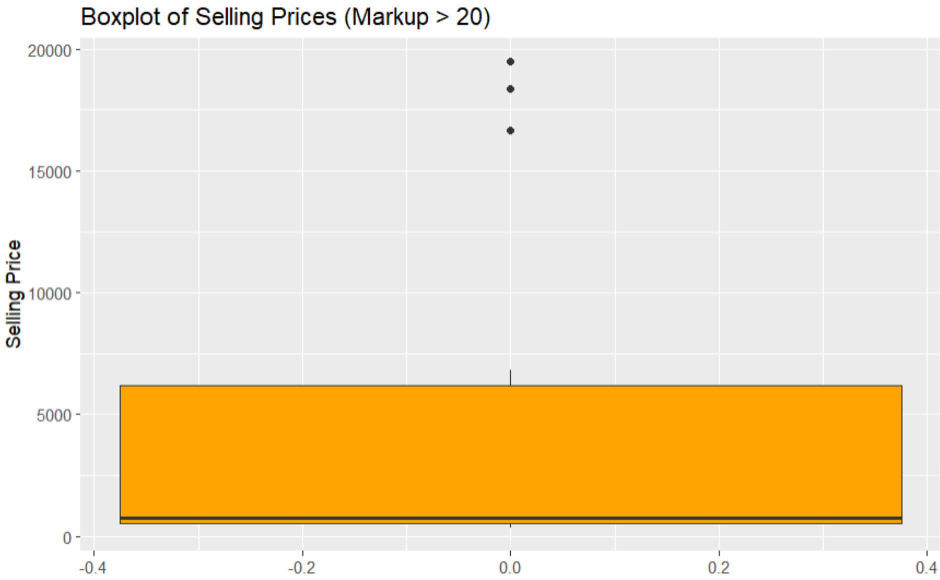

Figure 24: Boxplot distribution Selling Prices (markup > 20) for products_Headoffice2025

As products_Headoffice had significant changes to the selling price and markup value, the investigation of the distribution of selling prices for products with a markup of greater than 20% was now conducted using the updated products_Headoffice dataset rather the adjusted products_data dataset. This distribution shows only three outlying products that exhibit selling prices greater approximately 16 000. The third quartile value is now presented as being approximately 6250 and the mean value of selling prices is just below 1000, suggesting that a large portion of products have a low selling price and a slightly smaller range of products have higher selling prices of greater than 2000 but smaller than 7000.

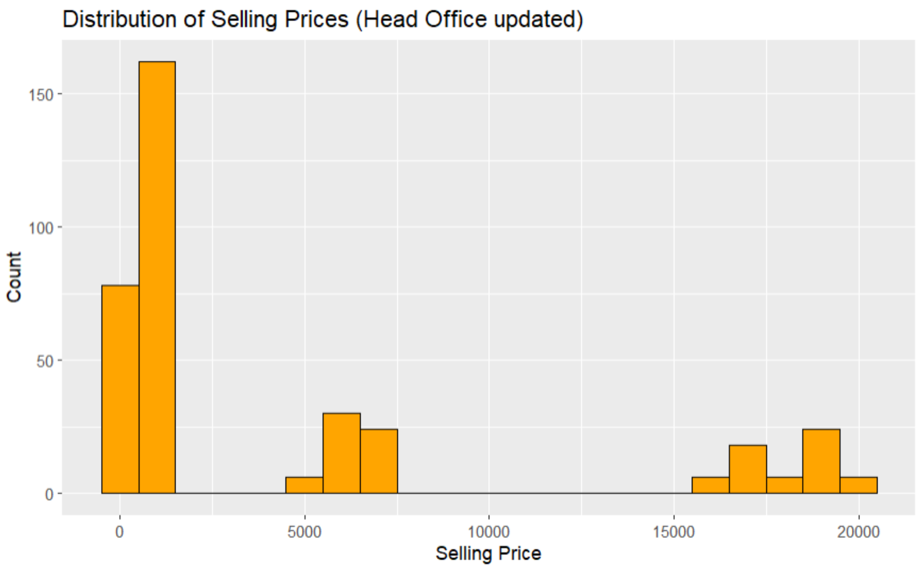

Figure 25: Histogram of Selling Prices for products_Headoffice2025

The shape of the selling price distribution for products_Headoffice is now identical to that of the products_data histogram distribution. This is due to the adjustment of matching the selling prices and markup values within the dataset to those in the products_data dataset and repeating these values for every tenth product for each category. However, the frequency of products differs between the respective histograms as products_Headoffice contains 360 products whilst products_data only displays 60 products of these products (60*6 = 360).

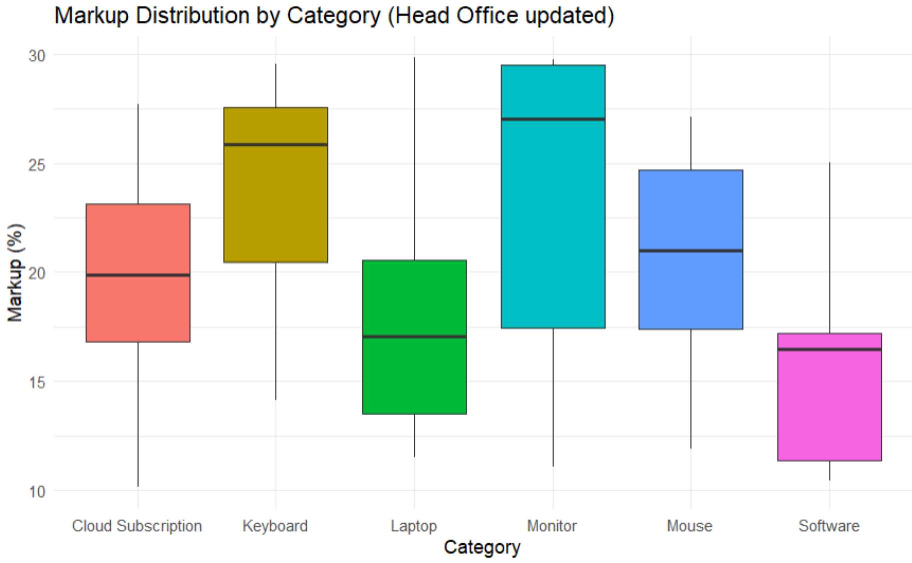

Figure 26: Boxplot Markup distribution by Category for products_Headoffice2025

The boxplots of the markup values for each category now present varying average markup values and distributions in comparison to that shown by Figure 11. The mean markup values per category are now displayed as follows: Cloud Subscription ~ 20%, Keyboard ~ 26%, Laptop ~ 17%, Monitor ~ 27%, Mouse ~ 21% and Software ~ 16.5%. It can now be deduced that monitors tend to have higher markup values but additionally, they also have a larger interquartile range in comparison to other products, suggesting a higher variation in markup values. Cloud subscriptions, laptops and mouses present more evenly spread distributions of markup values whilst keyboards, monitors and software products can be expected to have more left-skewed distributions with mean markup values tending more closely towards their respective third quartile values. A large majority of software products in particular exhibit higher markup values of approximately 16.5% within the interquartile range from around 11.5% to just below 17.5%

