Quality Assurance

ECSA Project GA4

Olivier, M, Mr [25962604@sun.ac.za]
Stellenbosch University

Table of Contents

Part 1 Descriptive statistics. 3

Data visualization 3

Scatterplots 3

Box plots 6

Scatterplots 9

Relationships 12

Part 3 15

X-bar and S-charts 15

X-bar chart 15

S-chart 15

Process Capabilities 19

Process Control Issues (Rule based) 20

Rule A 20

Rule B 21

Rule C 21

Final recommendation 22

Part 4 23

Probability of Type I error for A, B, and C 23

Type I error 23

Probability 23

Probability of consecutive samples above the centreline 23

Estimating the probability of making a Type II error 23

Updated files basic data analysis 24

Summary of the original and fixed data set 24

Boxplots 25

Average selling price vs cost 26

Total profit by category 27

Part 5. 28

Optimizing Profit 28

Step 1 28

Step 2 29

Step 3 31

Part 6 33

Introduction 33

Aim of the analysis 33
Experimental design and hypothesis 33
Data preparation. 34
Interpretation........ 34
Diagnostics check 34
Graphical analysis........ 35
Outcome 39
Part 7 40
Expected number of reliable days per year 40
Profit optimization 40
Model setup........ 40
Final recommendation 40
References........ 41

Part 1 Descriptive statistics.

Data visualization

Scatterplots

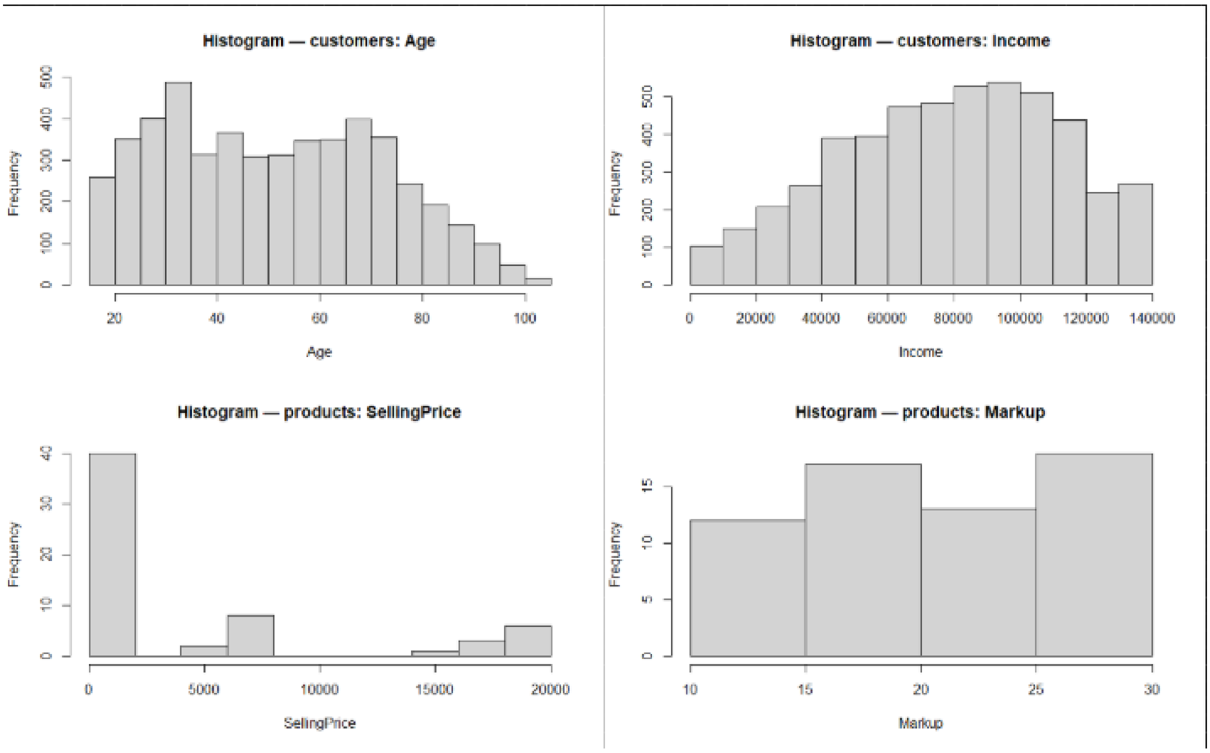


Figure 1

Customer Age: The distribution is slightly different right skewed with most of the customers falling between ages 30 and 60 years old. This indicates that the market lies mostly in the middle aged group.

Customer income: The distribution is normal with the majority of customers earning between R 50 000 and R 120 000 per year. There is a small number of lower income and also a few high income outliers that earn more than R 130 000 per year. This spread suggests a fairly balanced but moderately affluent customer base.

Products selling price: The distribution is highly right skewed with most product listed at low prices and a small amount of product being more expensive that R 10 000 – R 20 000. This indicated the set is dominated by affordable products with view premium high valued products.

Products Markup: The markup distribution appears fairly uniform, with most products maintaining markups between 10 % and 30 %. This suggests consistent pricing policies across categories and limited evidence of extreme under- or over-pricing.

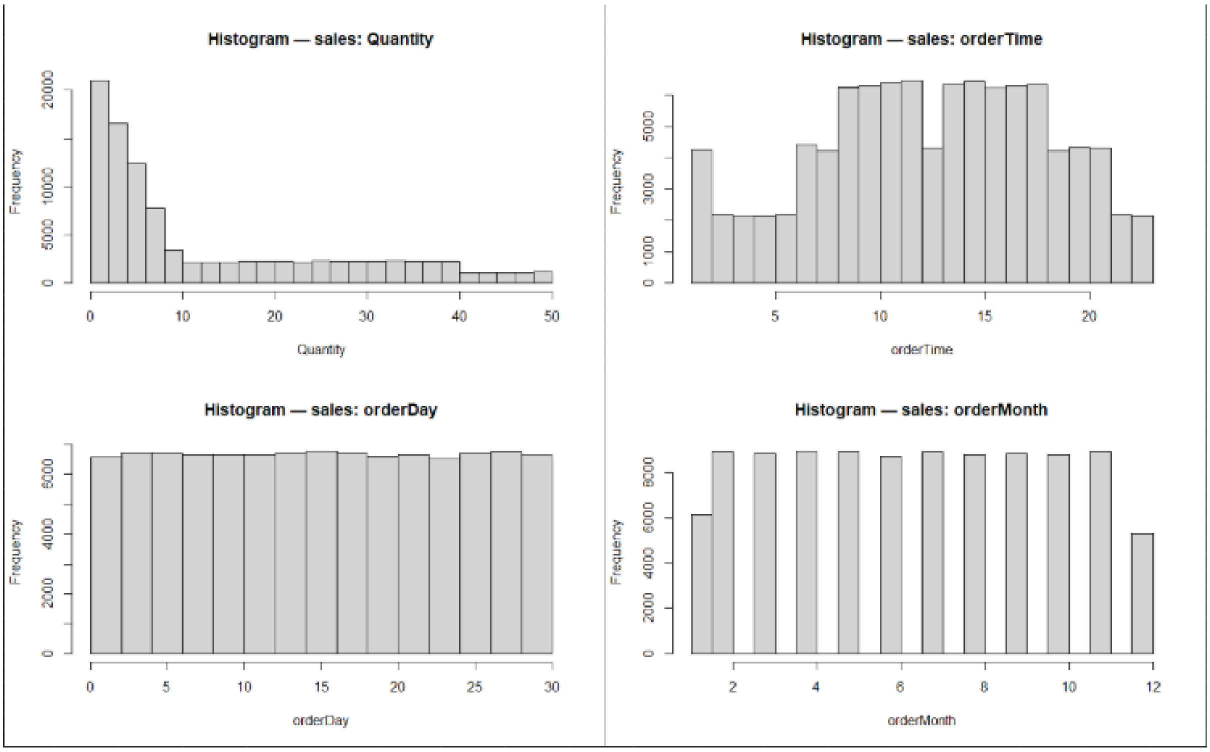


Figure 2

Sales Quantity: The quantity distribution is strongly right skewed where most of the order sizes being smaller order sizes between 1-5 units and progressively less large order sizes. This is normal to see in the retail environment where smaller order sizes usually dominate larger order sizes.

Sales orderTime: Order time has a rough uniform to almost normal distribution with the highest frequencies between 10-20. This is an indication that most orders take place during daytime/business hours.

Sales orderDay: We can see that order day has a even uniform distribution across the days of the month indicating that sales stay consistent during the month with no days as outliers. This supports operational stability as the workload is spread evenly throughout the month.

Sales orderMonth: Sales stay quite consistent per month during the year. This is an indication of steady sale performance with no strong seasonality during the year. There are drops in volume during the months of January and December which can be an indication of fewer working days or the business being closed in holidays.

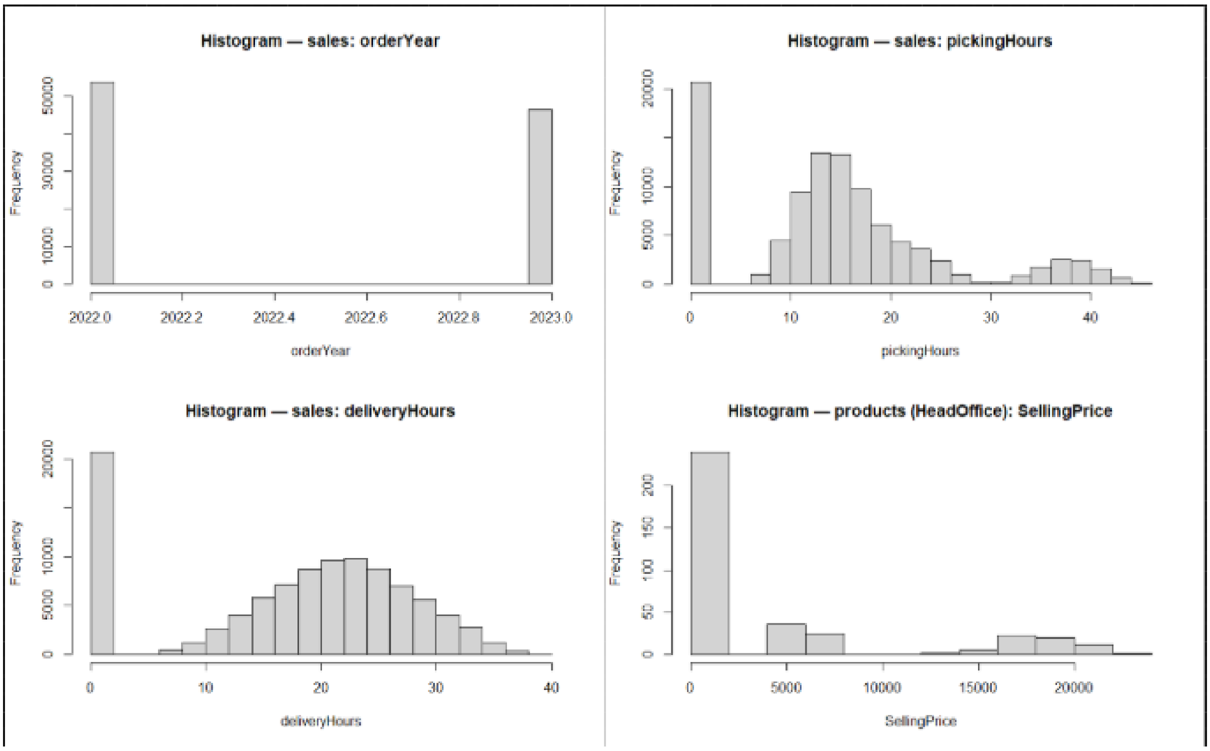


Figure 3

Sales orderYear: We can see that sales have a uniform distribution between the two years 2022 and 2023 with the volume having only a small difference. This is an indication of consistent sales across the two years.

Sales pickingHours: Picking hours has a right skewed distribution where many orders require less than 20 picking hours and only a few that require more than 30 hours. The long tail is an indication of occasional operational delays or complex orders, but most of the orders picking hours occur efficiently in a narrow time window.

Sales deliveryHours: There is a large spike at 0, which is an indication that most of the deliveries take less that an hour to be delivered. Then the rest of the histogram follow a normal distribution with most of the deliveries taking around 20 hours. This indicates a stable process mean and variance.

Products headOffice SellingPrice: We can see that the head office selling prices are highly right skewed dominated by low prices and few products that cost more than R15 000. This indicates a strategy that prioritizes affordability of products while maintaining a smaller portfolio of premium higher priced products.

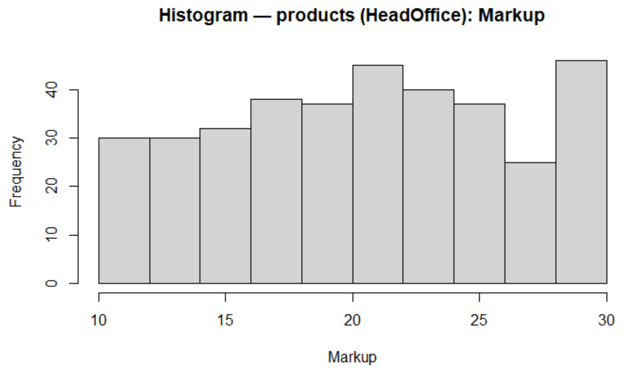


Figure 4

Products Head Office Markup: We can see that markup follow a roughly uniform distribution where there is no markup range that really dominates the other. This is an indication that the company maintains a balanced and standardized markup policy across its product portfolio.

Box plots

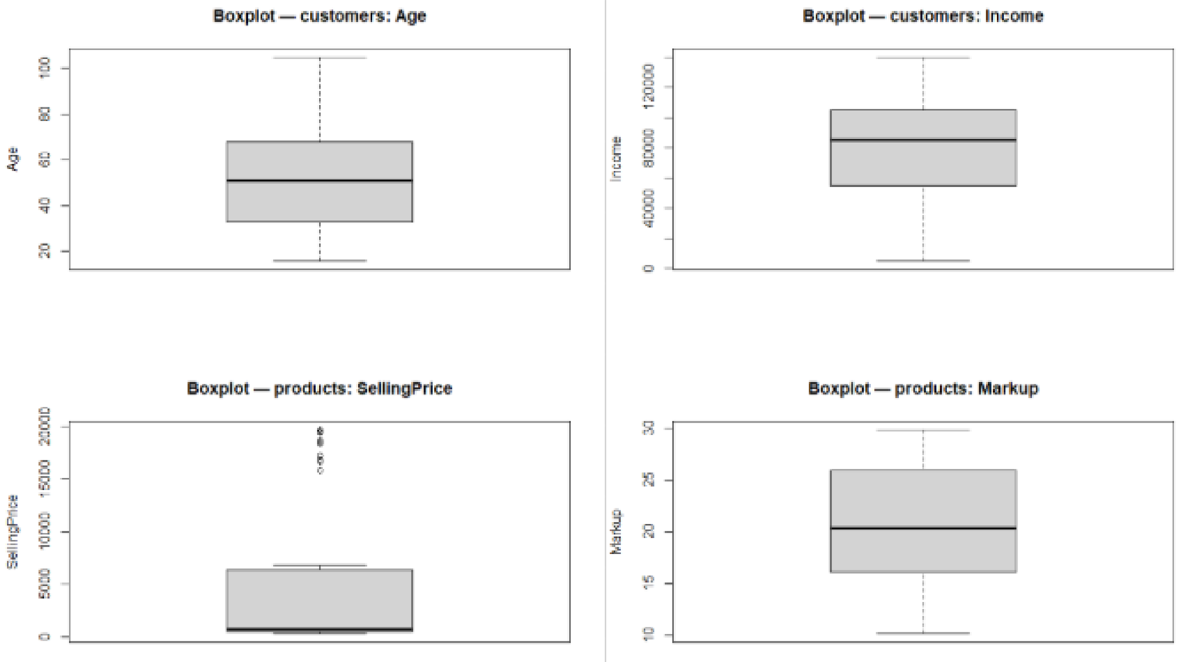


Figure 5

Customers Age: We can see that customer ages range from roughly 20 to a 100 years with the median near 45 years. The interquartile range is moderate showing a balanced age spread around the median age with no extreme outliers. This suggests that customer demographics are stable and representative of typical adult market segment.

Customer Income: The customer annual income indicate a broader spread form roughly R40 000 to R130 0000 with a median near R80 000. The whiskers indicate a mild variability and the absence of extreme outliers suggest consistent income reporting. This is an indication that the customer base in largely middle to upper class.

Products Selling Price: The selling price is skewed and we can see that there are several high price outliers above R10 000. Most of the products fall below R5 000 which again that the company prioritizes affordable products over premium higher priced products. The outliers have to be considered as they represent the higher value products which severely influences the companies' revenue.

Products Markup: The markup values range between 10 and 30 with the median to be found near 20%. The boxplot is symmetric and compact. This is a confirmation that products markup are consistent and well controlled across categories.