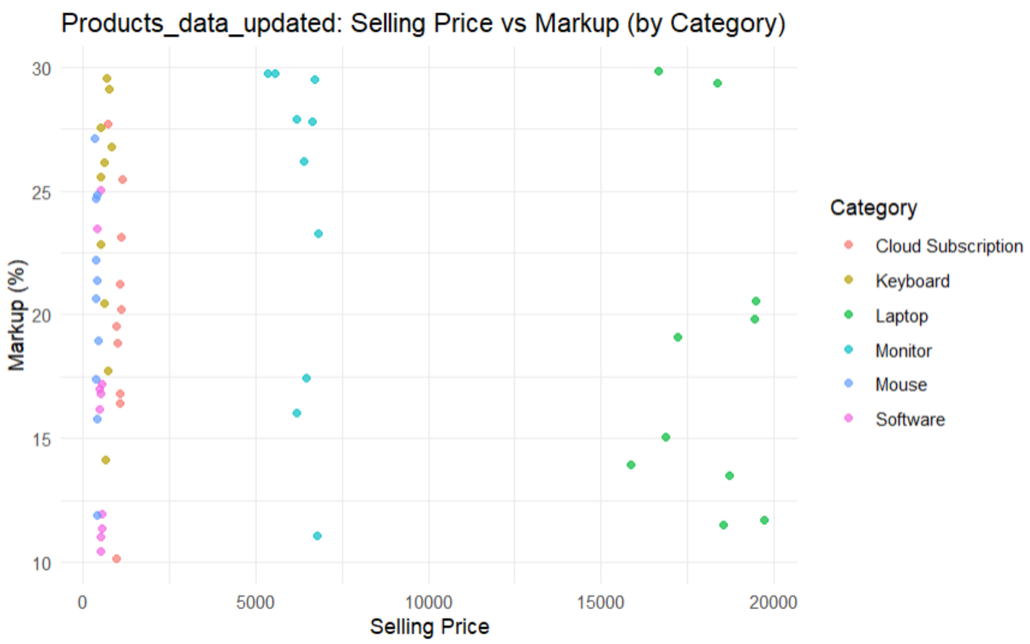

Figure 27: Scatter plot of Selling Price vs Markup (by Category) for products_data_updated

Figure 27 illustrates the revised scatter plot of the selling prices against the markup, differentiated by category (colour). In comparison Figure 8, there is now a distinct relationship between the selling prices, markup values and categories. Cloud subscriptions, keyboards, mouses and software display lower selling prices between approximately 200 and 1000 in comparison to monitors and laptops, suggesting that these products are of lower value. Monitors are shown to have a selling price range between 5100 and 7000 and laptops have a selling price range from approximately 16000 to just below 20000, indicating that laptops are the highest valued items amongst all the product categories.

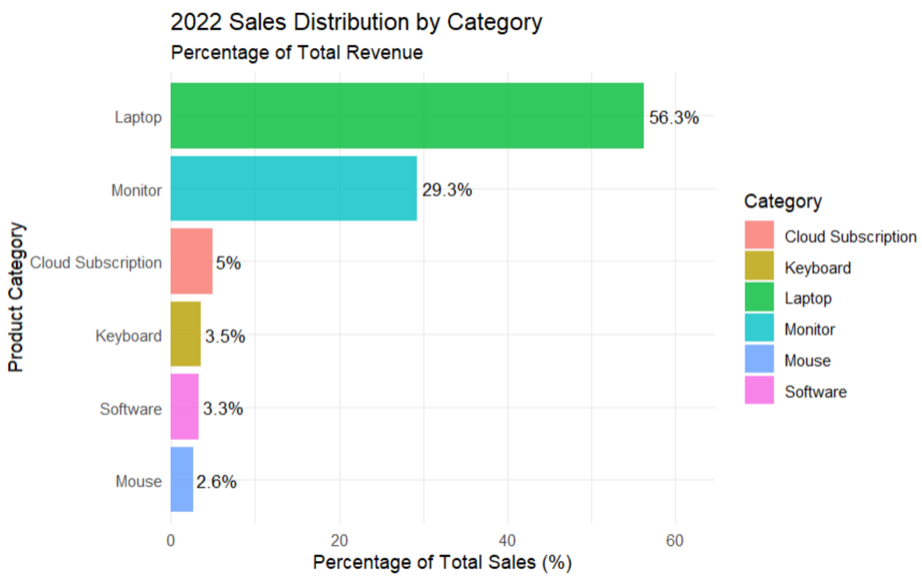

Figure 28: 2022 Sales (percentage) Distribution by Category

The 2022 sales distribution by category bar plot illustrates that a majority of revenue earned is made from the selling of laptops as they contributed a 56.3% to the total revenue. Monitors contributed the second largest amount of 29.3% to the total revenue earned. This could be indicative of a strong positive correlation between the selling prices of items and the total amount of revenue they generate (i.e. higher value items generate the most revenue). In correspondence to this, mouses only contributed 2.6% to the total revenue likely due to their lower selling prices.

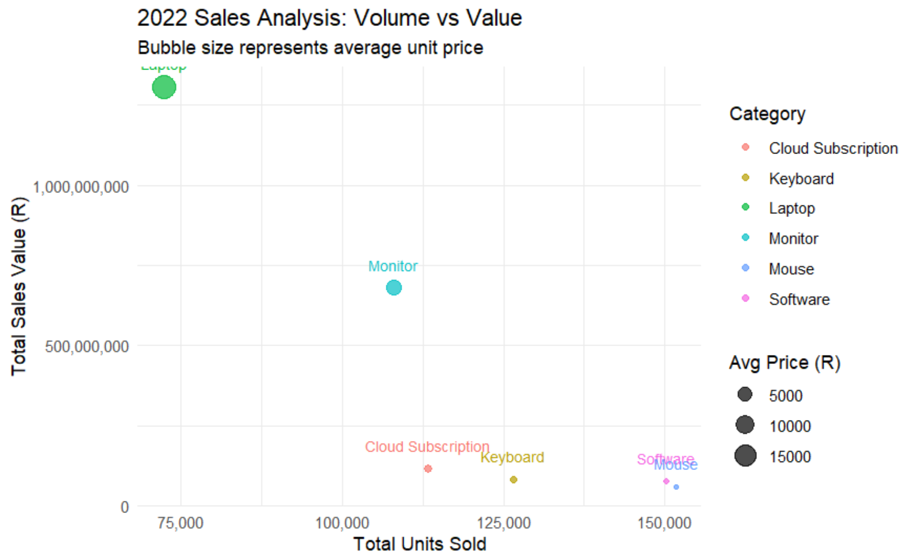

Figure 29: 2022 Sales Analysis of Volume vs Price Value

The analysis of sales volume from 2022 against the total sales value shows a negative correlation of higher value items having lower total units sold. This is presented by laptops which has the highest selling price, but lowest total number of units sold of just under 75000, whereas mouses have the lowest average selling prices but the highest total units sold of over 150 000. However, laptops are still shown to contribute the highest total sales value to the total revenue whilst mouses contribute the lowest.

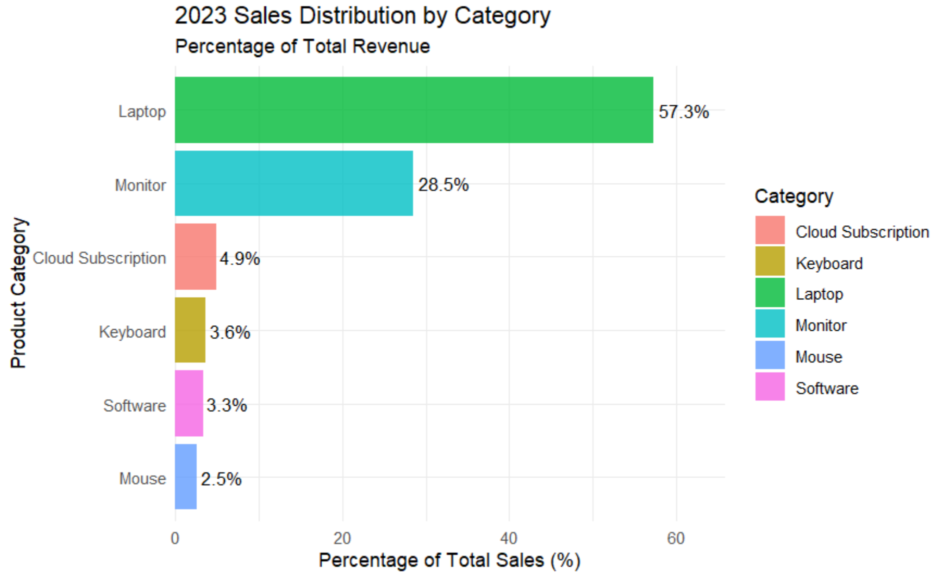

Figure 30: 2023 Sales (percentage) Distribution by Category

The contribution to the total revenue per product category did not show any significant changes in 2023. Whilst laptops contributed 1% more to the total revenue in 2023, monitors contributed 1% less to the total revenue. The contribution made by software did not shift, but cloud subscriptions and mouses contributed 0.1% less and keyboards contributed 0.1% more.

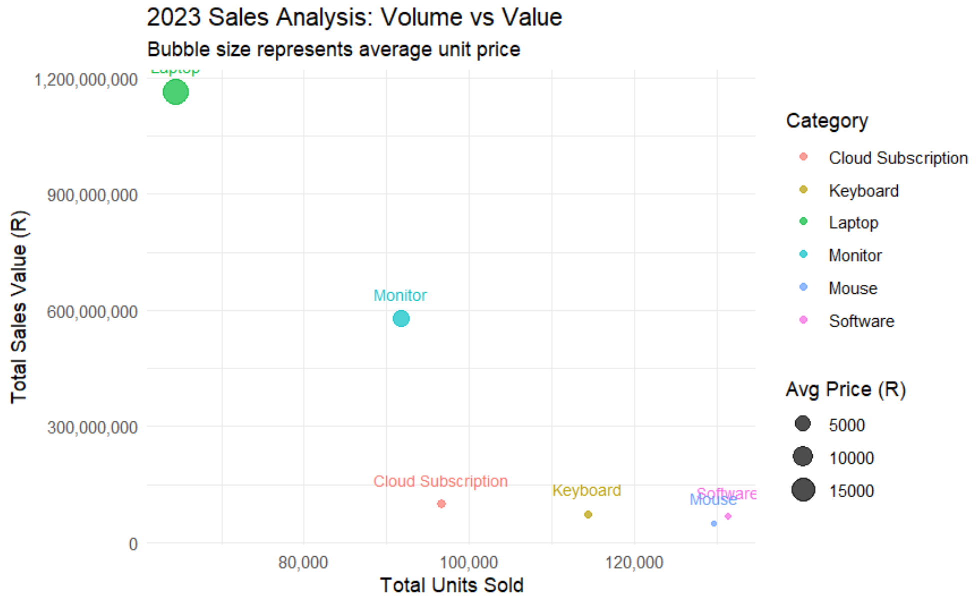

Figure 31: 2023 Sales Analysis of Volume vs Price Value

A significantly lower number of all products were sold in 2023. For instance, only 70 000 laptops were sold, whereas almost 75000 laptops were sold in the previous year. The volume of mouses sold was also overtaken by software, which only displayed had a sales volume of just over 130 000. This may indicate the need for more promotions and campaigns for all products in order to potentially increase sales volumes.

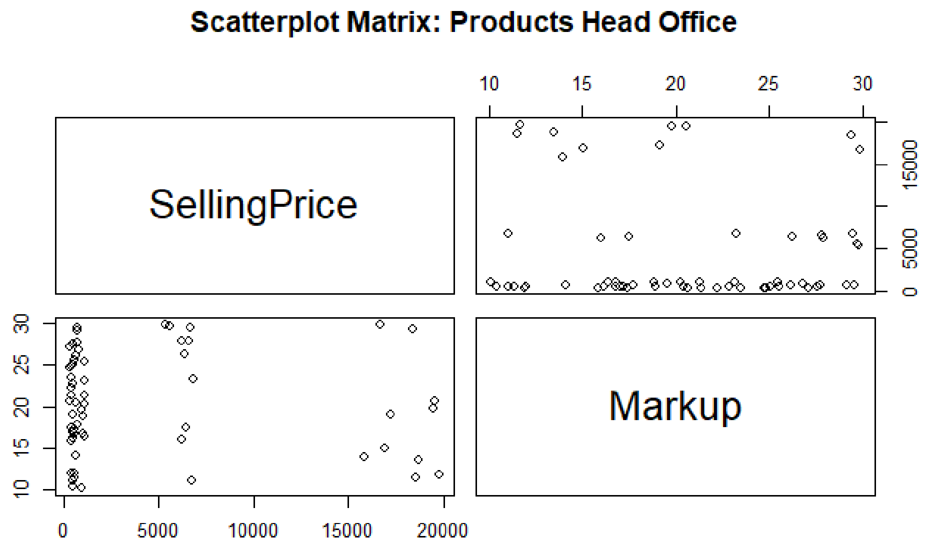

Figure 32: Scatterplot Matrix of products_Headoffice2025

The scatter plot matrix for products_Headoffice displays the relationship between the selling prices and markup values, which is identical to that of Figure 27, owing to the particular adjustments made to the products_Headoffice dataset, as discussed previously.

Section 5

5.1 Shop 1 analysis

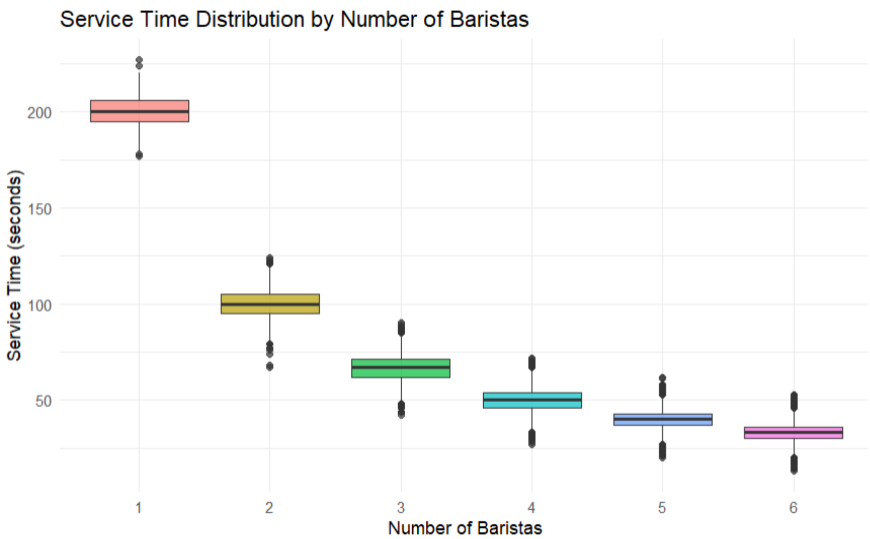

Figure 33: Service Time Distribution by Number of Baristas

Figure 33 displays the service time distributions per number of baristas which shows that as the number of baristas increases from one to six, the variation in service time decreases and the mean service time decreases along an exponential curve. This indicates improved service consistency with more baristas; however, the average service times taper off towards more than 4 baristas.

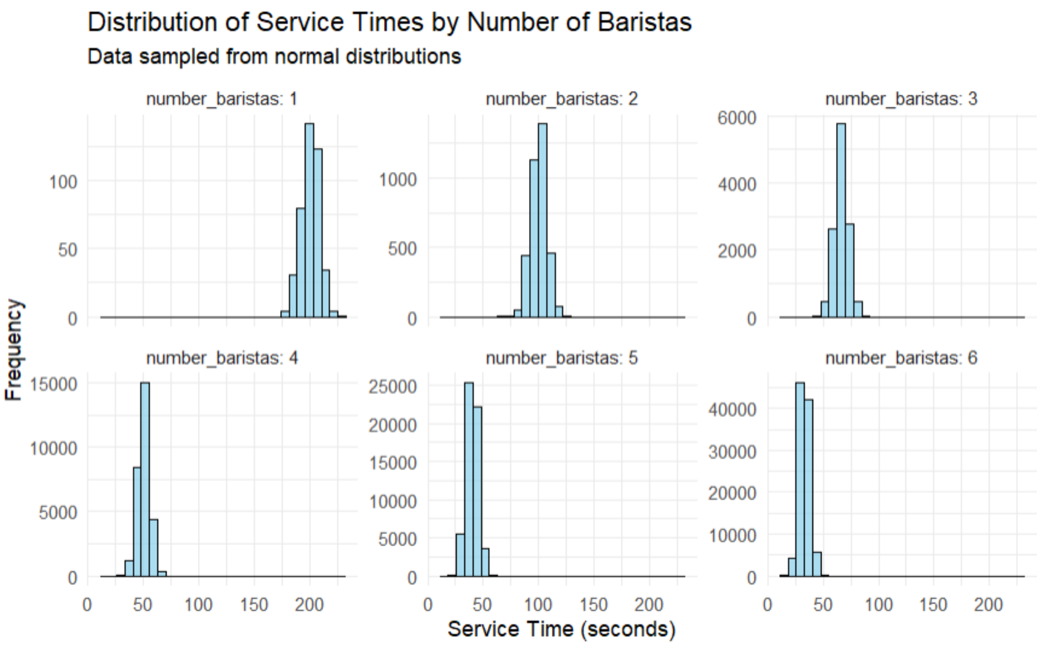

Figure 34: Histogram Distribution of Service Times by Number of Baristas

Figure 34 once again illustrates that service times tend to be longer and have a broader variation for less than 3 baristas. This suggests that peak demand times would benefit from increasing the staffing level