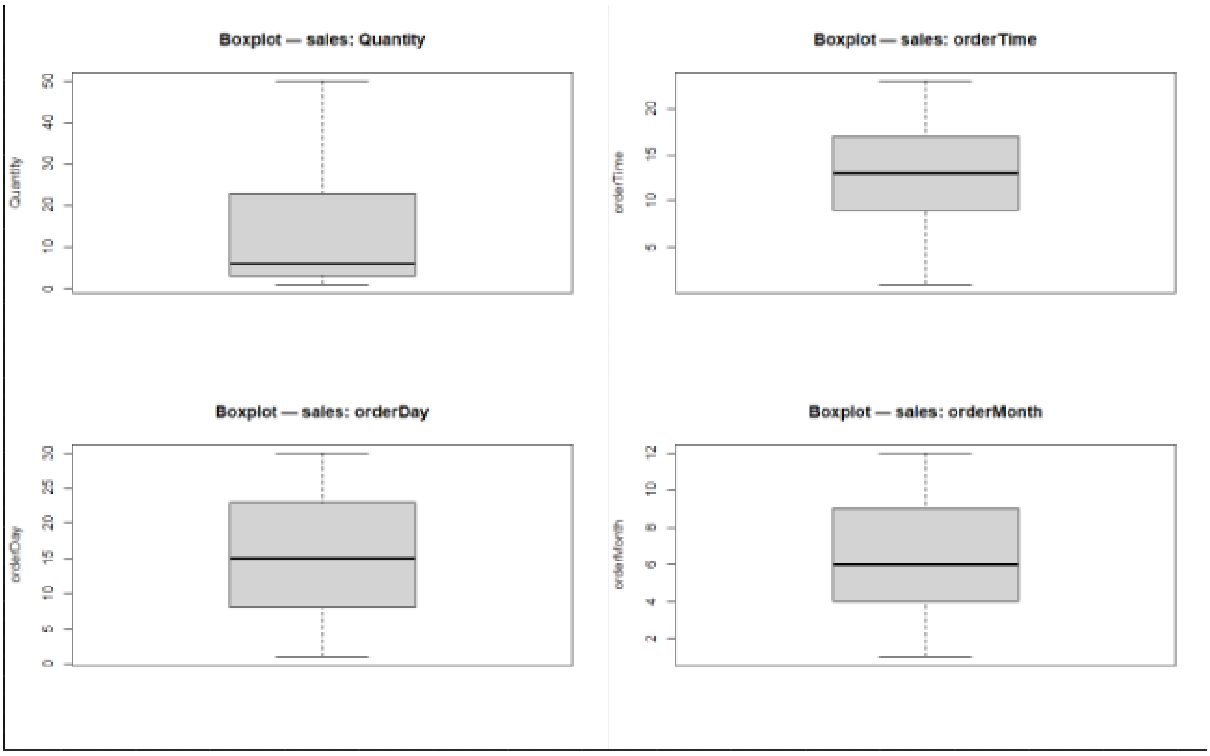


Figure 6

Sales Quantity: From the box plot we can see that quantity has a wide spread with values ranging from 0 to 50 and has a median near 5. This is a indication that most of the orders are small with some large volume outliers.

Sales orderTime: We can see that order time has a consistent spread with the median near 14:00. There are no extreme time outliers. This is an indication that most of the sales take place during normal business hours.

Sales orderDay: The box plot indicates that sales a spread evenly throughout the month with the median near day 15. This is a indication that customer demand remains steady throughout the month across the days.

Sales orderMonth: We can see that the sales stay consistent throughout the year across the 12 months with a slight increase mid year. The median can be found near month 6. There are no strong seasonal peak or dips indicating stable demand throughout the year.

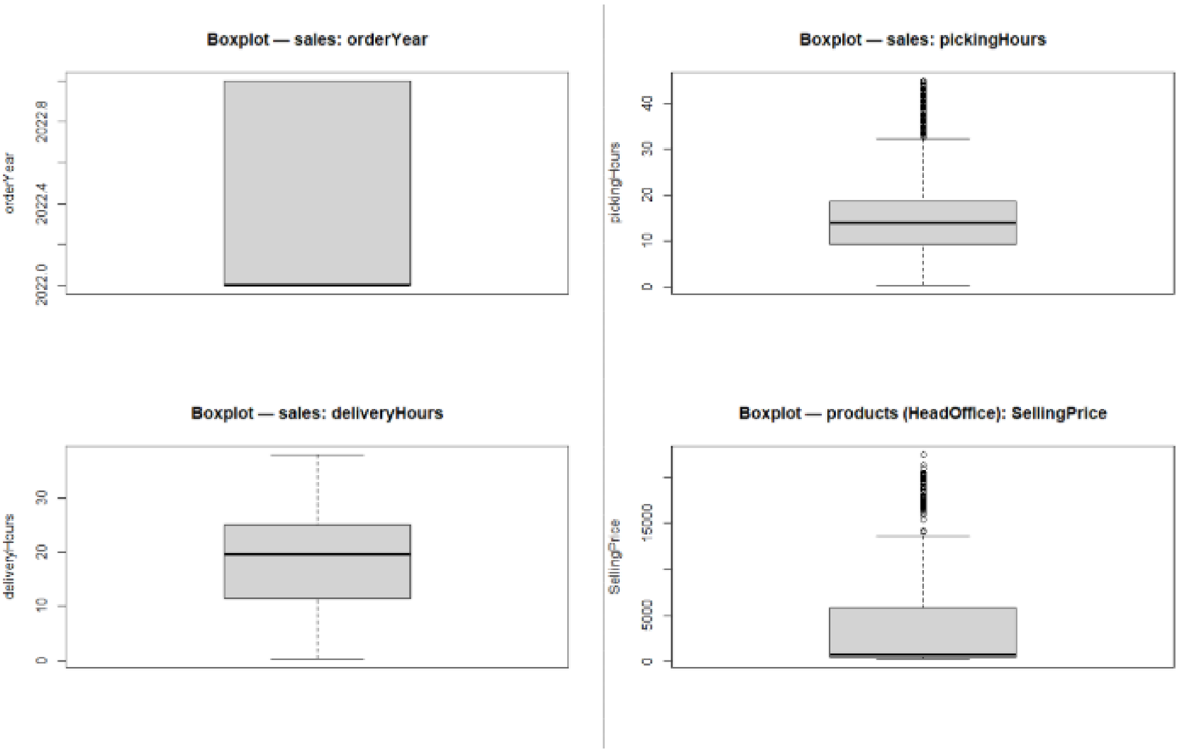


Figure 7

Sales orderYear: From the boxplot we can see that there is two clear data points, 2022 and 2023 confirming that the data spreads across these two years. We can also see that there is no variability which is an indication that the data was correctly filtered for this period.

Sales pickingHours: The boxplot shows a cluster of values below 20 with the median near 10 hours. There are several outliers that can be found 35-50 hours which can be seen as a indication of occasional longer picking durations.

Sales deliveryHours: The boxplot indicates that delivery hours have a moderate spread with the median near 20 hours. There are no extreme outliers. We can see that the box and whiskers are almost symmetrical suggesting a stable and predictable delivery process.

Products HeadOffice SellingPrice: Most of the prices can be found below R5 000 but there are also many extreme outliers present which can be found above R20 000. This shows that the product mix is dominated by lower priced affordable products.

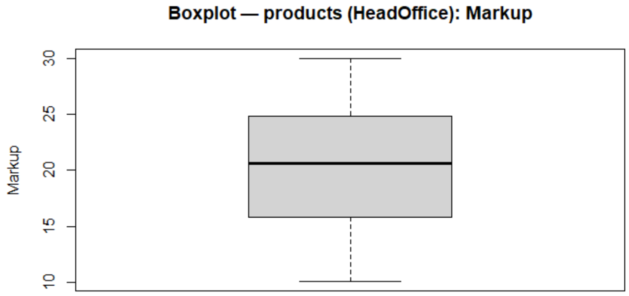


Figure 8

Products HeadOffice Markup: We can see for markup most of the values range from 15%-25% with the median around 20%. The box and whiskers are symmetrical indicating consistent pricing margins across products. There are no outliers present showing that markups are tightly controlled.

Scatterplots

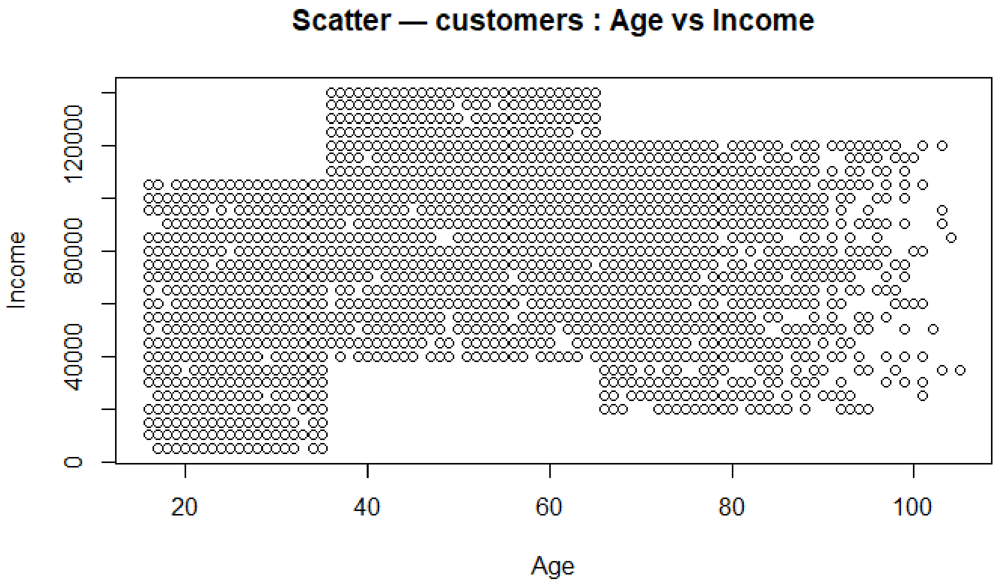


Figure 9

Customers Age vs Income: We can see that the customers age and income have a slightly positive relationship where customers under the age of 35-38 earn around R60 000 and middle aged customers earn around R70 000 – R100 000. We can then also see a slight drop in income after the age of 63 – 65. This is an indication of a realistic demographic indicating that the data is reliable.

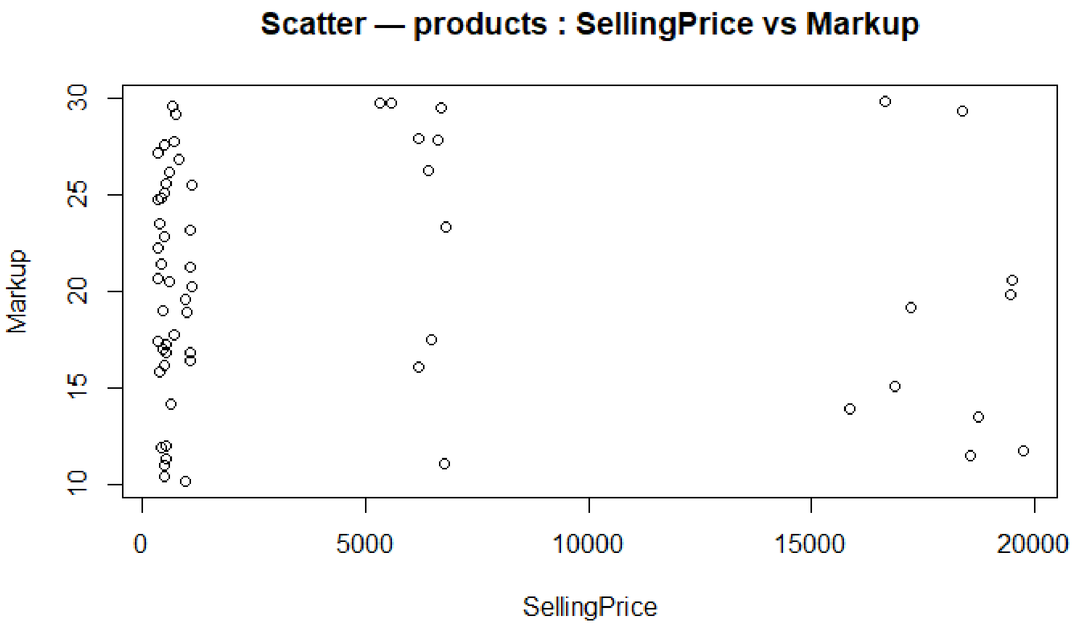


Figure 10

Prodcuts Selling price vs Markup: There is no clear relationship between the products selling price and markup. Product with a lower price (below R5 000) show a wide range of markups between 10%-30% and higher priced products show lower and more consistent markups. This is an indication that the company puts more flexible pricing on lower priced products for competitive reasons.

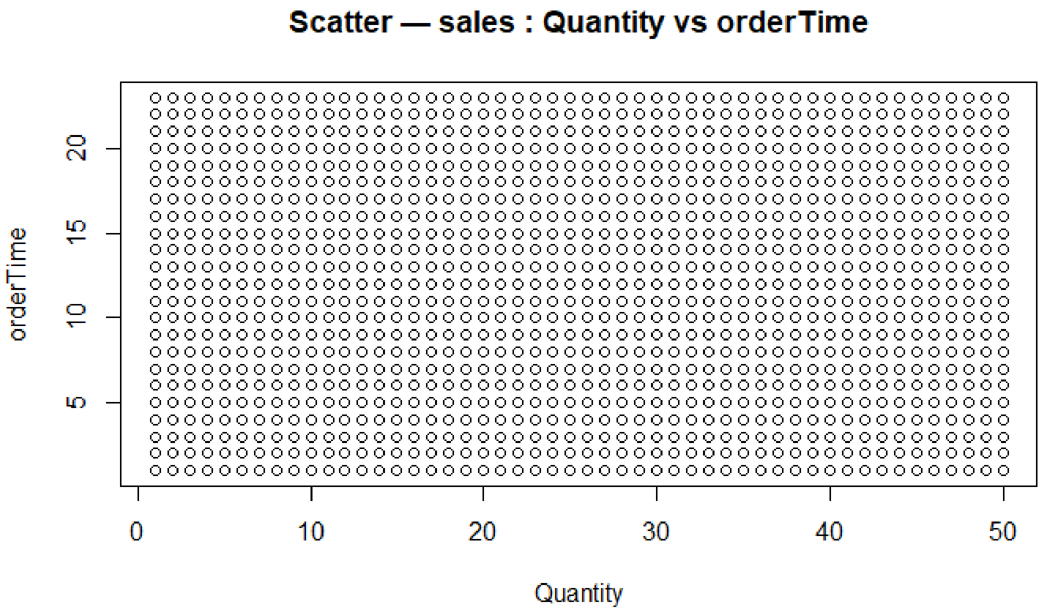


Figure 11

Sales Quantity vs orderTime: The scatter plot show that there is no visible relationship between order time and quantity. This indicates that orders quantities are consistent during the day meaning that order quantity does not depend on the time of the day.

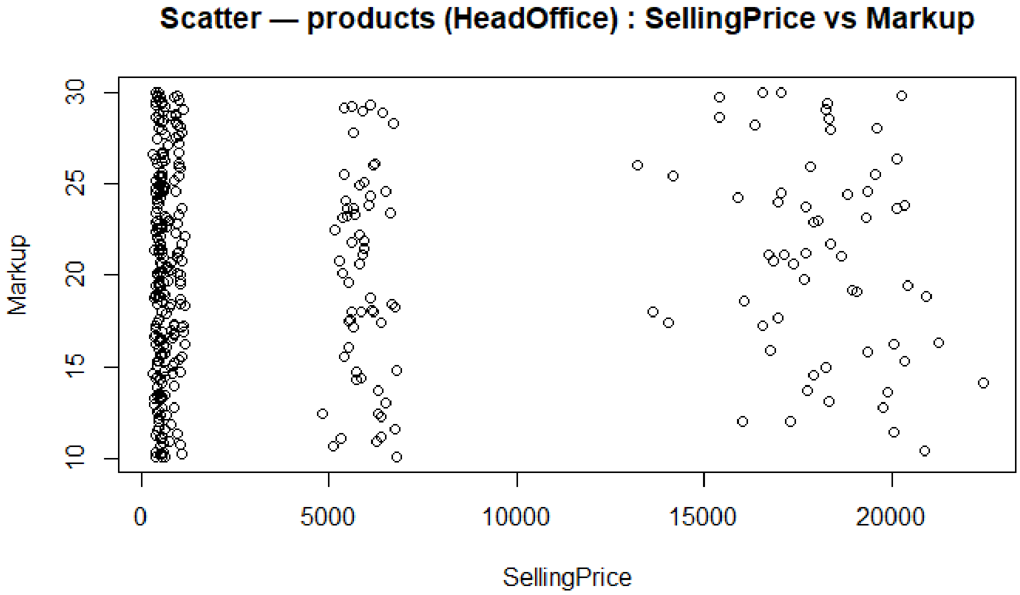


Figure 12

Products HeadOffice SellingPrice vs Markup: We can see that there is no clear relationship between selling price and markup. Products below the price of R5 000 tend to have a very broad markup variation of between 10%-30% and higher priced products maintain more moderate and consistent markups. The flexibility in pricing for lower priced products are for competitive advantage.

Relationships

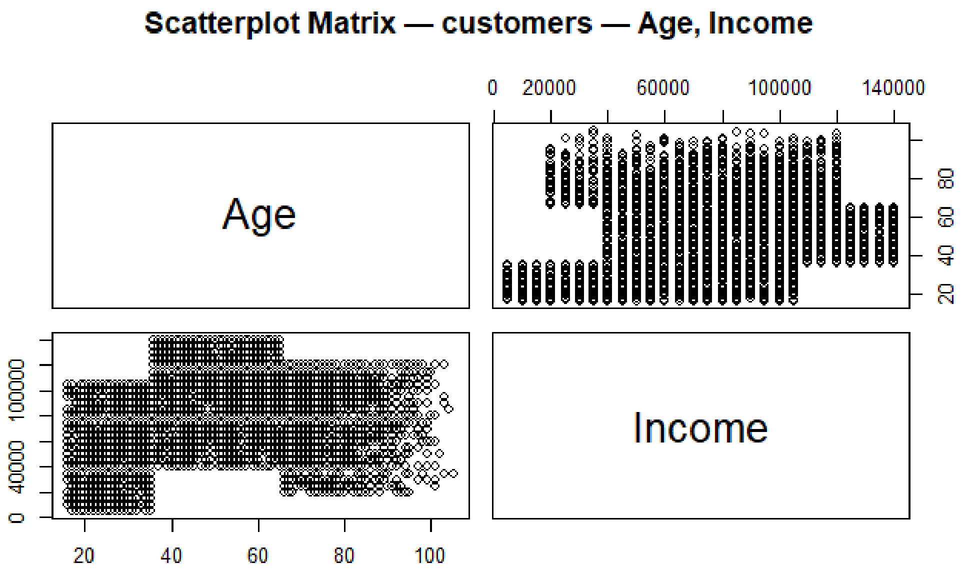


Figure 13

Customers Age, Income: From the plot we can see that there is a weak positive relationship between age and income of customers. The relation ship indicates that the middle age group has the highest income. Income increases after the age of 35-38 up until the age of 63-65 and then lowers again. There is a big spread at each income and age group which is a indication of diversity in financial profiles within similar age groups.