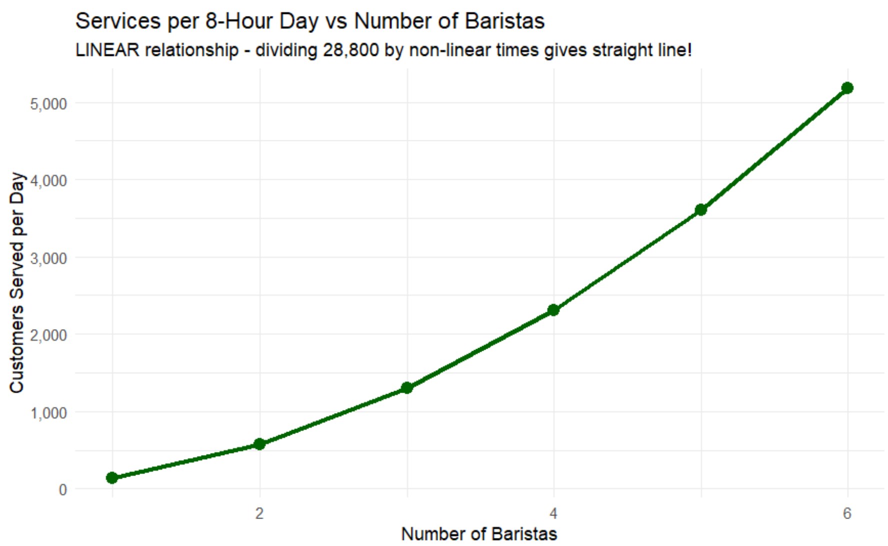

Figure 35: Number of Customers Served per 8_Hour Day vs Number of Baristas

The expected number of customers served per 8-hour day per number of baristas was calculated as 28800 divided by the mean service time and then multiplied by the corresponding number of baristas. Figure 35 displays an increase in the number of customers, reaching up to just over 5000 customers, as the number of baristas increases.

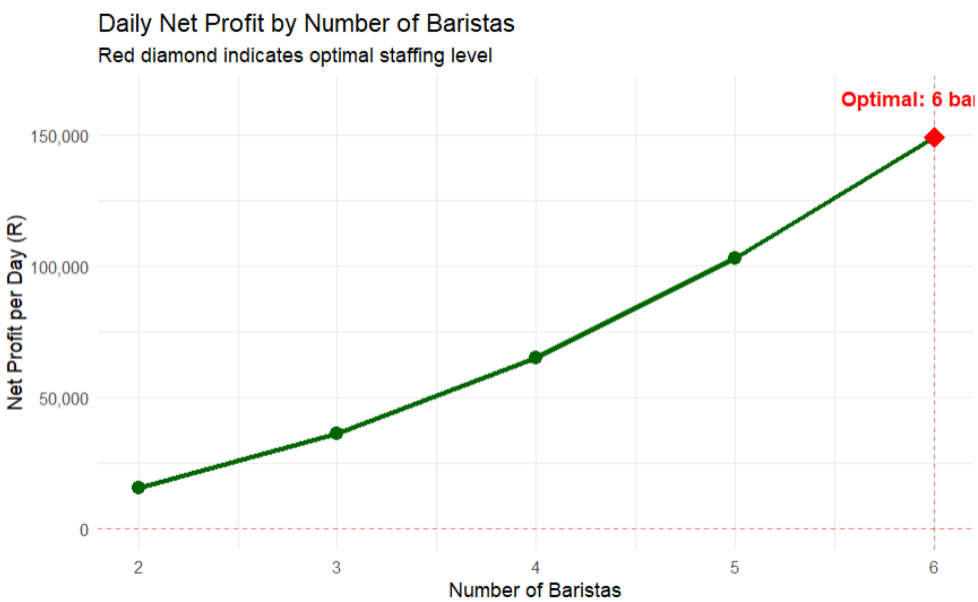

Figure 36: Daily Net Profit by Number of Baristas

The net profit per day was calculated using the gross revenue that could be generated based on the number of customers per day and subtracting the cost of hiring staff. Figure 36 shows the maximum daily net profit that could be generated as being approximately R150000, which can only be attained by hiring 6 baristas.

Table 17: Profit analysis based on the number of baristas employed

number_baristas <dbl>	CustomersServed <dbl>	GrossRevenue <dbl>	PersonnelCost <dbl>	NetProfitPerDay <dbl>	AnnualProfit <dbl>
2	575.0168	17250.51	2000	15250.51	5566434
3	1297.0686	38912.06	3000	35912.06	13107901
4	2304.9045	69147.14	4000	65147.14	23778704
5	3603.4381	108103.14	5000	103103.14	37632648
6	5180.5322	155415.97	6000	149415.97	54536828

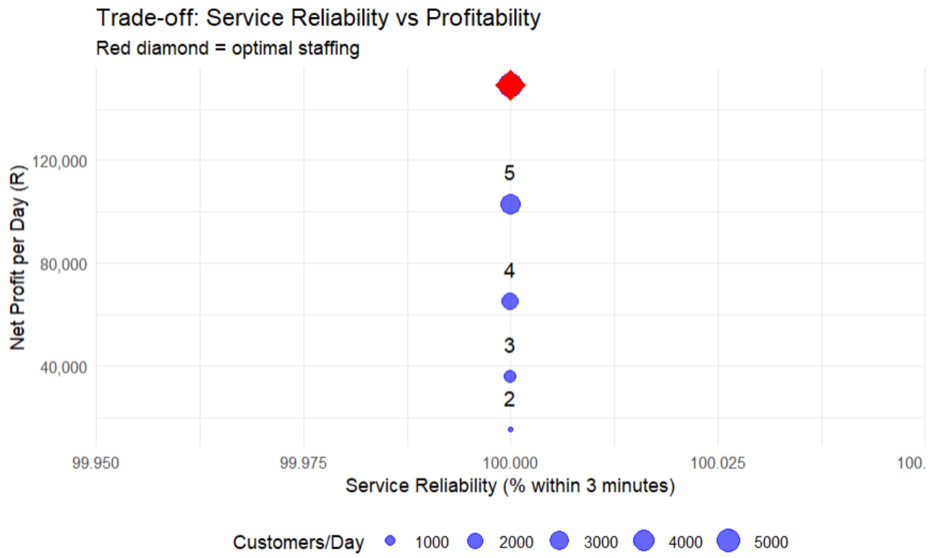

Figure 37: Service Reliability vs Profitability

Figure 37 displays that the service reliability, using a three-minute standard, is approximately 100% for all levels of staffing, however, hiring six baristas still displays the highest daily net profitability and so overall, the total number of baristas that should be hired is six baristas.