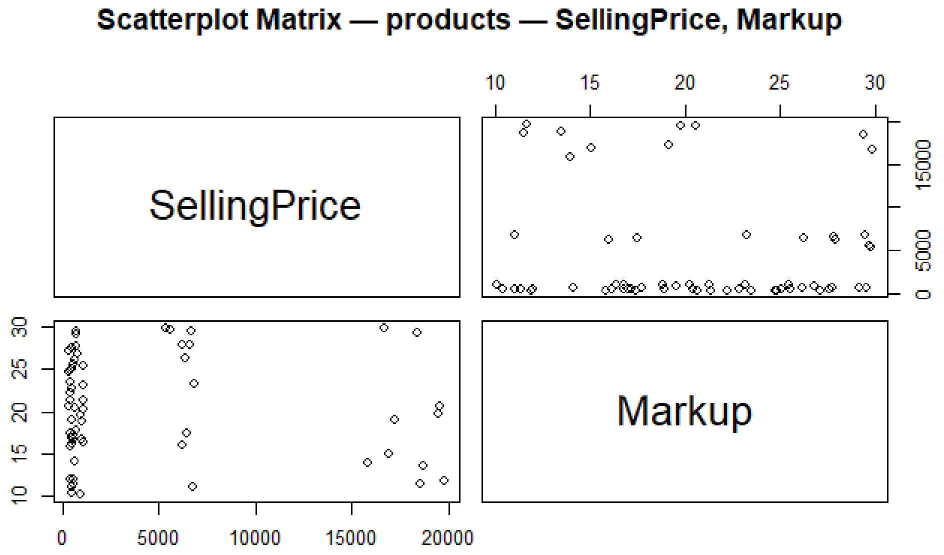


Figure 14

Products SellingPrice, Markup: From the plot we can see that there is no strong correlation between selling price and markups. Most of the products are clustered at lower prices below R5 000 but markup is spread widely between 10%-30% which is an indication of a flexible pricing strategy.

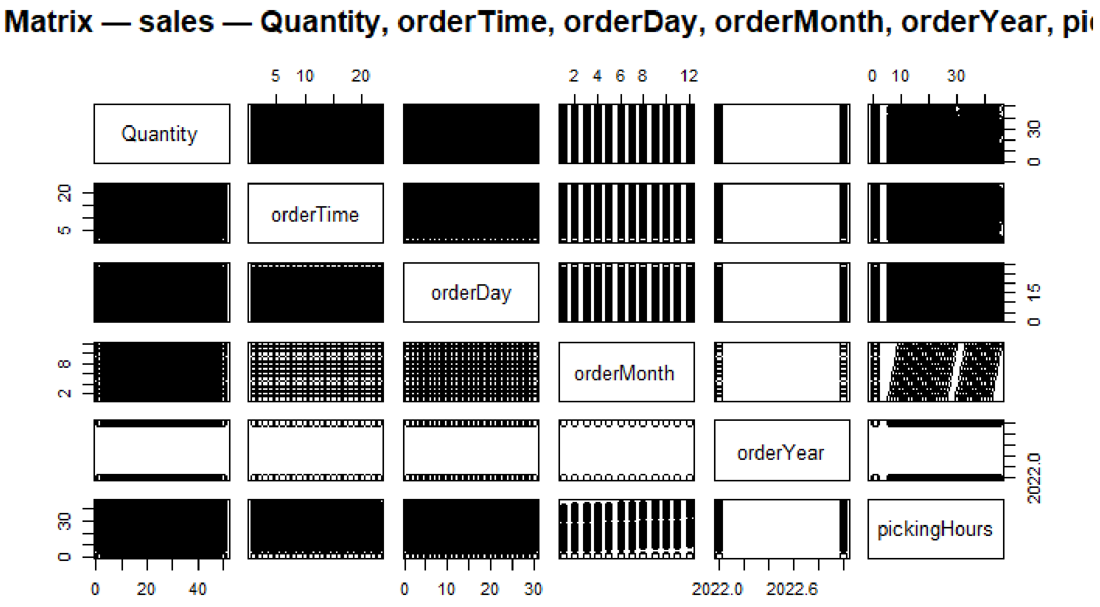


Figure 15

Sales Variables: We can see from the scatterplot that there is no visible linear relationship between the sales variables. We can see that order day and order month appear uniformly distributed which is a indication of stable sales over time periods. Quantity, order time, and picking hours do not show any association which indicates that order volume and timing are independent of operational factors. The data shows steady workflow patterns and not seasonal time based dependencies.

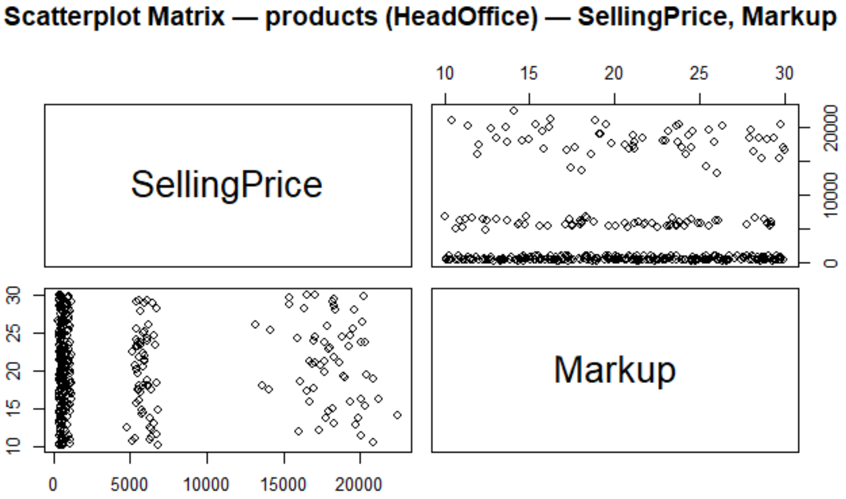


Figure 16

Products HeadOffice Selling price, Markup: There is no strong indication a correlation between selling price and markup. Most of the products can be found at the lower price range below R5 000 while markups as a wide distribution between 10% and 30%.

Part 3

X-bar and S-charts

X-bar chart

Two essential tools in Statistical Process Control (SPC) for tracking and preserving a process's stability over time are an X-bar chart and an s-chart. Changes in the process average are monitored using the X-bar chart, sometimes referred to as the mean chart. The mean of each sample or subgroup, usually consisting of 24 items, is plotted across time in a sequential manner. On an X-bar chart, the upper and lower control limits are usually placed at three standard errors above and below the mean, while the center line shows the grand mean of all subgroup means. The predicted range of typical process variation is delineated by these boundaries. Sample means indicate a possible shift or particular cause variation that requires examination if they deviate from the control limits or show non-random patterns, such as multiple points in a row above the centre line.

S-chart

Alongside the X-bar chart, the standard deviation chart, or s-chart, is used to track process variability or dispersion. It plots each subgroup's standard deviation (s) rather than its means. The average of the subgroup standard deviations (s̄) is represented by the centre line of the s-chart, and its control limits are determined using constants (B3 and B4) that are dependent on the sample size: LCL = B3 * s and UCL = B4 * s. The s-chart is always assessed first because the inferences made from the X-bar chart on changes in the process mean may not be reliable if the process variation is unstable (too much or inconsistent spread).