5.2 Shop 2 analysis

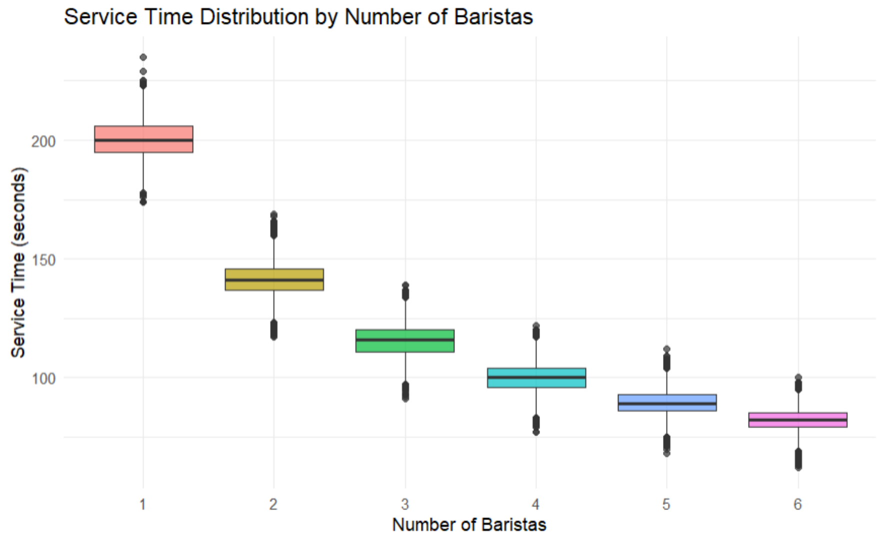

Figure 38: Service Time Distribution by Number of Baristas (Shop 2)

The service time distribution by number of baristas is similar to that of shop 1 but with a shallower decrease from one to three baristas. For example, with 2 baristas employed, the mean service time for shop 2 is approximately 140 seconds, whereas for shop 1 it was approximately 100 seconds.

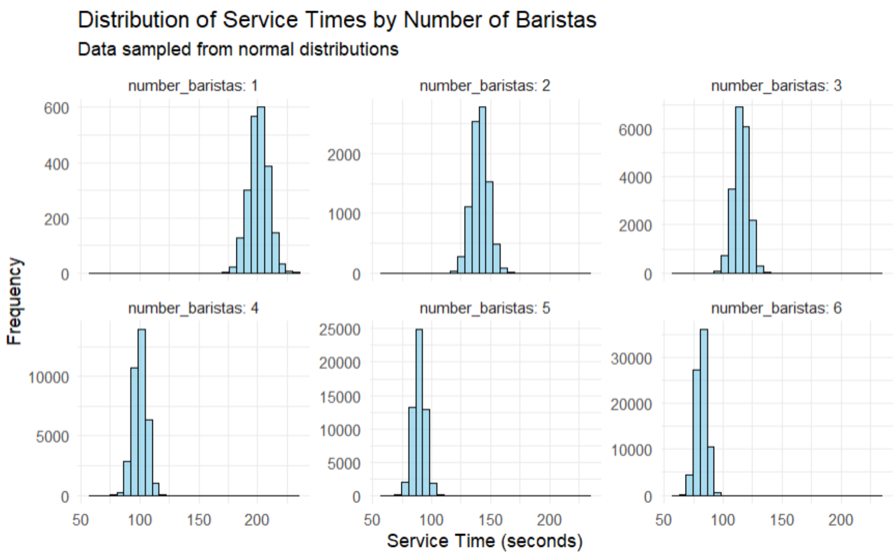

Figure 39: Histogram Distribution of Service Times by Number of Baristas (Shop 2)

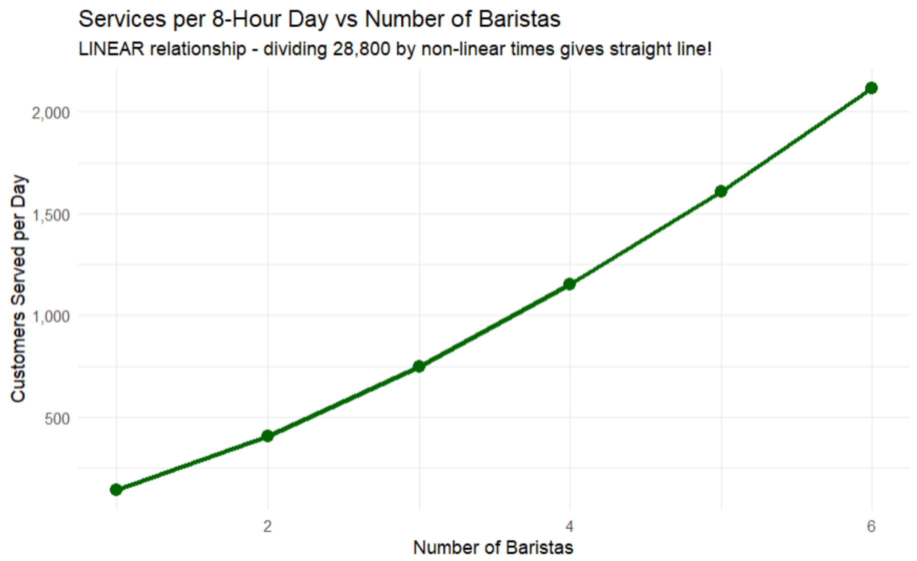

Figure 40: Number of Customers Served per 8_Hour Day vs Number of Baristas (Shop 2)

As with shop 1, the expected number of customers served per 8-hour day per number of baristas was calculated as 28800 divided by the mean service time and then multiplied by the corresponding number of baristas. However, compared to shop 1, the maximum number of customers that could be served in a day is now relatively greater than 2000, which is approximately 3000 less than that of shop 1.

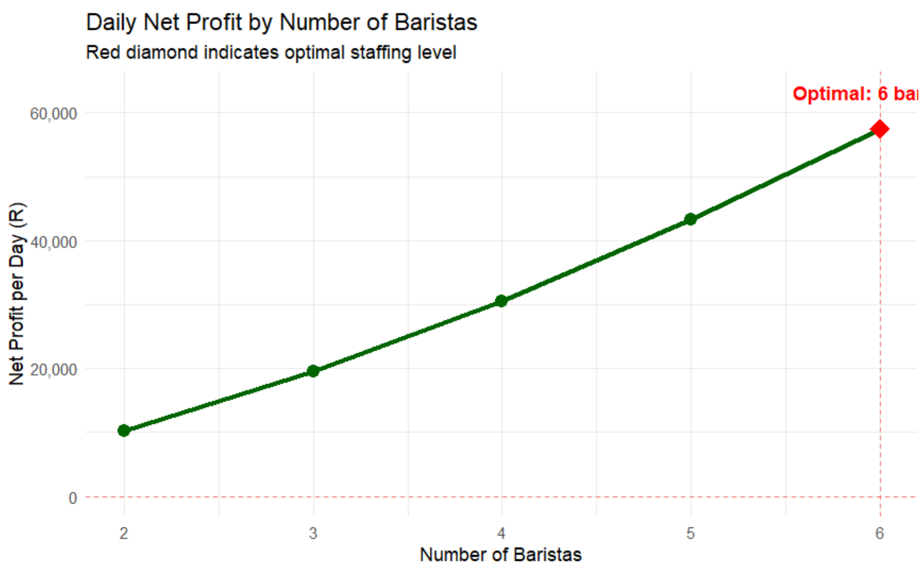

Figure 41: Daily Net Profit by Number of Baristas (Shop 2)

Similar to shop 1, the optimal number of baristas to hire is six, as this number generates the greatest daily net profit. The corresponding daily net profit for shop 2, however, is just under R60000, which is much less than that of shop 1. This is likely due to the lower number of customers per day.