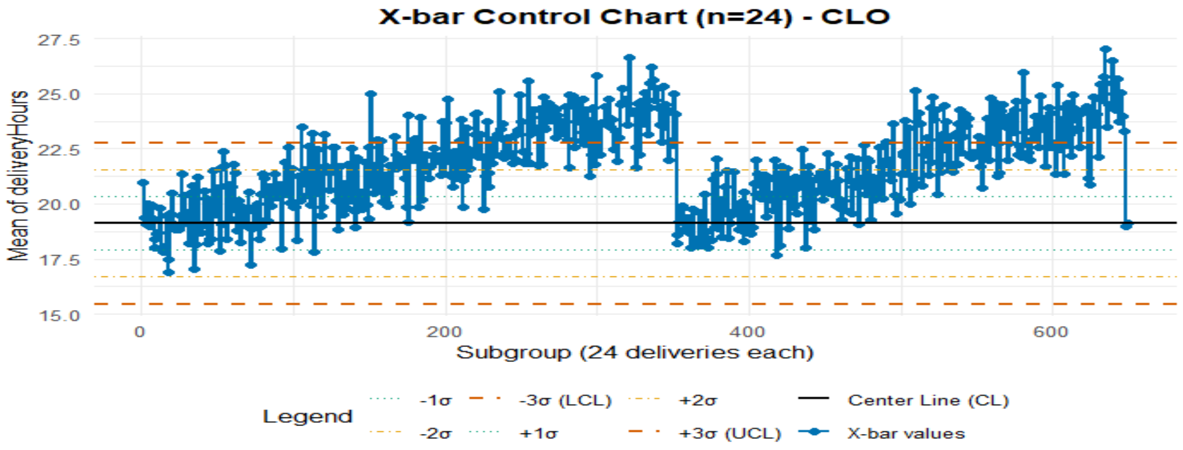


Figure 17

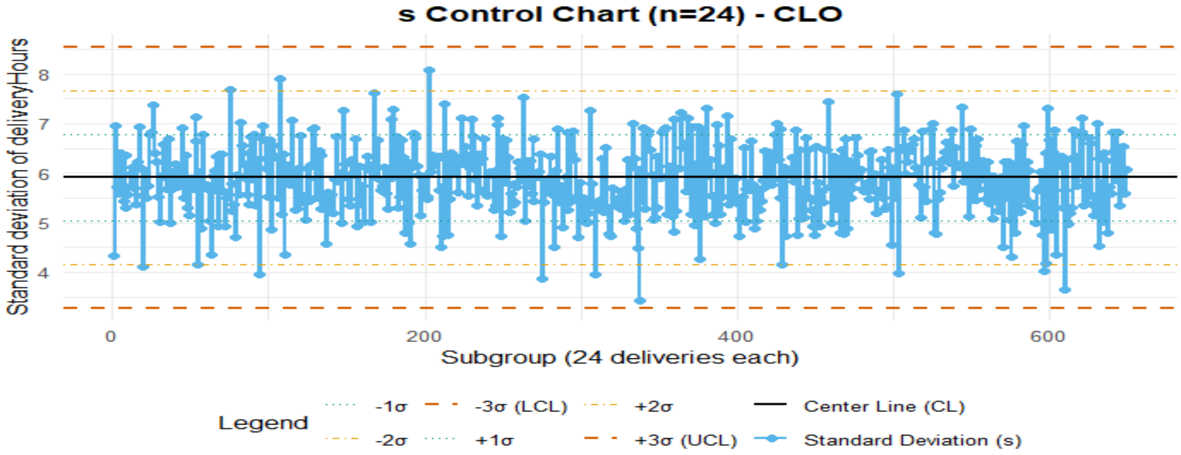


Figure 18

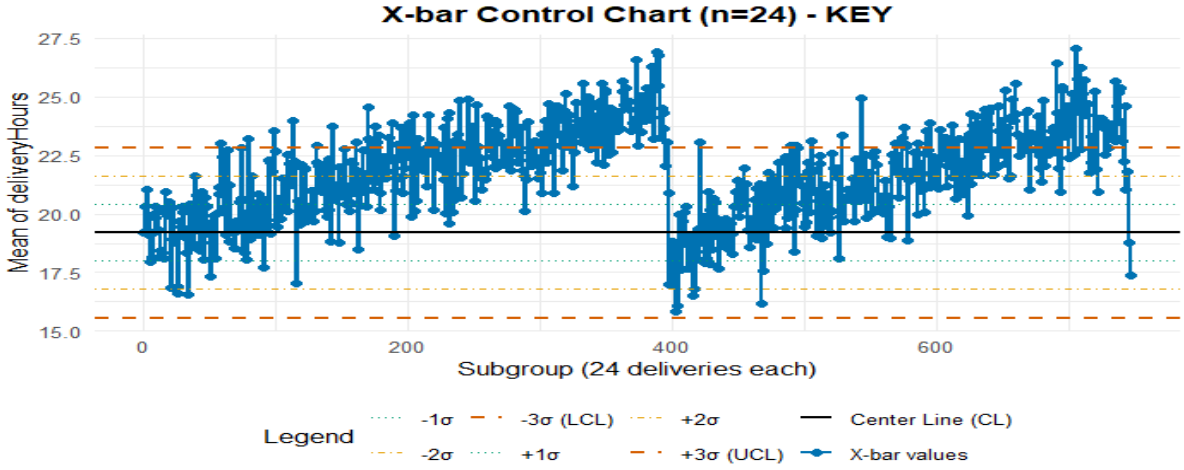


Figure 19

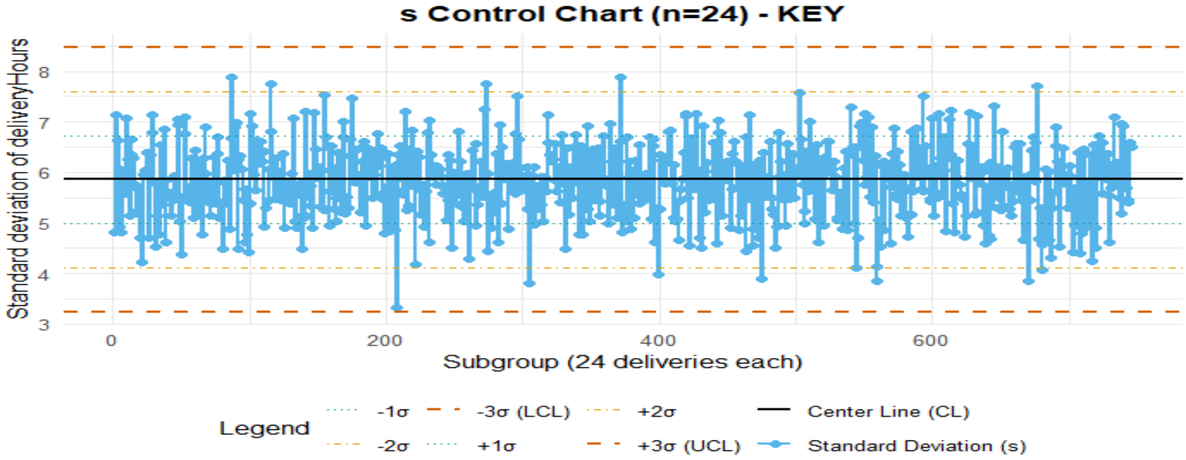


Figure 20

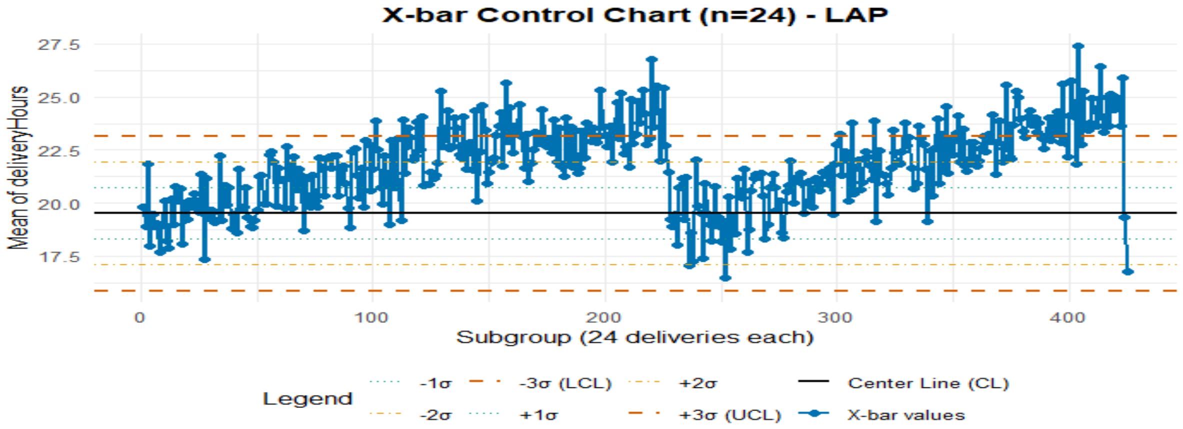


Figure 21

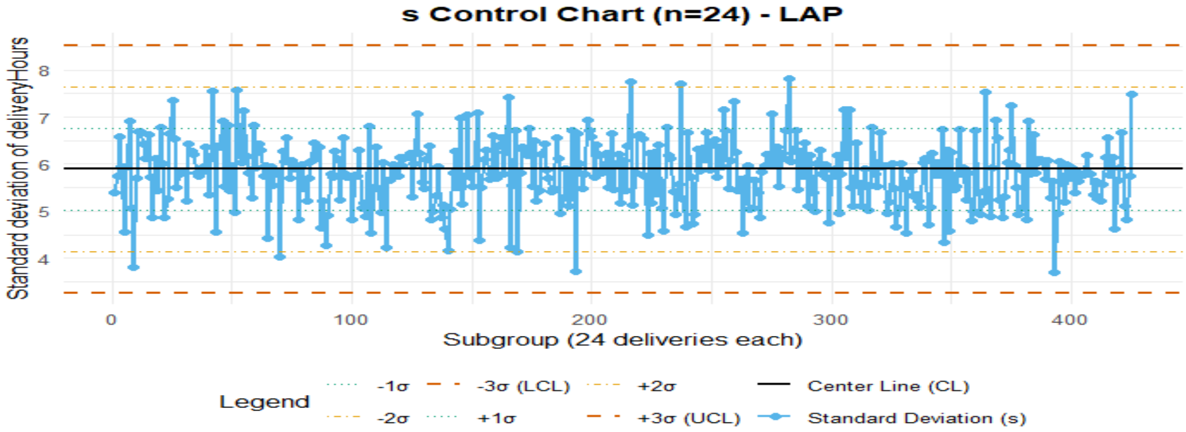


Figure 22

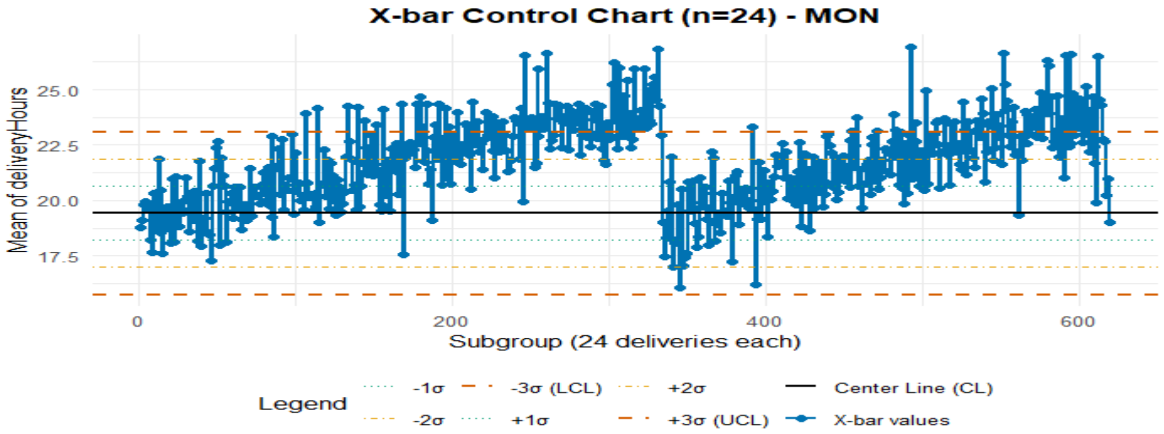


Figure 23

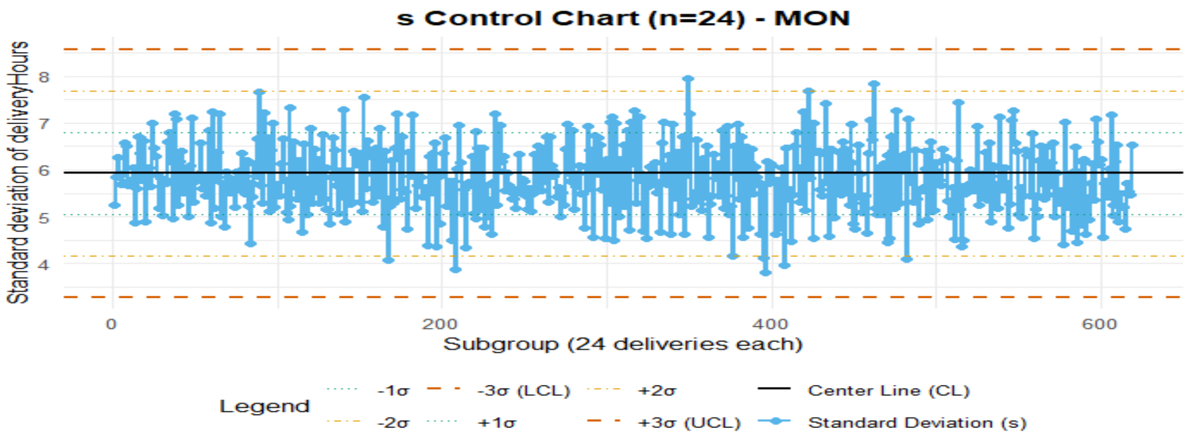


Figure 24

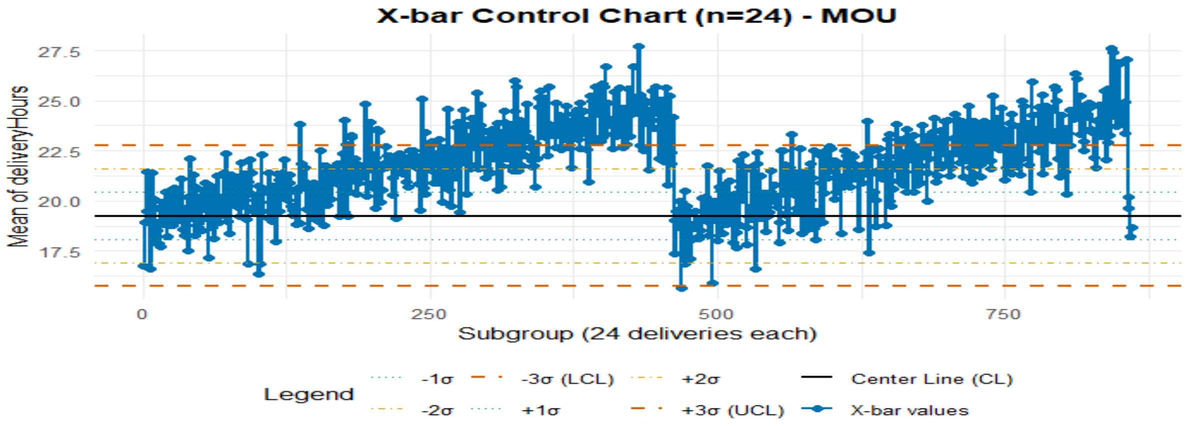


Figure 25

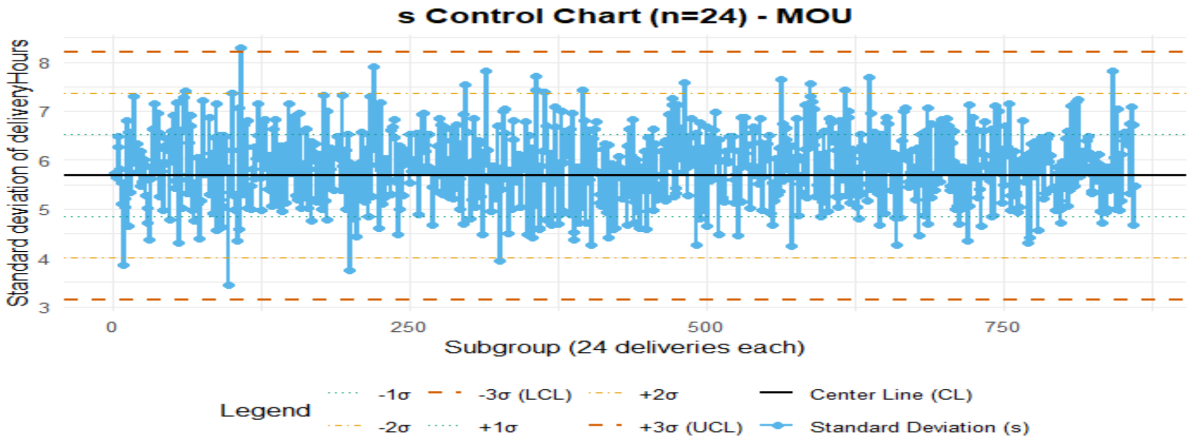


Figure 26

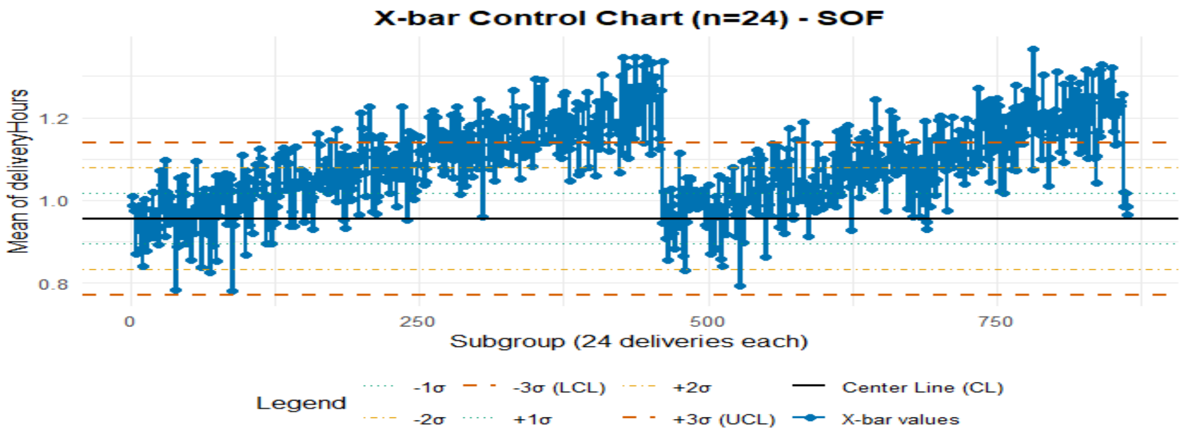


Figure 27

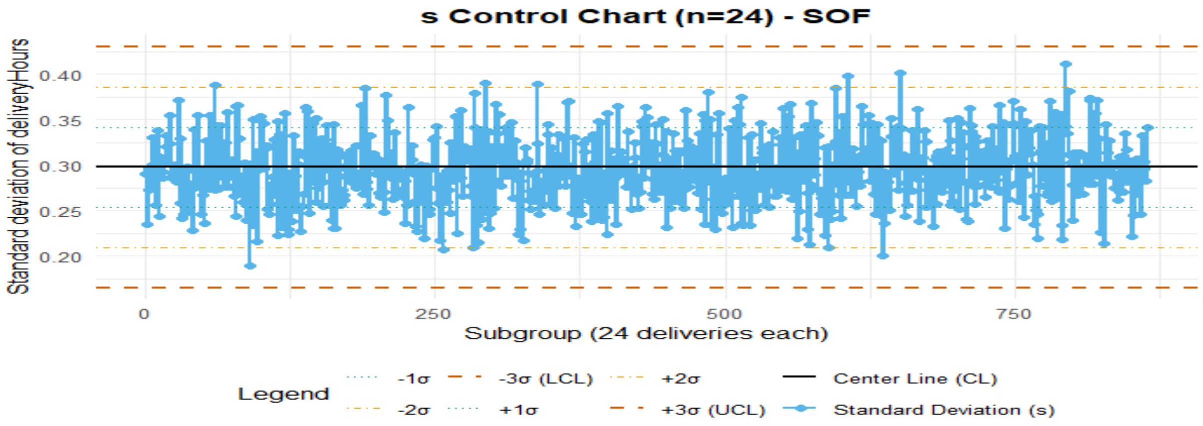
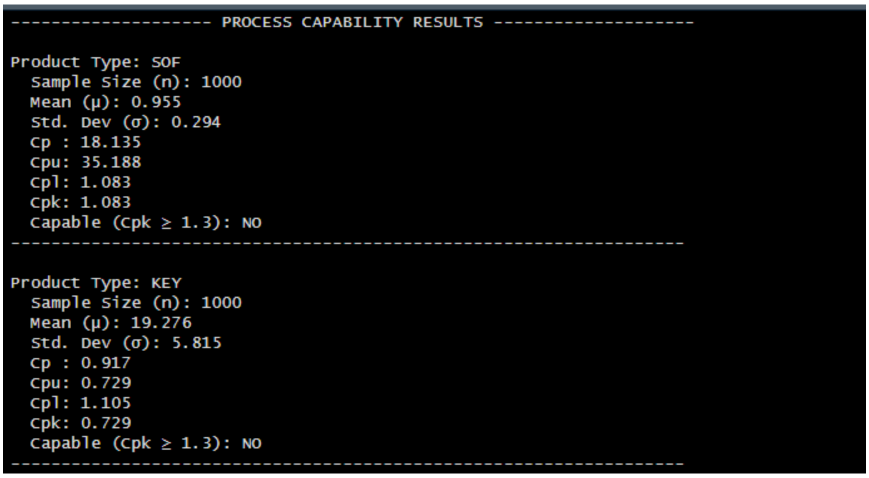


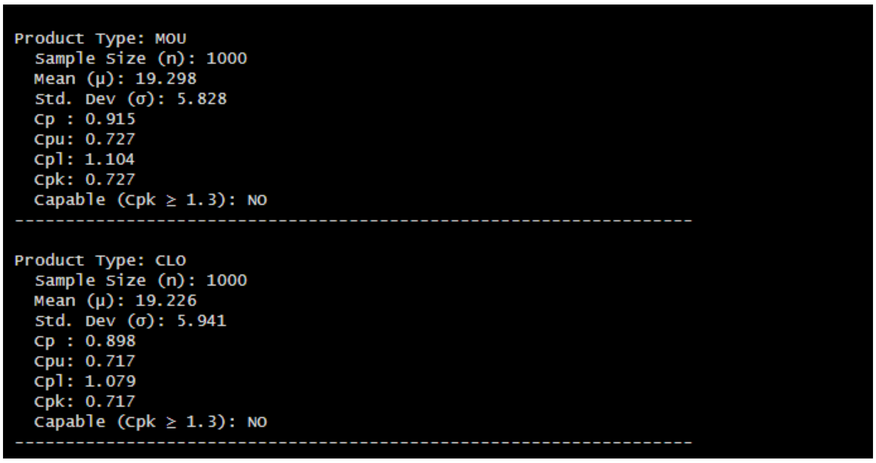
Figure 28

Process Capabilities



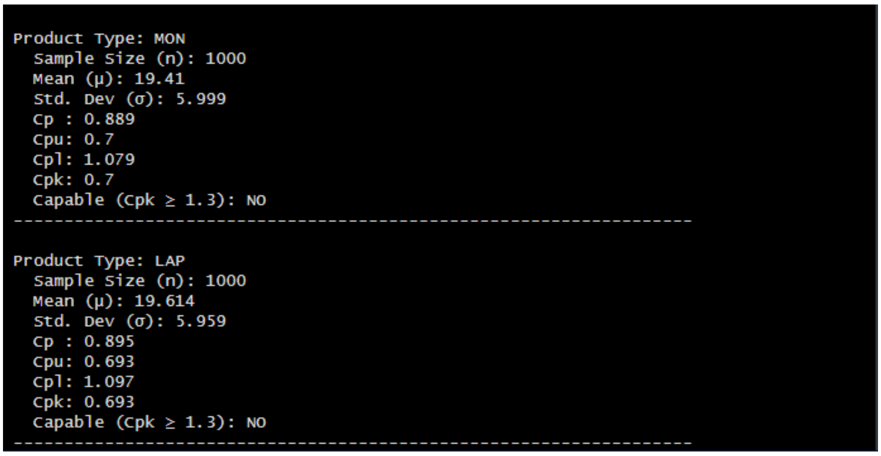
```
-------------------- PROCESS CAPABILITY RESULTS --------------------

Product Type: SOF
  Sample Size (n): 1000
  Mean (μ): 0.955
  Std. Dev (σ): 0.294
  Cp : 18.135
  Cpu: 35.188
  Cpl: 1.083
  Cpk: 1.083
  Capable (Cpk ≥ 1.3): NO
--------------------------------------------------------------------

Product Type: KEY
  Sample Size (n): 1000
  Mean (μ): 19.276
  Std. Dev (σ): 5.815
  Cp : 0.917
  Cpu: 0.729
  Cpl: 1.105
  Cpk: 0.729
  Capable (Cpk ≥ 1.3): NO
--------------------------------------------------------------------
```

Figure 29



```
Product Type: MOU
  Sample Size (n): 1000
  Mean (μ): 19.298
  Std. Dev (σ): 5.828
  Cp : 0.915
  Cpu: 0.727
  Cpl: 1.104
  Cpk: 0.727
  Capable (Cpk ≥ 1.3): NO
------------------------------------------------------------------

Product Type: CLO
  Sample Size (n): 1000
  Mean (μ): 19.226
  Std. Dev (σ): 5.941
  Cp : 0.898
  Cpu: 0.717
  Cpl: 1.079
  Cpk: 0.717
  Capable (Cpk ≥ 1.3): NO
------------------------------------------------------------------
```

Figure 30



```
Product Type: MON
  Sample Size (n): 1000
  Mean (μ): 19.41
  Std. Dev (σ): 5.999
  Cp : 0.889
  Cpu: 0.7
  Cpl: 1.079
  Cpk: 0.7
  Capable (Cpk ≥ 1.3): NO
------------------------------------------------------------------

Product Type: LAP
  Sample Size (n): 1000
  Mean (μ): 19.614
  Std. Dev (σ): 5.959
  Cp : 0.895
  Cpu: 0.693
  Cpl: 1.097
  Cpk: 0.693
  Capable (Cpk ≥ 1.3): NO
------------------------------------------------------------------
```

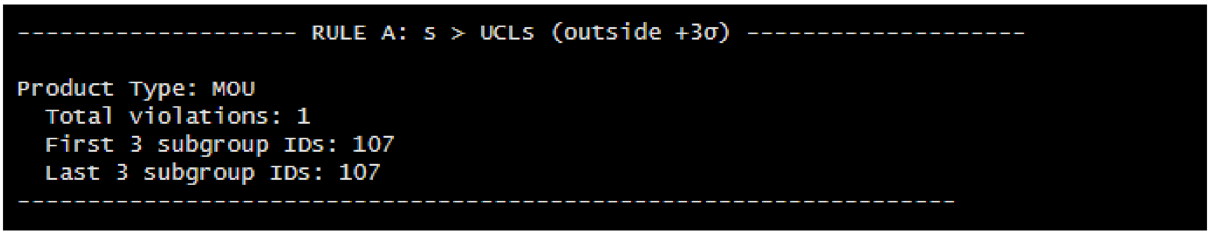
Figure 31

We can see form figures 29 to 31 that all of the product types have a cpk < 1.3 which means that none of them are able to meet the VOC requirements.

Process Control Issues (Rule based)

Rule A

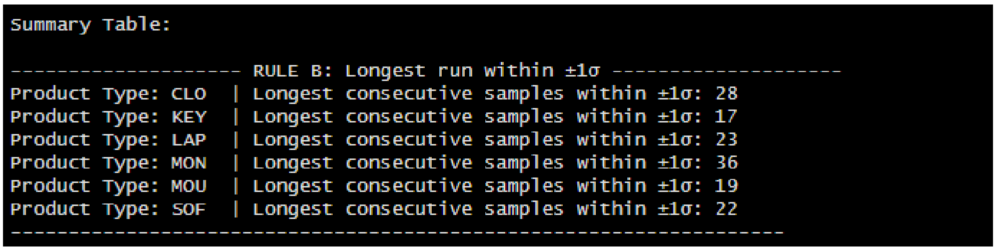
We can see from the top of figure 32 below that there is only one product type that violated rule A. This product type is MOU which violated this rule once with one sample falling outside of the limits. The sample violating this rule is sample ID 107. This sample should be investigated. All of the other samples did not violate this rule indicating good statistical stability.



```
-------------------- RULE A: s > UCLs (outside +3σ) --------------------

Product Type: MOU
  Total violations: 1
  First 3 subgroup IDs: 107
  Last 3 subgroup IDs: 107
------------------------------------------------------------------
```

Figure 32

Rule B

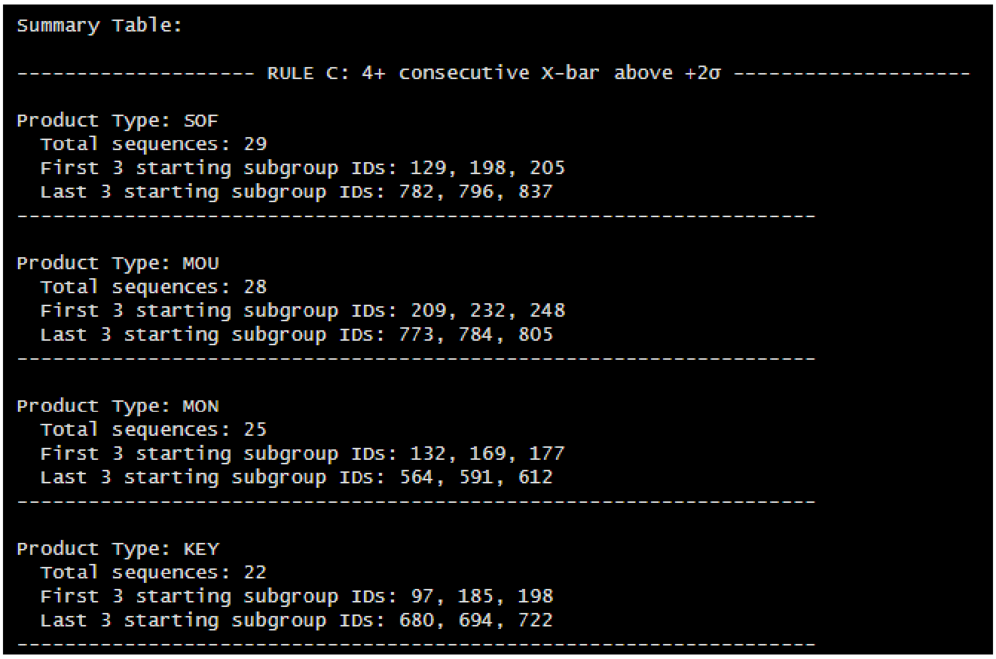


```
Summary Table:

-------------------- RULE B: Longest run within ±1σ --------------------
Product Type: CLO  | Longest consecutive samples within ±1σ: 28
Product Type: KEY  | Longest consecutive samples within ±1σ: 17
Product Type: LAP  | Longest consecutive samples within ±1σ: 23
Product Type: MON  | Longest consecutive samples within ±1σ: 36
Product Type: MOU  | Longest consecutive samples within ±1σ: 19
Product Type: SOF  | Longest consecutive samples within ±1σ: 22
-----------------------------------------------------------------
```

Figure 33

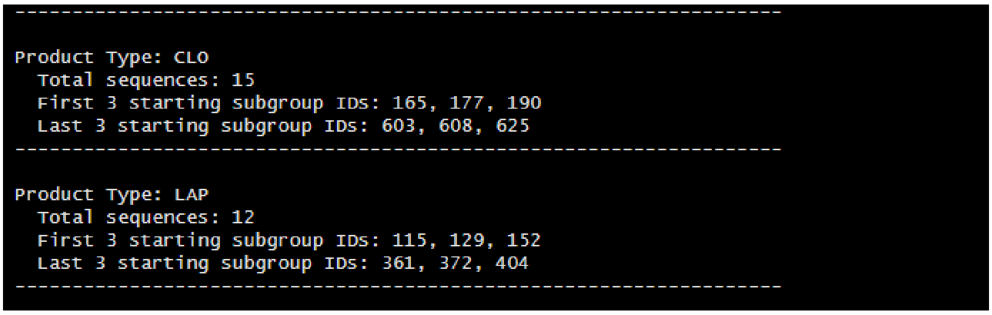
From the R result in figure 33 we can see that all of the product types had a good and long run within the +- standard deviation especially CLO and MON indicating good process line internal control for long periods of time.

Rule C



```
Summary Table:

-------------------- RULE C: 4+ consecutive X-bar above +2σ --------------------

Product Type: SOF
  Total sequences: 29
  First 3 starting subgroup IDs: 129, 198, 205
  Last 3 starting subgroup IDs: 782, 796, 837
-------------------------------------------------------------------

Product Type: MOU
  Total sequences: 28
  First 3 starting subgroup IDs: 209, 232, 248
  Last 3 starting subgroup IDs: 773, 784, 805
-------------------------------------------------------------------

Product Type: MON
  Total sequences: 25
  First 3 starting subgroup IDs: 132, 169, 177
  Last 3 starting subgroup IDs: 564, 591, 612
-------------------------------------------------------------------

Product Type: KEY
  Total sequences: 22
  First 3 starting subgroup IDs: 97, 185, 198
  Last 3 starting subgroup IDs: 680, 694, 722
-------------------------------------------------------------------
```

Figure 34



```
-------------------------------------------------------------------

Product Type: CLO
  Total sequences: 15
  First 3 starting subgroup IDs: 165, 177, 190
  Last 3 starting subgroup IDs: 603, 608, 625
-------------------------------------------------------------------

Product Type: LAP
  Total sequences: 12
  First 3 starting subgroup IDs: 115, 129, 152
  Last 3 starting subgroup IDs: 361, 372, 404
-------------------------------------------------------------------
```

Figure 35

From figures 34 and 35 we can see the following:

- Regular sequences over +2 * standard deviation indicate "special-cause" variation or transient mean shifts.
- Numerous such sequences are displayed in SOF, MOU, and MON; these indicate probable over-delivery (extended delivery durations) at specific intervals.
- Because CLO and LAP display fewer sequences, their mean performance seems more stable.

Final recommendation

The majority of product types are quite stable, according to the SPC results, but none of them can yet fully satisfy client needs (Cpk >= 1.3). The SOF product category is notable for having a high Cp value and minimal variation, indicating consistency. However, its average delivery time is too near the lower limit, resulting in a lower Cpk. With Cp and Cpk values less than 1, the other items (CLO, KEY, LAP, MON, and MOU) all produce greater changes, indicating that their distribution procedures aren't strict enough to meet the regulations.

From the control rules, only MOU showed a single outlier beyond the +3* standard deviation limit, suggesting that most processes are under control. MON and CLO had the longest runs within +-1 * standard deviation, which indicates good short-term consistency, while SOF, MOU, and MON showed several mean shifts above +2 * standard deviation, hinting at occasional process drift.

We can see that the processes are stable but not yet capable. SOF's mean should be shifted slightly in a way that it will be closer to the centre. The other product types should be focussing on lowering the variations. The outlier of MOU spotted for rule A should be inspected.

Part 4

Probability of Type I error for A, B, and C

Type I error

A type I error is when a process is being flagged for being out of control while it is actually in control. This is also know as manufacturer's risk. In context of our problem the null hypothesis states that the is in control and the alternative hypothesis states that the process has shifter and the variation has increased. The probability of making a type I error is theoretical as it depends on the control rule used and the statistical properties of the normal distribution

Probability

Rule A, where a single sample point falls outside the 3*standard deviation control limits, the theoretical probability for this to happen by random chance in a normally distributed process is about 0.0027 or 0.27%. This indicates that even if the process is under perfect control and we plot 1000 points there will still be at least 3 points that outside of the control limits. Rule B which identifies long runs of points within the +-1*standard deviation range, does not actually signal a out of control process and therefore the probability of making a type I error for it is ultimately zero. When four consecutive sample means slip beyond the +2*standard deviation line, Rule C raises the possibility of a problem. The likelihood of a single point lying above +2*standard deviation is 0.0228, and the likelihood of four consecutive points lying above +2*standard deviation is (0.0228)^4 = 2.7 * 10^(-7), assuming a typical, stable process. This is a very uncommon occurrence, with a false alarm probability of about one in 3.7 million.

Probability of consecutive samples above the centreline

The centreline on the control chart indicates the process mean for a perfectly cantered and normally dispersed process. Any single sample has an equal chance (0.5) of falling above or below the centreline since a normal distribution is symmetrical about its mean. Consequently, there is a 0.5 chance that one sample will be above the centreline. For seven successive samples, the likelihood that they will all fall above the centreline is 0.5^7 = 0.0078, or around 8 in 1000. This is the probability that, in a situation where the process is in control, such a pattern would emerge by pure chance.

Estimating the probability of making a Type II error

Making a type II error occurs when the process is actually out of control and we fail to pick it up and assume that it is still in control. This is also know as a consumers error. The probability of making a type II error for the bottle filling process is as follows:

Beta = P(25.011 < Xbar < 25.089)

u = 25.028 and standard deviation of Xbar = 0.017

Zlcl = (25.011 - 25.028)/0.017 = -1

Zucl = (25.089 – 25.028)/0.017 = 3.59

From the standard normal tables:

P(Z<3.59) = 0.9998

P(Z<-1) = 0.1587

Beta = 0.9998 – 0.1587 = 0.8411

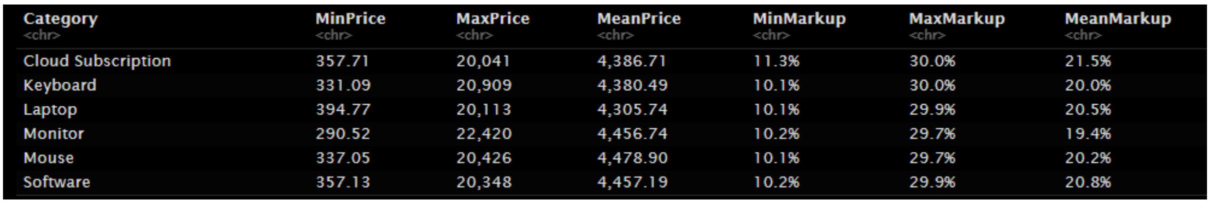
Therefore the probability of making a type II error in the bottle filling process is 0.8411 or 84.11%.

Updated files basic data analysis

In week one there were several differences between products_Headoffice2025.csv and products_data.csv, product IDs, prices and markups were incorrectly recorded. The biggest problems were the product IDs where instead of their code like MOU and SOF there were NA values, and wrong selling price and markup values. We addressed these issues and fixed the data set.

Summary of the original and fixed data set

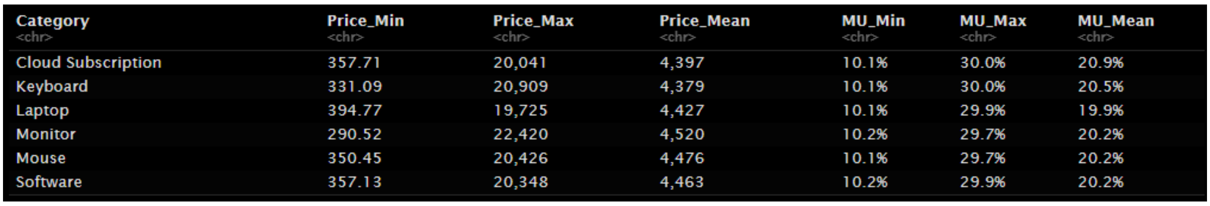
Original:



Category <chr>	MinPrice <chr>	MaxPrice <chr>	MeanPrice <chr>	MinMarkup <chr>	MaxMarkup <chr>	MeanMarkup <chr>
Cloud Subscription	357.71	20,041	4,386.71	11.3%	30.0%	21.5%
Keyboard	331.09	20,909	4,380.49	10.1%	30.0%	20.0%
Laptop	394.77	20,113	4,305.74	10.1%	29.9%	20.5%
Monitor	290.52	22,420	4,456.74	10.2%	29.7%	19.4%
Mouse	337.05	20,426	4,478.90	10.1%	29.7%	20.2%
Software	357.13	20,348	4,457.19	10.2%	29.9%	20.8%

Figure 36

Fixed:



Category <chr>	Price_Min <chr>	Price_Max <chr>	Price_Mean <chr>	MU_Min <chr>	MU_Max <chr>	MU_Mean <chr>
Cloud Subscription	357.71	20,041	4,397	10.1%	30.0%	20.9%
Keyboard	331.09	20,909	4,379	10.1%	30.0%	20.5%
Laptop	394.77	19,725	4,427	10.1%	29.9%	19.9%
Monitor	290.52	22,420	4,520	10.2%	29.7%	20.2%
Mouse	350.45	20,426	4,476	10.1%	29.7%	20.2%
Software	357.13	20,348	4,463	10.2%	29.9%	20.2%

Figure 37

We can see that after fixing the data set that the values of the summary tables stay fairly the same but the fixed data's summary provides more consistent and realistic values.

Boxplots

Original:

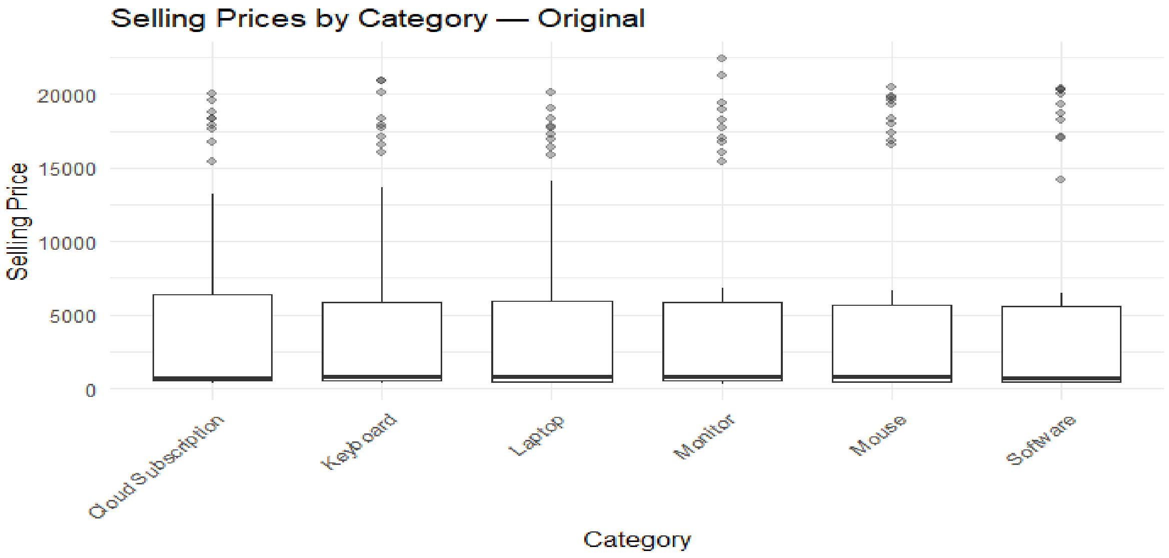


Figure 38

Fixed:

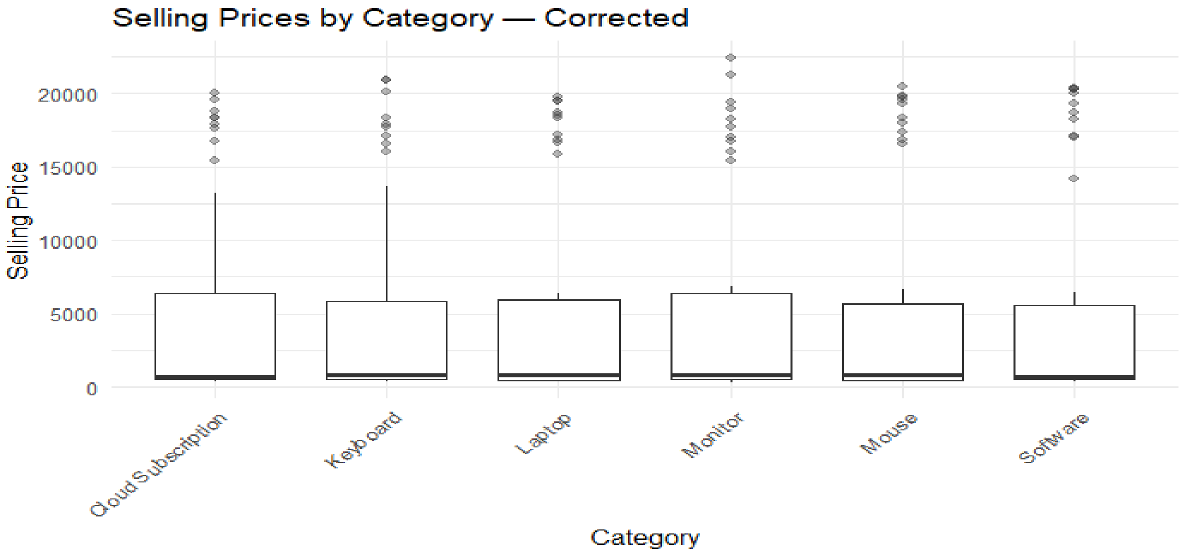


Figure 39

From the plot we can see that the fixed and original selling price still follow the same distribution indicating that the fixes did not have a major influence on the distribution.