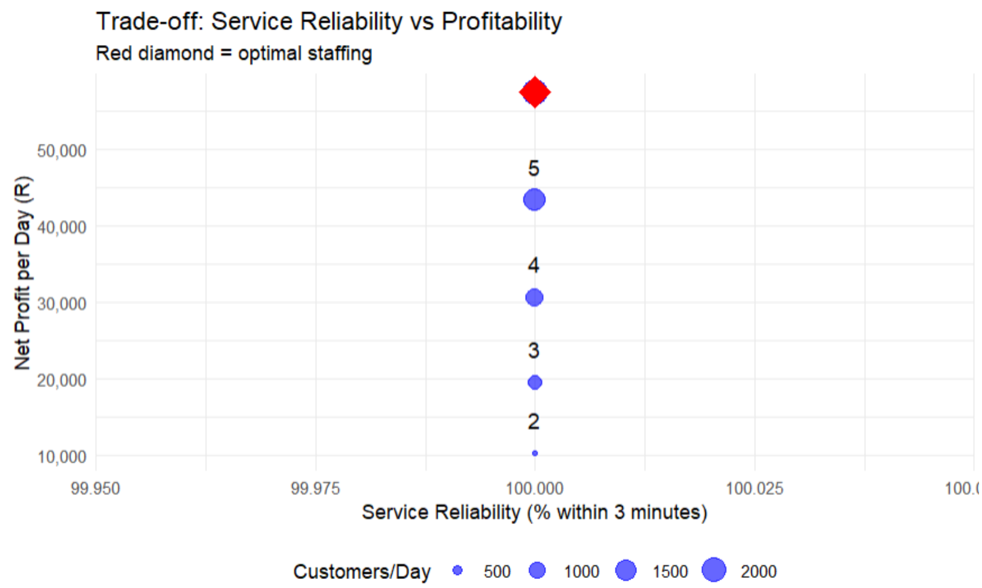

Figure 42: Service Reliability vs Profitability (Shop 2)

Section 6

6.2 Testing hypotheses using ANOVA

Based on the results from Section 3, all products classified under the software category (SOF001 – SOF010) were shown to have Cpk values that were greater than 1 and thus identified as being capable. The three products with the highest Cpk values were then chosen in order to represent the most stable and capable delivery processes, which allowed for clearer interpretation of the differences between the years and months. The products with the highest Cpk values were SOF003, SOF008 and SOF010. Using these three products, the following three null hypotheses were tested using ANOVA:

- H0_year: There is no significant difference in delivery times between 2022 and 2023.
- H0_month: There is no significant difference in delivery times across months.
- H0_interaction: There are no interaction effects between factors.

Table 18: Summary Values of Top 3 Cpk Products by year

	.product	year_factor	n	mean_delivery	sd_delivery
1	SOF003	2022	1144	1.088782	0.3009322
2	SOF003	2023	915	1.100743	0.3203670
3	SOF008	2022	1116	1.093710	0.2974502
4	SOF008	2023	977	1.102889	0.2986806
5	SOF010	2022	1113	1.084635	0.3008854
6	SOF010	2023	992	1.091489	0.3160956

The yearly main summary statistics for the selected products are provided by Table 18. The differences in the mean delivery capability are examined by the ANOVA in order to assess the statistical significance.

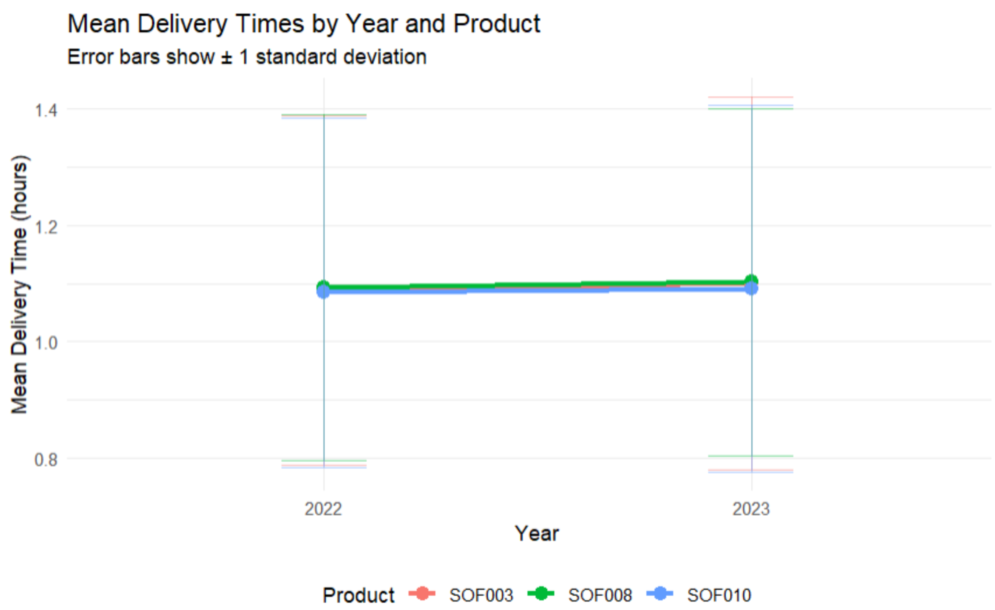

Figure 43: Mean Delivery Times by Year and Product

The mean delivery times by year and product illustrates very little deviation in the mean delivery times across the products. The figure also displays a slight increase from year 2022 to 2023. Visually, this slight increase does not indicate great significance between the two years.

```
ANOVA Results (Year Effect):
              Df Sum Sq Mean Sq F value Pr(>F)
year_factor    1   0.02 0.02464   0.259  0.611
Residuals   2103 199.69 0.09495

Levene's Test for Homogeneity of Variance:
Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    1  3.0885  0.079 .
      2103
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 44: ANOVA Results (Year Effect)