Average selling price vs cost

Original:

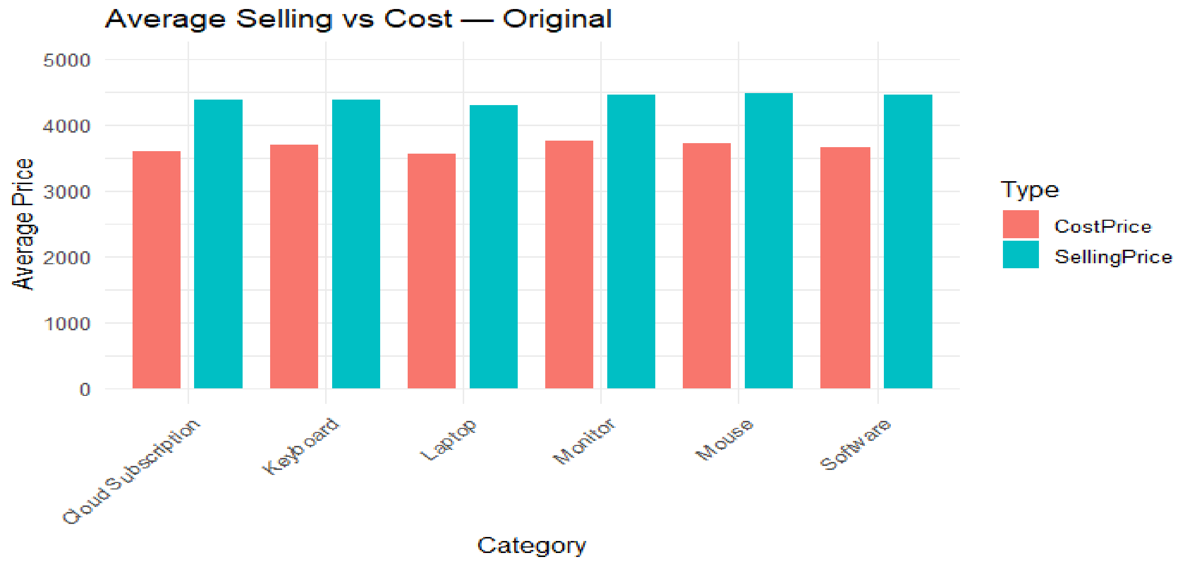


Figure 40

Fixed:

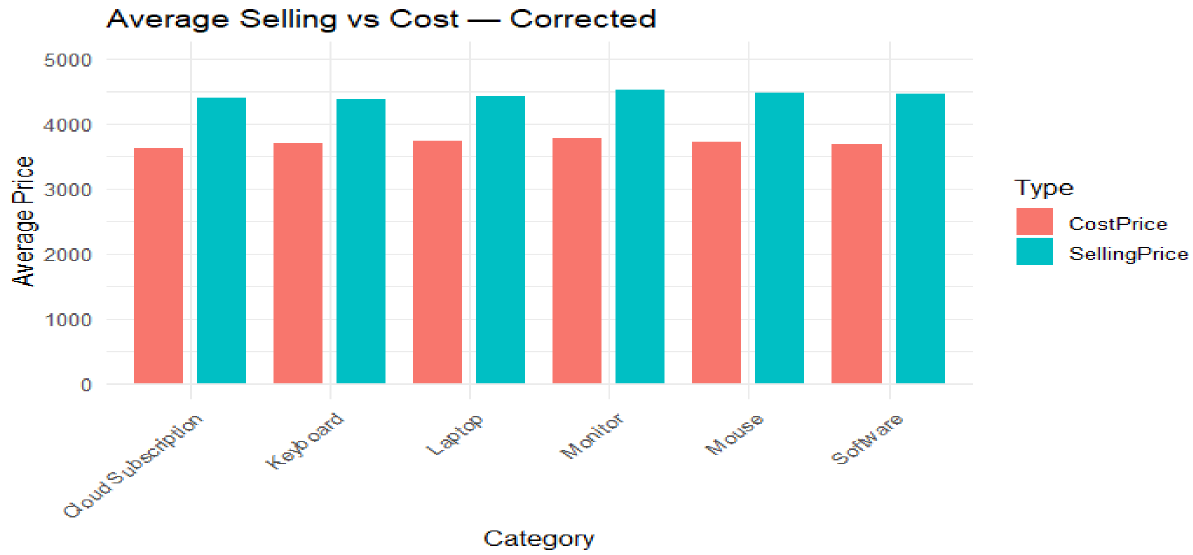


Figure 41

From the plots we can see that the selling price and cost price for both the original and fixed data set have consistent gaps but we can also see that the fixed data sets’ bars appear to be more even and align better, which is an indication that the cost and selling price data is now more accurate and systematic.

Total profit by category

Original:

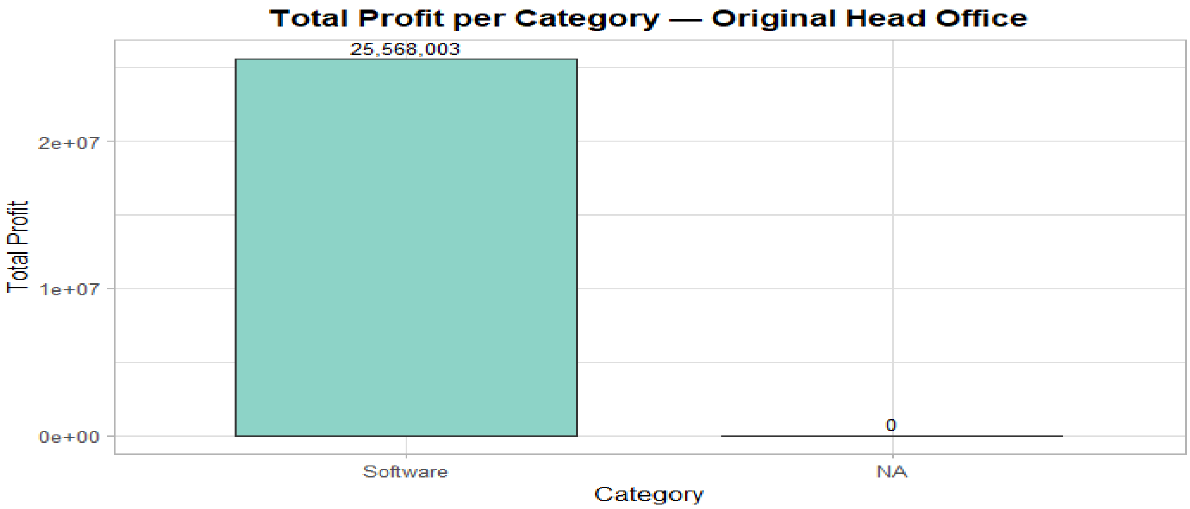


Figure 42

Fixed:

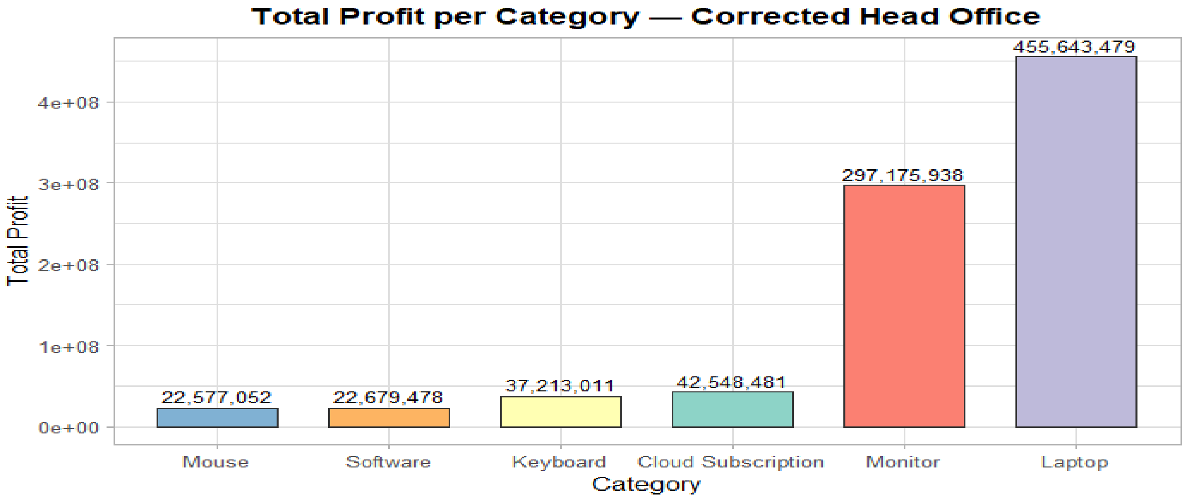


Figure 43

From the plots above we can see that there is a clear data issue with the original data set. It could only find recordings for one product type which is software. This is due to the error in the Product ID column which contained NA values. After fixing this issue we can see that we get a far more accurate and realistic output. We can now clearly see the profits related to each product type. This indicated that the data fixing process was successful and the data is now updated and safe for use.

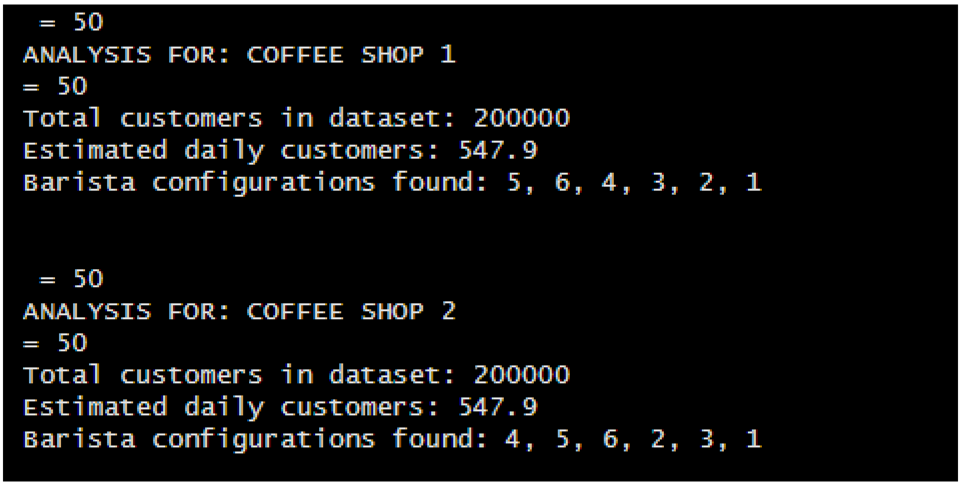
Part 5.

Optimizing Profit

For this part we were given two data sets timeToServe.csv and timeToServe2.csv. Each of these two csv's contain two columns called V1 and V2 with V1 representing the number of baristas working at that instance and V2 representing the number of seconds it took to serve the customer. We were asked to build the model suggested by the previous data analyst and use it to optimize the profit by choosing the optimal number of baristas per weekday. I will now go through the steps of how I did this.

Step 1

I firstly loaded the data from the two csv files to r for the modelling. After loading the data I set up some summaries for each of the two csv's. These summaries can be found below where it gives us the total number of customers for the year, the average daily number of customers, and the barista configuration found.

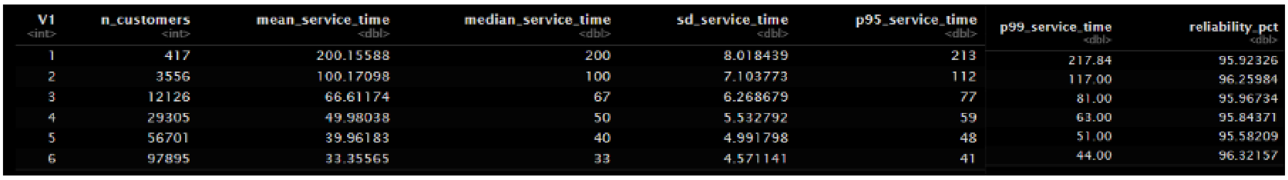


```
 = 50
ANALYSIS FOR: COFFEE SHOP 1
= 50
Total customers in dataset: 200000
Estimated daily customers: 547.9
Barista configurations found: 5, 6, 4, 3, 2, 1


 = 50
ANALYSIS FOR: COFFEE SHOP 2
= 50
Total customers in dataset: 200000
Estimated daily customers: 547.9
Barista configurations found: 4, 5, 6, 2, 3, 1
```

Figure 44

Then for each number of baristas from 2 to 6 I determine the following statistics: the number of customers served, the mean service time, the median service time, the standard deviation of service time, the 95 percentile service time, the 99 percentile service time, and the reliability. The image below shows all these values.

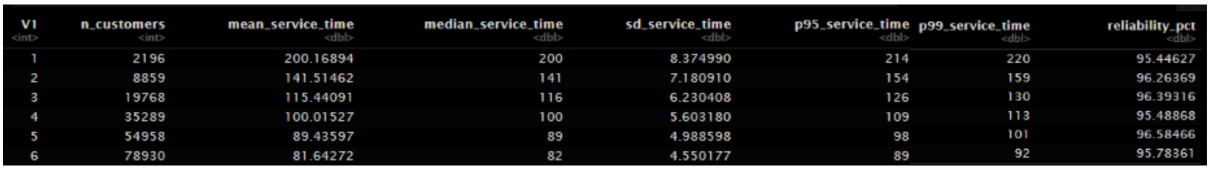
timeToServe:



V1 <int>	n_customers <int>	mean_service_time <dbl>	median_service_time <dbl>	sd_service_time <dbl>	p95_service_time <dbl>	p99_service_time <dbl>	reliability_pct <dbl>
1	417	200.15588	200	8.018439	213	217.84	95.92326
2	3556	100.17098	100	7.103773	112	117.00	96.25984
3	12126	66.61174	67	6.268679	77	81.00	95.96734
4	29305	49.98038	50	5.532792	59	63.00	95.84371
5	56701	39.96183	40	4.991798	48	51.00	95.58209
6	97895	33.35565	33	4.571141	41	44.00	96.32157

Figure 45

timeToServe2:



V1	n_customers	mean_service_time	median_service_time	sd_service_time	p95_service_time	p99_service_time	reliability_pct
<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2196	200.16894	200	8.374990	214	220	95.44627
2	8859	141.51462	141	7.180910	154	159	96.26369
3	19768	115.44091	116	6.230408	126	130	96.39316
4	35289	100.01527	100	5.603180	109	113	95.48868
5	54958	89.43597	89	4.988598	98	101	96.58466
6	78930	81.64272	82	4.550177	89	92	95.78361

Figure 46

Step 2

In this step I visualized the data for each shop. We represented the median service time distribution on box plots, the mean service time for a different number of baristas, and the service reliability associated with each number of baristas.

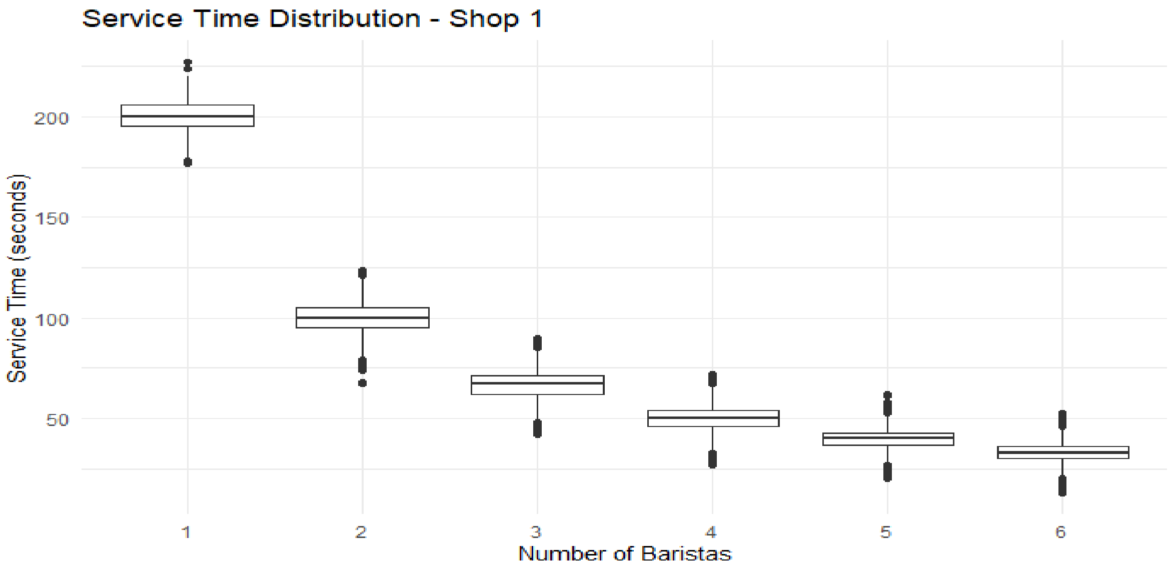


Figure 47

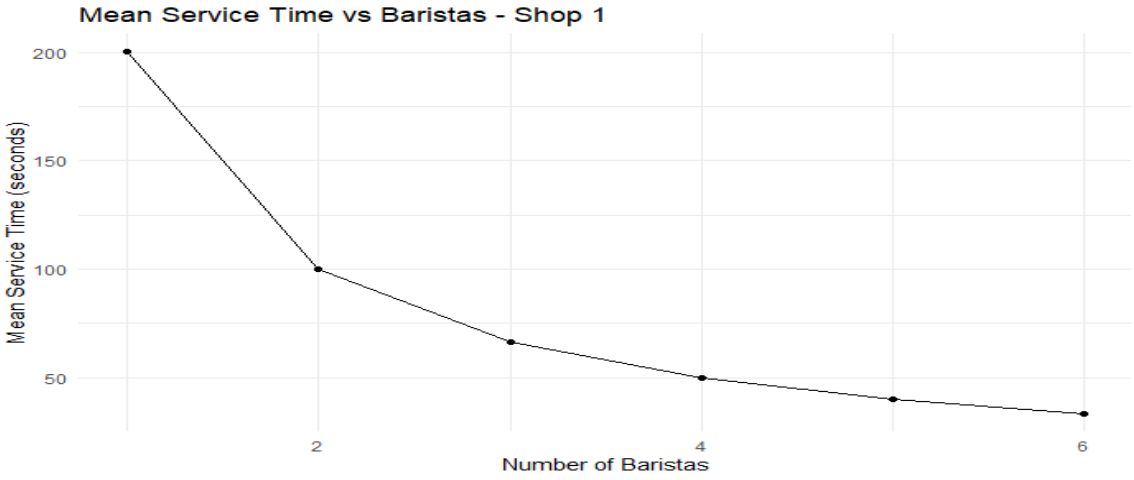


Figure 48

These two plots give us a clear indication of the effect of the number of baristas working on mean service time. The more baristas there are working the smaller the service time mean becomes. Which is a clear indication that the stronger the work force the faster customers can be served.

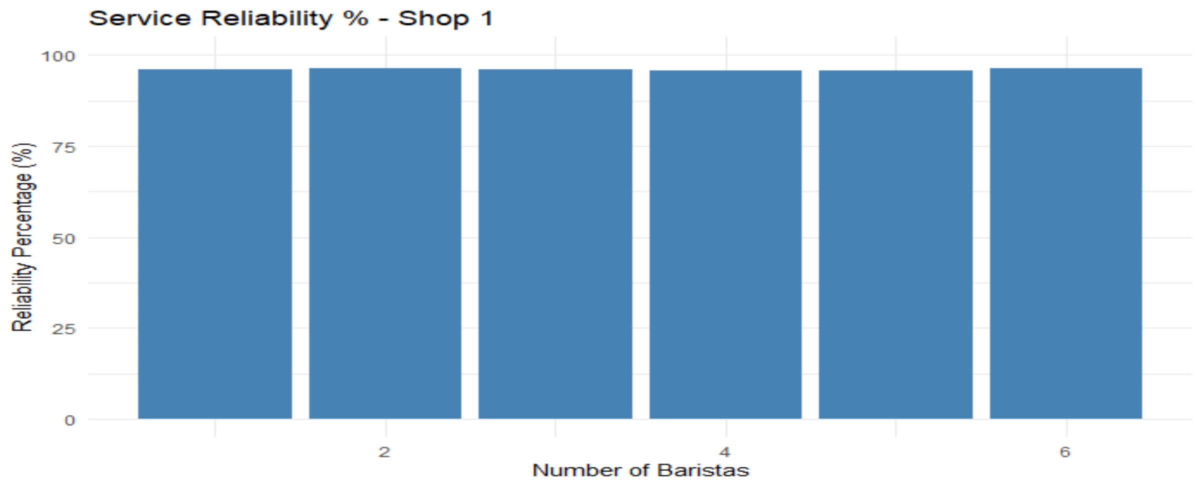


Figure 49

This chart shows the reliability associated with each number of baristas working. From this we can see that the reliability is at its highest when there are six baristas working but all of the reliabilities are very close to each other with around only a 1 % or less difference.

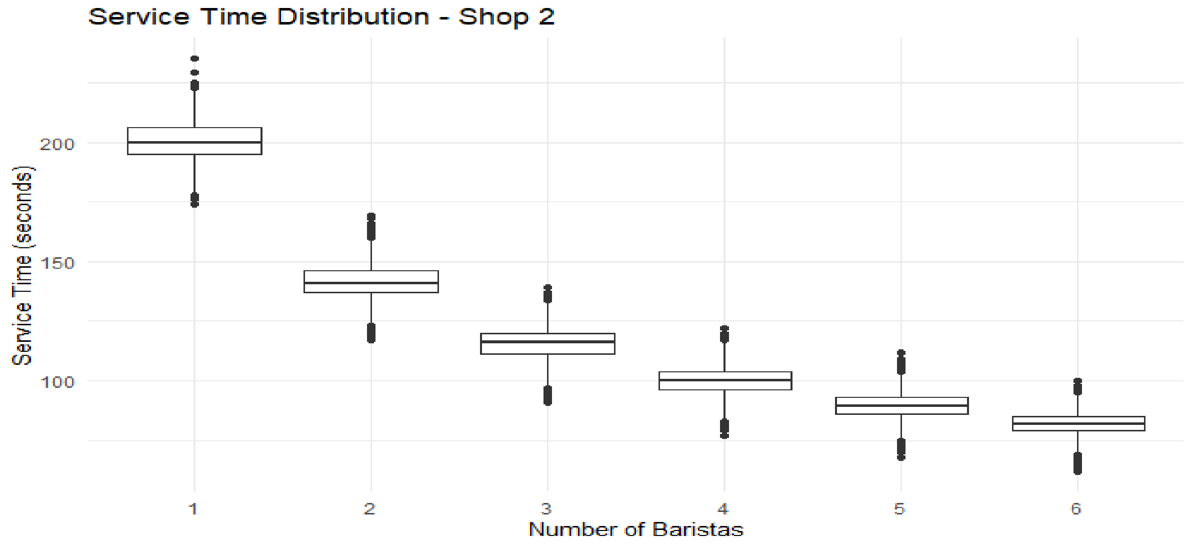


Figure 50

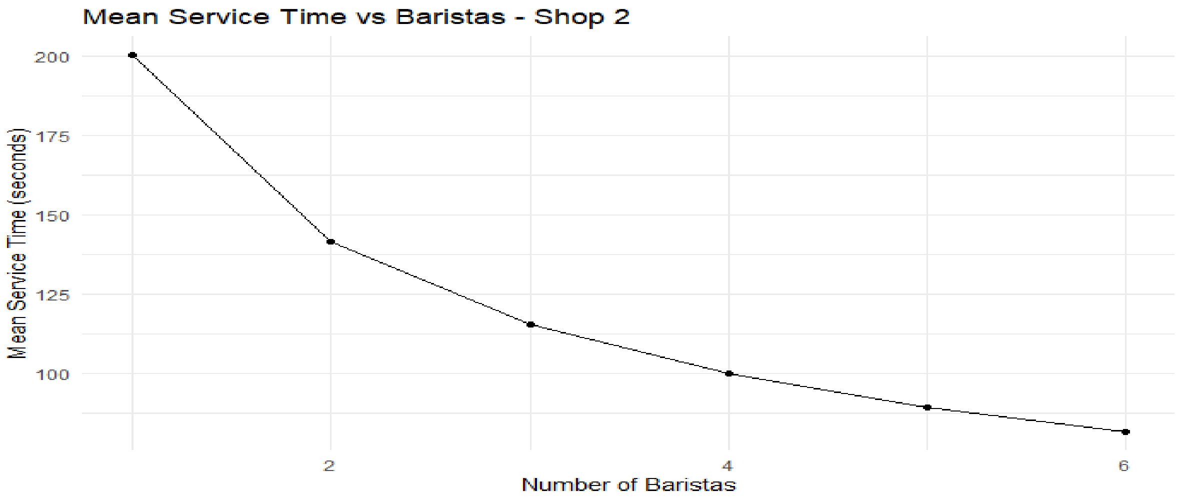


Figure 51

From this plot we can see that the number of baristas working have a significant influence on the mean service time. The mean service time decreases as the number of baristas working increases. This makes because the larger the work force the quicker work can be done.

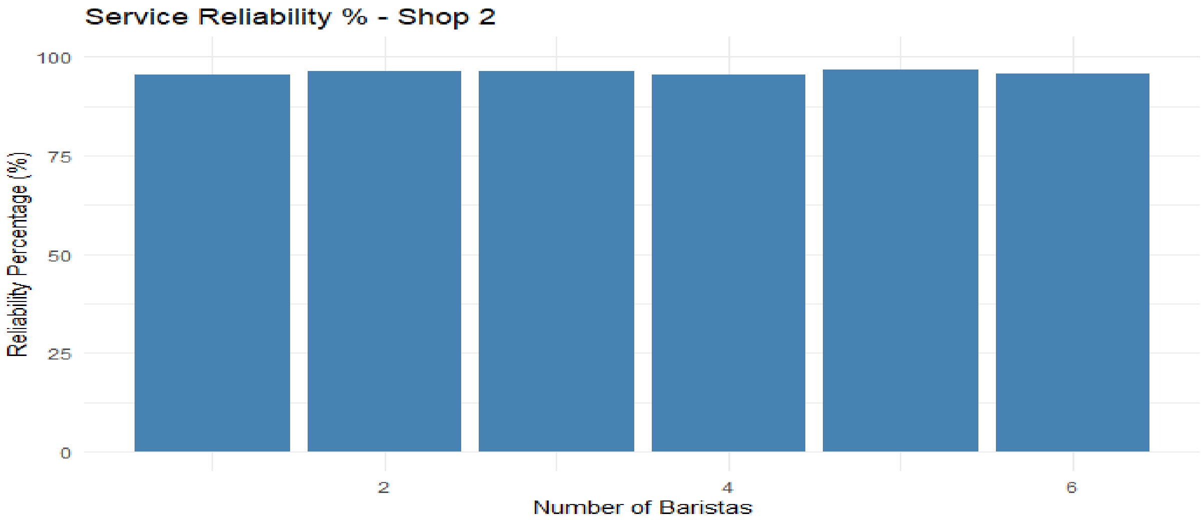


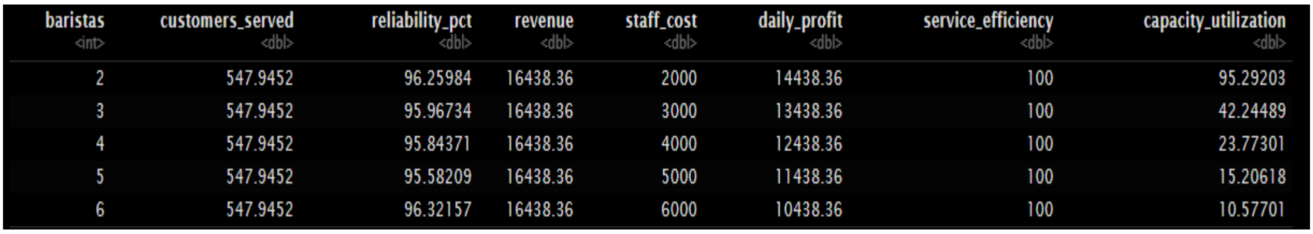
Figure 52

This chart indicates the reliability associated with the number of baristas working. We can see that 5 baristas working yields the highest reliability with again the reliability not changing with about more that 1% for each.

Step 3

I then determined the number of customers served per day, reliability, revenue, staff cost, daily profit, service efficiency, and capacity utilization of each shop with their associated number of baristas working. This is represented by the image below.

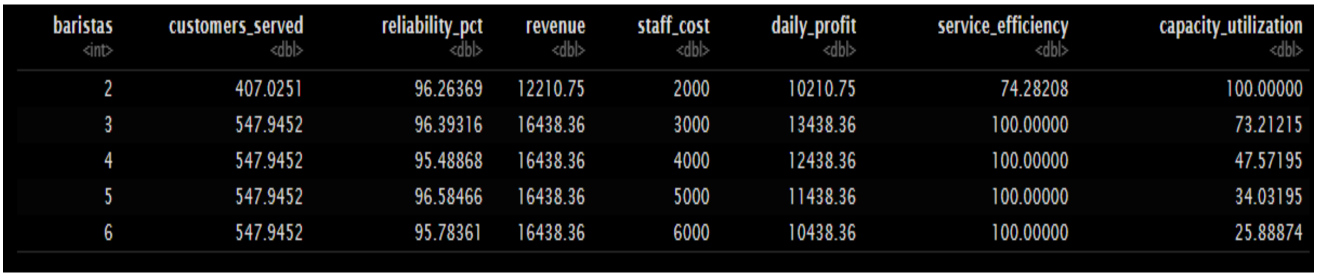
Shop 1:



baristas <int>	customers_served <dbl>	reliability_pct <dbl>	revenue <dbl>	staff_cost <dbl>	daily_profit <dbl>	service_efficiency <dbl>	capacity_utilization <dbl>
2	547.9452	96.25984	16438.36	2000	14438.36	100	95.29203
3	547.9452	95.96734	16438.36	3000	13438.36	100	42.24489
4	547.9452	95.84371	16438.36	4000	12438.36	100	23.77301
5	547.9452	95.58209	16438.36	5000	11438.36	100	15.20618
6	547.9452	96.32157	16438.36	6000	10438.36	100	10.57701

Figure 53

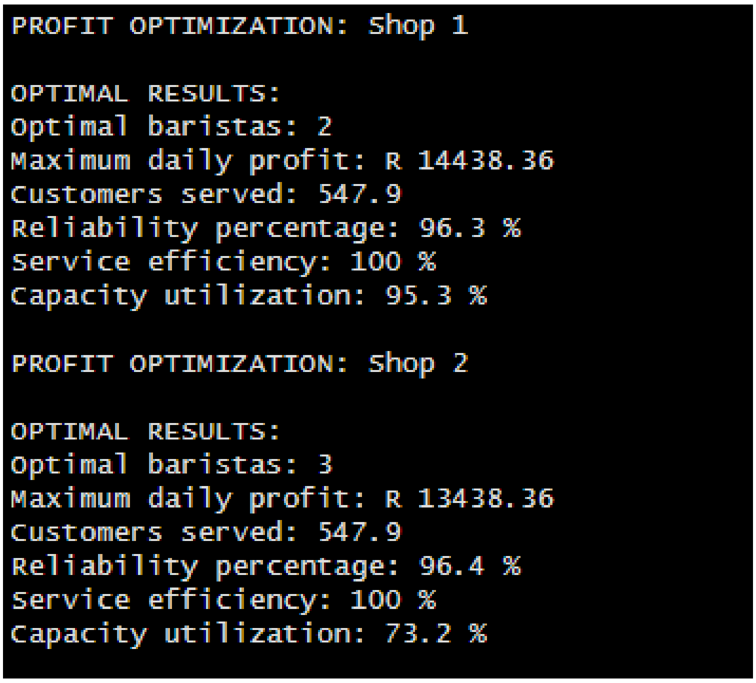
Shop 2:



baristas <int>	customers_served <dbl>	reliability_pct <dbl>	revenue <dbl>	staff_cost <dbl>	daily_profit <dbl>	service_efficiency <dbl>	capacity_utilization <dbl>
2	407.0251	96.26369	12210.75	2000	10210.75	74.28208	100.00000
3	547.9452	96.39316	16438.36	3000	13438.36	100.00000	73.21215
4	547.9452	95.48868	16438.36	4000	12438.36	100.00000	47.57195
5	547.9452	96.58466	16438.36	5000	11438.36	100.00000	34.03195
6	547.9452	95.78361	16438.36	6000	10438.36	100.00000	25.88874

Figure 54

I then used the model to determine the optimized number of baristas for each shop based on the above-mentioned factors. The model gave the output in the figure below:



```
PROFIT OPTIMIZATION: Shop 1

OPTIMAL RESULTS:
Optimal baristas: 2
Maximum daily profit: R 14438.36
Customers served: 547.9
Reliability percentage: 96.3 %
Service efficiency: 100 %
Capacity utilization: 95.3 %

PROFIT OPTIMIZATION: Shop 2

OPTIMAL RESULTS:
Optimal baristas: 3
Maximum daily profit: R 13438.36
Customers served: 547.9
Reliability percentage: 96.4 %
Service efficiency: 100 %
Capacity utilization: 73.2 %
```

Figure 55

We can see that the model I created determined that the optimal number of barista workers for shop one is 2 and for shop 2 it is 3.

Part 6

Introduction

DOE (Design of Experiments) is an approached used to determine how different factors influence a process outcome. For this DOE we used ANOVA to assess whether delivery process performance differed across years and products types.

This was done using the sales2026and2027.csv data set which contained transaction level information like, customer ID, product ID, quantities, order dates, and time based performance measures like pickingHours and delivery hours.

Aim of the analysis

- We determined whether delivery performance changed between years 2026 and 2027.
- Determined whether delivery hours differ between different product types.
- Determined whether the effect of year depends on product type.

Experimental design and hypothesis

I employed a two factorial design with the following structure.

Factor	**Description**	**Levels**
Year	Order year	2026 and 2027
Product type	Product category (First three letters of product ID)	CLO, KEY, LAP, MON, MOU, SOF

We set the response variable to be deliveryHours, representing the time between when a order was placed and when the order was successfully delivered.