The p-value obtained from the ANOVA is equal to 0.611, which is greater than 0.05, indicating that we fail to reject the null hypothesis: There is no significant difference in delivery times between 2022 and 2023. This is supported visually by Figure 43.

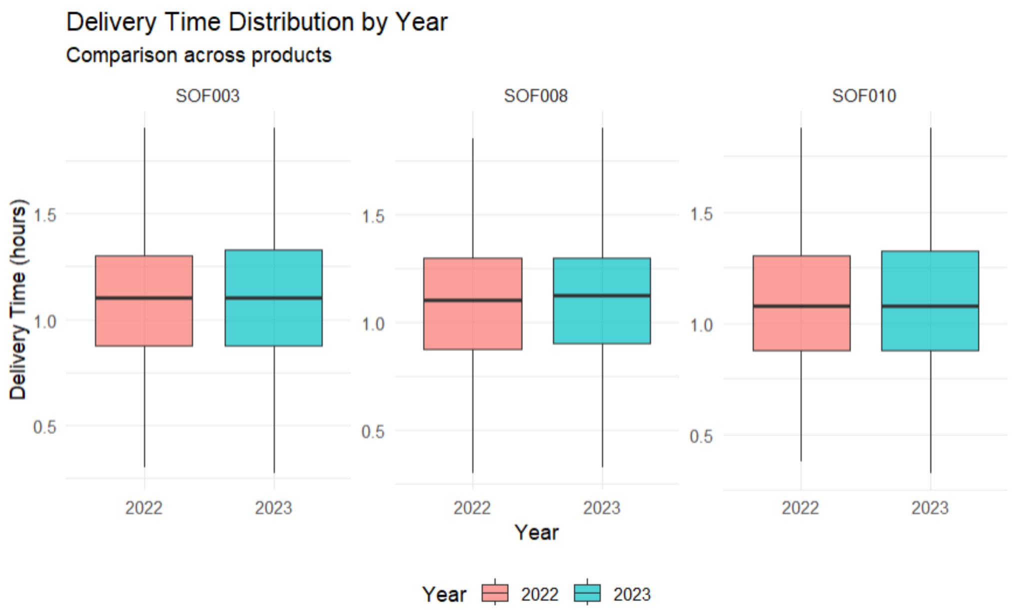

Figure 45: Boxplots of Delivery Distribution by Year

Figure 45 displays the distribution of delivery times (in hours) between the two years for each product. There is very low variation in the mean delivery times and overall distributions shown by the boxplots. This suggests that there is likely little to no significance in the delivery times between the years across each of the products.

Table 19: Summary Values of Top 3 Cpk Products by month

	.product	month_factor	n	mean_delivery	sd_delivery
1	SOF003	1	115	0.9637643	0.2978919
2	SOF003	2	200	0.9672435	0.3108190
3	SOF003	3	199	1.0264844	0.2913946
4	SOF003	4	207	1.0353348	0.2869388
5	SOF003	5	195	1.0580138	0.2963794
6	SOF003	6	174	1.1036828	0.3237890
7	SOF003	7	174	1.1068448	0.2825332
8	SOF003	8	180	1.1300239	0.2797693
9	SOF003	9	190	1.1554016	0.2960268
10	SOF003	10	154	1.1931565	0.3030106
11	SOF003	11	158	1.1933873	0.3216805
12	SOF003	12	113	1.2717133	0.2950117
13	SOF008	1	129	0.9582527	0.2832022
14	SOF008	2	202	1.0193223	0.2817323
15	SOF008	3	201	1.0110786	0.3135658
16	SOF008	4	168	1.0477798	0.2789441
17	SOF008	5	176	1.0607699	0.3076843
18	SOF008	6	180	1.0947450	0.2689458
19	SOF008	7	182	1.0764280	0.2791328
20	SOF008	8	187	1.1145481	0.2821517
21	SOF008	9	193	1.1619585	0.3065852
22	SOF008	10	186	1.1936457	0.2726405
23	SOF008	11	177	1.1975881	0.2797293
24	SOF008	12	112	1.2768000	0.2842669
25	SOF010	1	152	0.9436974	0.2896509
26	SOF010	2	181	0.9442243	0.2846611
27	SOF010	3	187	1.0237813	0.2888698
28	SOF010	4	173	1.0565809	0.2811273
29	SOF010	5	186	1.0858446	0.3009963
30	SOF010	6	167	1.0917689	0.2824076
31	SOF010	7	199	1.0692065	0.3136777
32	SOF010	8	203	1.1251552	0.3083082
33	SOF010	9	182	1.1305440	0.2824875
34	SOF010	10	182	1.1665335	0.3099033
35	SOF010	11	183	1.2187760	0.3205824
36	SOF010	12	110	1.2254345	0.2916026

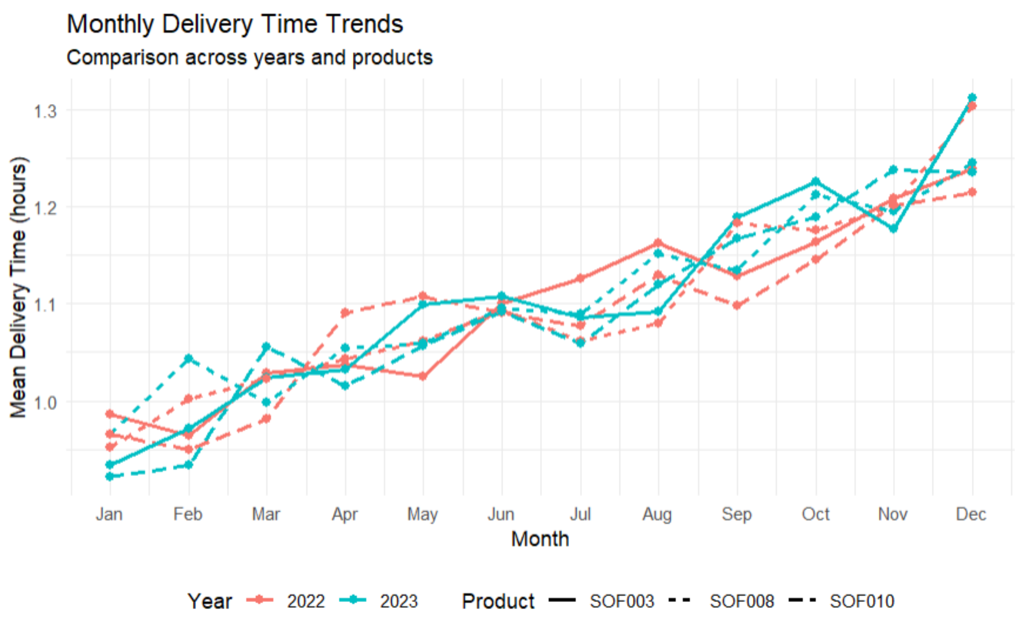

Figure 46: Monthly Delivery Time Trends

The monthly mean delivery times show a general increase across the months from January to December over both years for all three products. However, it also displays significant fluctuations between each month, which may indicate that there is significant difference in the mean delivery times across the months for each year.

```
===== TWO-WAY ANOVA: Year x Month Interaction =====
Analyzing product: SOF008

Two-way ANOVA Results:
                            Df Sum Sq Mean Sq F value Pr(>F)
year_factor                  1   0.04  0.0439    0.53  0.466
month_factor                11  13.91  1.2641   15.28 <2e-16 ***
year_factor:month_factor    11   0.65  0.0587    0.71  0.730
Residuals                 2069 171.17  0.0827
---
```

Figure 47: Two_way ANOVA results for Year x Month Interaction

The two-way ANOVA results confirm that there is significance in the mean delivery times across the months for each year for the product SOF008, as the p-value obtained for the month_factor is much lower than 0.05. This suggests that we reject that null hypothesis: There is no significant difference in delivery times across months. This indicates that an investigation into months showing consistent deviations may be required.Conversely, there is no significant difference in delivery times between the years as the p-value for the year_factor is 0.466. There is also no interaction between the year_factor and month_factor which suggests that they are not dependent on one another, which is presented by the year_factor:month_factor p-value of 0.730.

Table 20: Summary Results

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
year_factor	1	0.04388842	0.04388842	0.5304985	4.664798e-01
month_factor	11	13.90532509	1.26412046	15.2799766	5.323499e-29
year_factor:month_factor	11	0.64568972	0.05869907	0.7095213	7.304830e-01
Residuals	2069	171.16945268	0.08273052	NA	NA

Section 7

7.1 Number of days we can expect reliable service

Table 21: Staff Summary

	staff_on_duty	days	percentage
1	12	1	0.2518892
2	13	5	1.2594458
3	14	25	6.2972292
4	15	96	24.1813602
5	16	270	68.0100756

Number of days with reliable service (>= 15 staff) out of 397 days = 270 + 96 = 366 days

Percentage of days with reliable service = 92.19%

Estimation of the expected reliable service days per year = 336 days

7.2 Optimising profit for the company

Table 22: Cost-Benefit Analysis of Additional Staffing

Cost-Benefit Analysis of Additional Staffing

additional_staff	avg_staff	problem_days	personnel_cost	loss_from_problems	total_cost	savings	roi_pct
0	15.58438	28.5	0	570025.19	570025.2	0.00	NA
1	16.58438	5.5	300000	110327.46	410327.5	159697.73	53.232578
2	17.58438	0.9	600000	18387.91	618387.9	-48362.72	-8.060453
3	18.58438	0.0	900000	0.00	900000.0	-329974.81	-36.663868
4	19.58438	0.0	1200000	0.00	1200000.0	-629974.81	-52.497901
5	20.58438	0.0	1500000	0.00	1500000.0	-929974.81	-61.998321

There is currently and average staff number of 15.58 which results in 28.5 days where the company is understaffed. Although there are no additional costs from hiring more staff, there is a total loss of R570025.19 that is generated from the number of understaffed days. As the average number of employees increase by 1, the number of understaffed days decreases significantly which reduces the loss generated from those days. However, the cost of hiring more personnel increases significantly, which also negatively affects the return on investment. From Table 22, it can be deduced that the optimal number of additional employees that should be hired is 1 as it is the only option that results in a positive ROI of 53.2326%

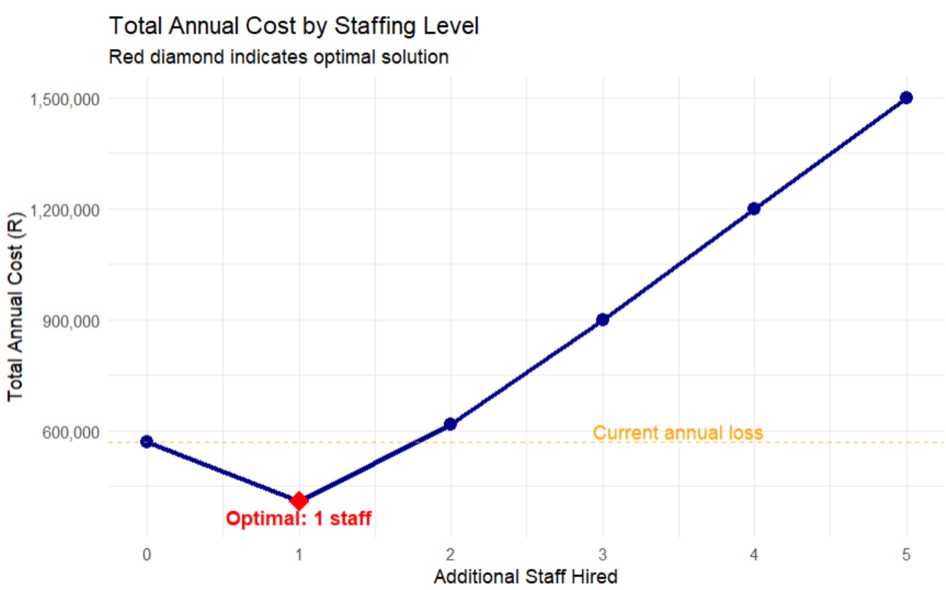

Figure 48: Total Annual Cost by Staffing Level

Figure 48 illustrates the total annual costs incurred per staffing level and further supports the notion that hiring one additional member of staff results in the lowest total annual cost. This is likely due to the lower cost generated from the number of understaffed days effectively balancing out with the slight increase in costs from hiring an additional employee.

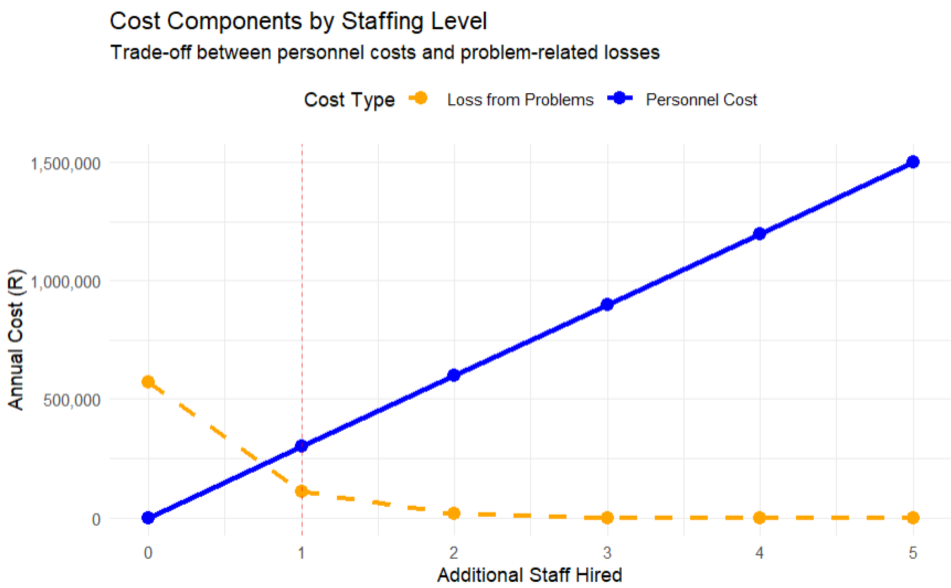

Figure 49: Cost Components by Staffing Level

Figure 49 illustrates the point at which the cost of hiring additional staff cancels with the loss generated from having days that are understaffed. As this point is closest to 1 additional staff member hired, it is further implied that hiring 1 employee will set off the two costs against each other the most effectively.

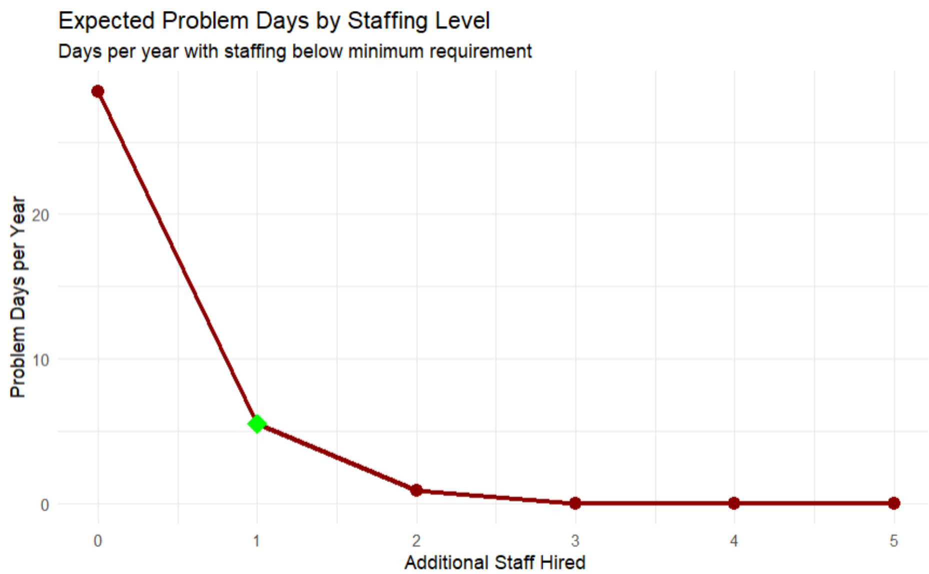

Figure 50: Expected Problem Days by Staffing Level

Figure 50 illustrates that even though the optimal solution is to hire one additional employee, the number of understaffed days will not be zero, but it will be significantly less than if the company were to not hire any additional employees.

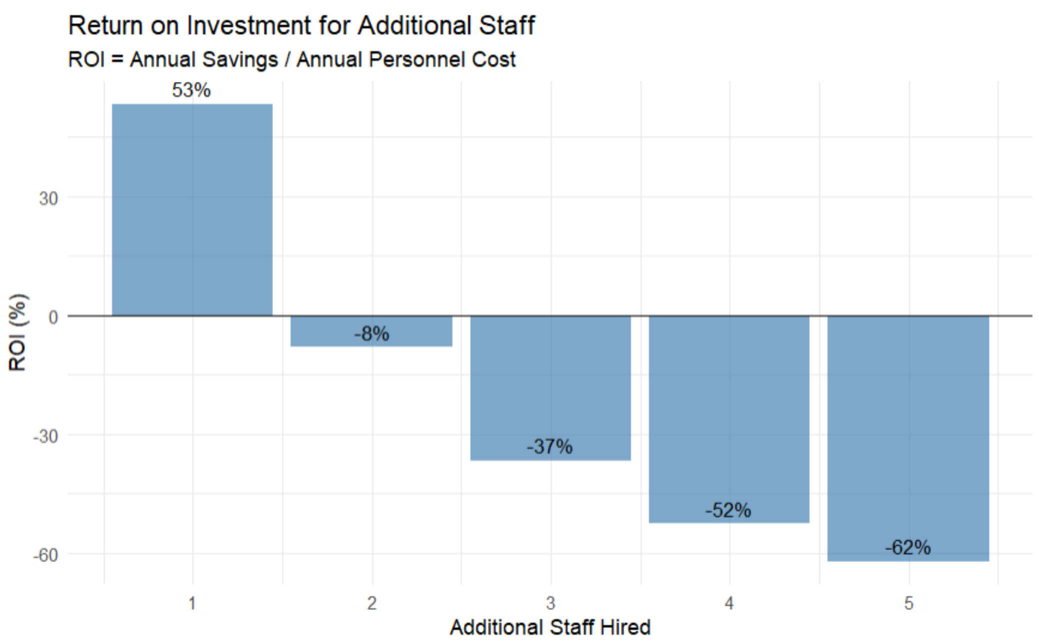

Figure 51: Return on Investment for Additional Staff

Figure 51 once again illustrates that the investment of hiring an additional staff member is beneficial for the company as it will results in an overall high positive return on investment in comparison to the other options.

Table 23: Optimization Summary Results

Metric	**Value**
Current Staff Average	15.6
Additional Staff Recommended	1
New Staff Average	16.6
Current Reliability	92.2%
Optimized Reliability	98.5%
Current Problem Days/Year	29
Optimized Problem Days/Year	6
Annual Personnel Investment	R 3e+05
Annual Savings	R 159,698
Net Annual Benefit	R -140,302

In conclusion, the optimal solution is to hire one additional employee, which will result in an increased reliability of 98.5% per year and 23 days less of potential understaffed days in a year. Although it will require an annual personnel investment of R300000, this increased cost is balances it with the decreased loss generated from the lower number of understaffed days in a year and results in savings of R159 698, resulting in the highest ROI of 53%.