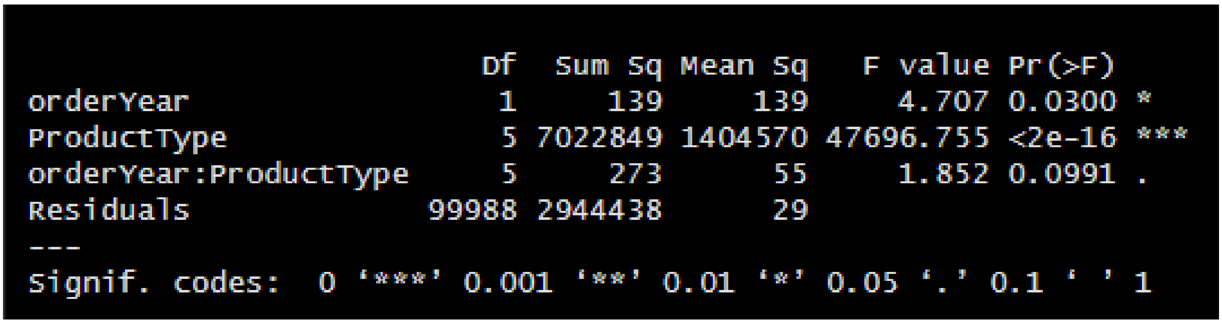
The model was as follows:

Hypothesis	**NULL (Ho)**	**Alternative (Ha)**
Year effect	u(2026) = u(2027)	u(2026) =/ u(2027)
Product type effect	u(CLO) = u(KEY) = u(LAP) = u(MON) = u(MOU) = u(SOF)	At least one product type u differs form the rest.
Interaction	No interaction between year and product type.	Significant interaction exists.

I used a significance level of 0.05 / 5%

Data preparation.

I first checked the data for completeness and structure. I derived product type from product ID’s using the first three characters as reference. Only valid finite delivery hours data was used. I treated order year and product type as categorical features. I used a sample that include more that 100 000 observations.



```
                             Df  Sum Sq Mean Sq   F value Pr(>F)    
orderYear                     1     139     139     4.707 0.0300 *  
ProductType                   5 7022849 1404570 47696.755 <2e-16 ***
orderYear:ProductType         5     273      55     1.852 0.0991 .  
Residuals                 99988 2944438      29                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

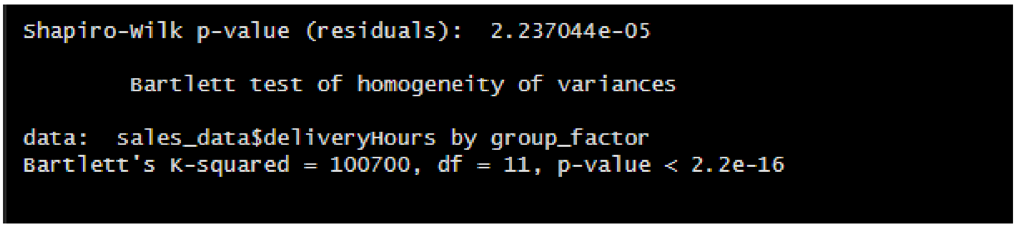
Figure 56

Interpretation

- Main effect of year:
  We can see that there is statistically significant difference of delivery hours between the two years. This is a indication of year to year improvement which can be a sign of better logistics coordination or process refinement in 2027.
- Main effect of product type:
  We can see that product type is highly significant. This is a indication that delivery hour differ a lot across different product types.
- Interaction (year vs product type):
  Since the interaction term was not statistically significant, all product kinds showed a similar pattern of annual change. It seems that operational enhancements have been consistent rather than product-specific.

Diagnostics check

To validate my ANOVA assumptions, residuals plots, Shapiro Wilk test, and Barlett’s test were used



```
Shapiro-Wilk p-value (residuals):  2.237044e-05

        Bartlett test of homogeneity of variances

data:  sales_data$deliveryHours by group_factor
Bartlett's K-squared = 100700, df = 11, p-value < 2.2e-16
```

Figure 57

Test	Static / P-value	Interpretation
Shapiro-Wilk (normally)	p = 2.24 * 10^-5	Residuals deviate from perfect normality, but the large sample size makes this negligible.

Barlett’s test (homogeneity)	p < 2.2 * 10^-16	Variance inequality detected, but robustness maintained due to balanced sample size.

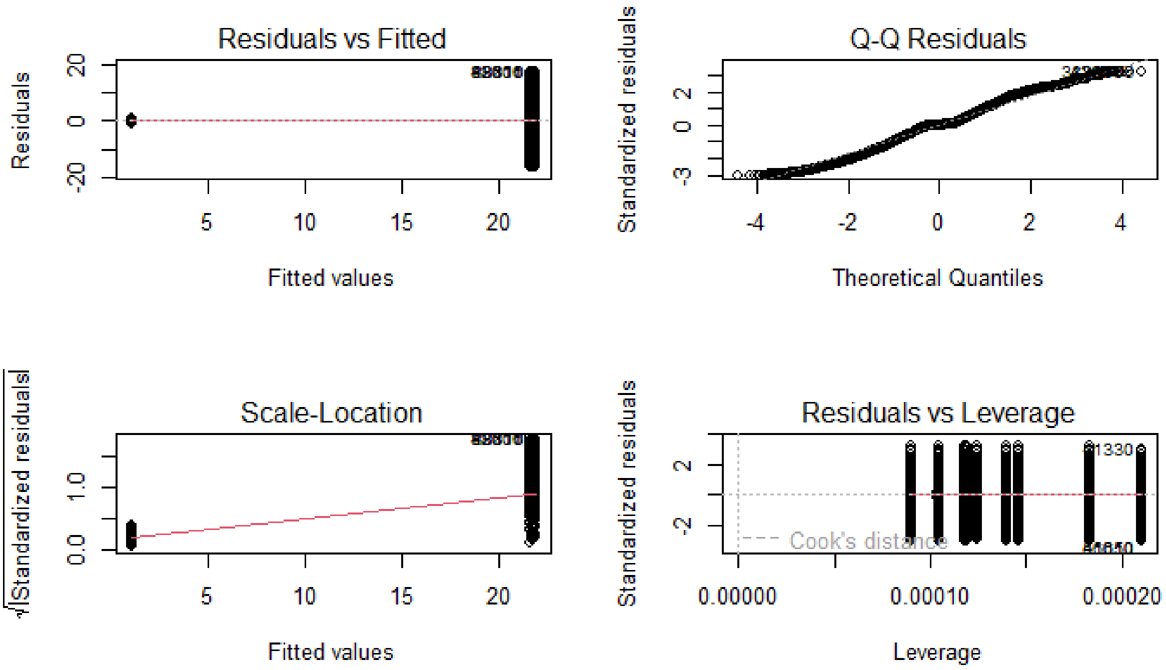


Figure 58

From the graphs above we can see the following:

- The Q-Q plot:
  This plot show moderate linearity and a slight tail deviation.
- Residual vs Fitted and scale location plots:
  From this plots we can see that there is a relative variance across fitted values and have no heteroscedastic trend.
- Residual vs leverage plot:
  From this plot we can see that there are no influential outliers.

From this we can see that the model assumptions are reasonably satisfied for partial interference.

Graphical analysis

Find below the set of visual support for the statistical findings.

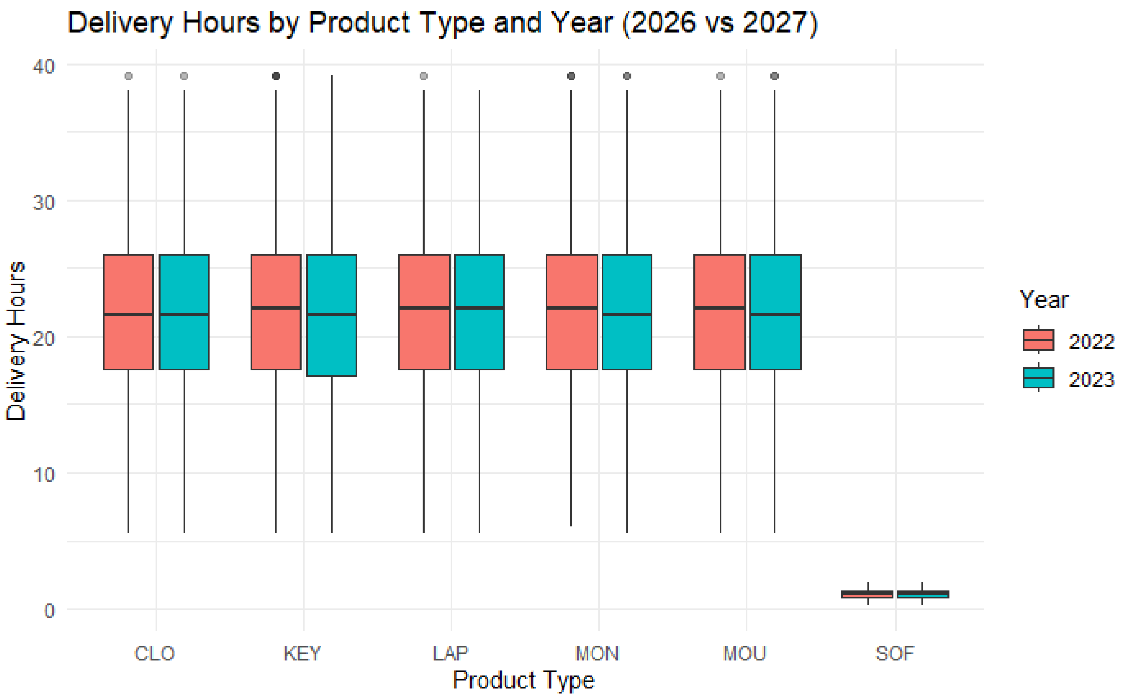


Figure 59

We can see that delivery hours are almost parallel for each product across the two years. We can also see that there is a significant lower demand for software products compared to all of the other products.

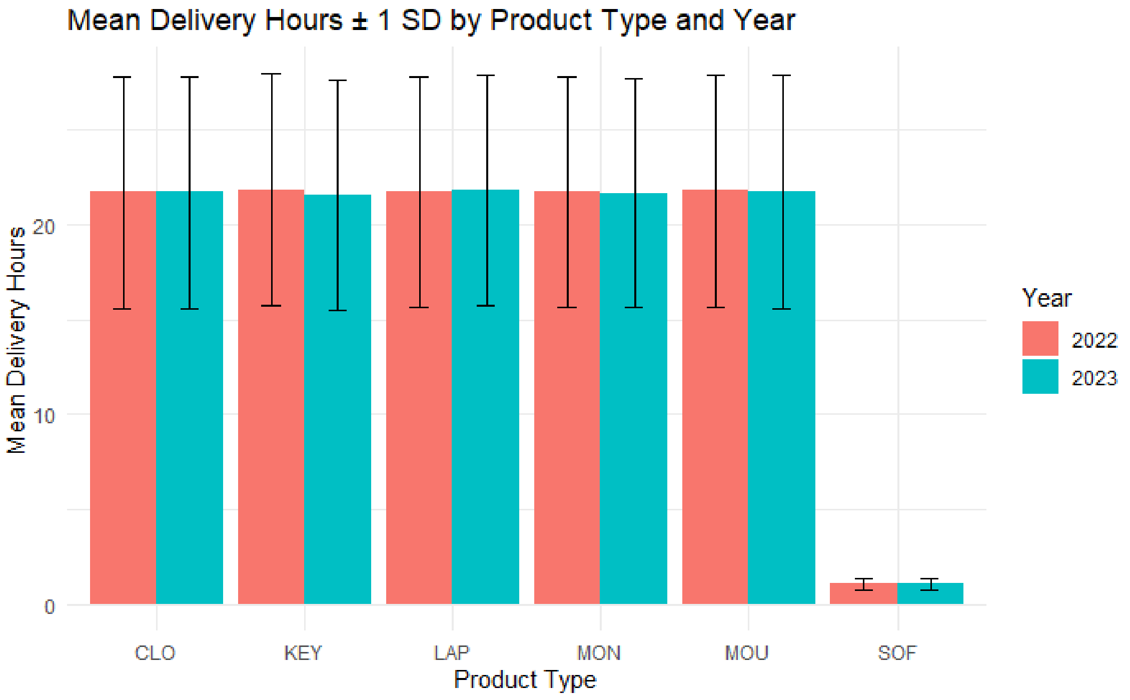


Figure 60

We can see that the mean of delivery hours across the years differ slightly and have overlapping error bars which is a indication of only moderate year effects.

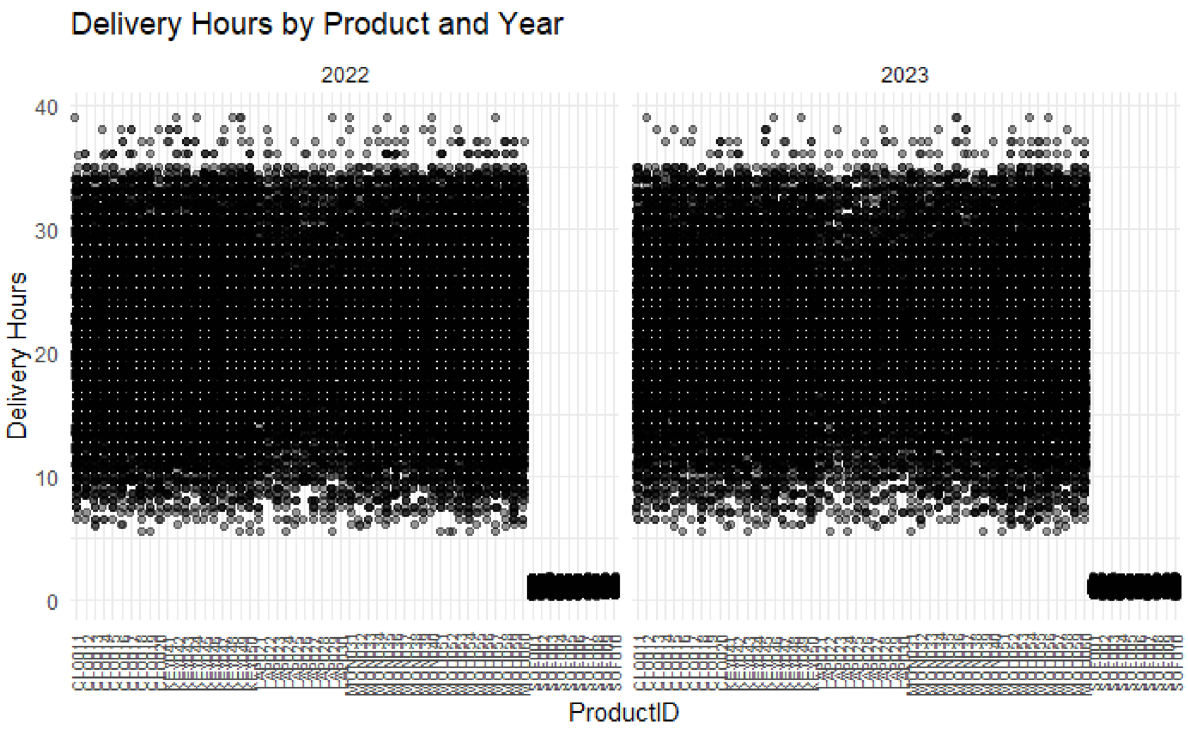


Figure 61

From this plot we can see consistent delivery times across products.

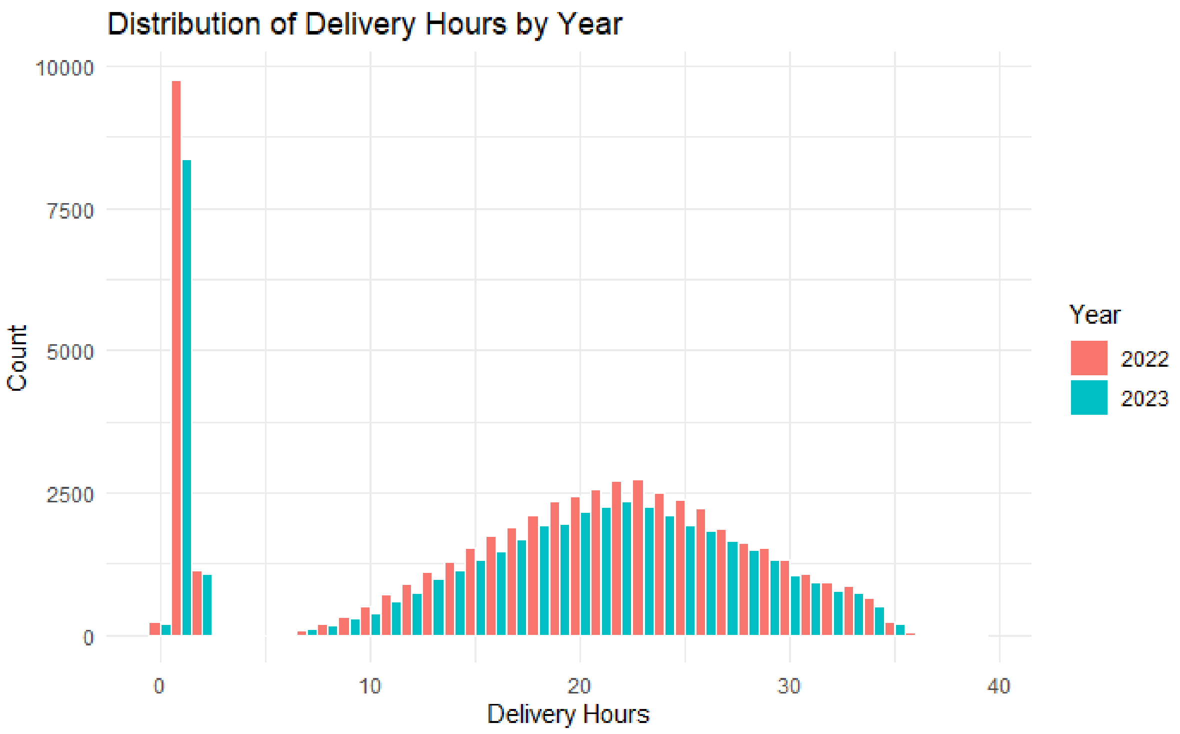


Figure 62

We can see from the plot that the two years produce almost a similar shape which confirms similar distribution behavior.

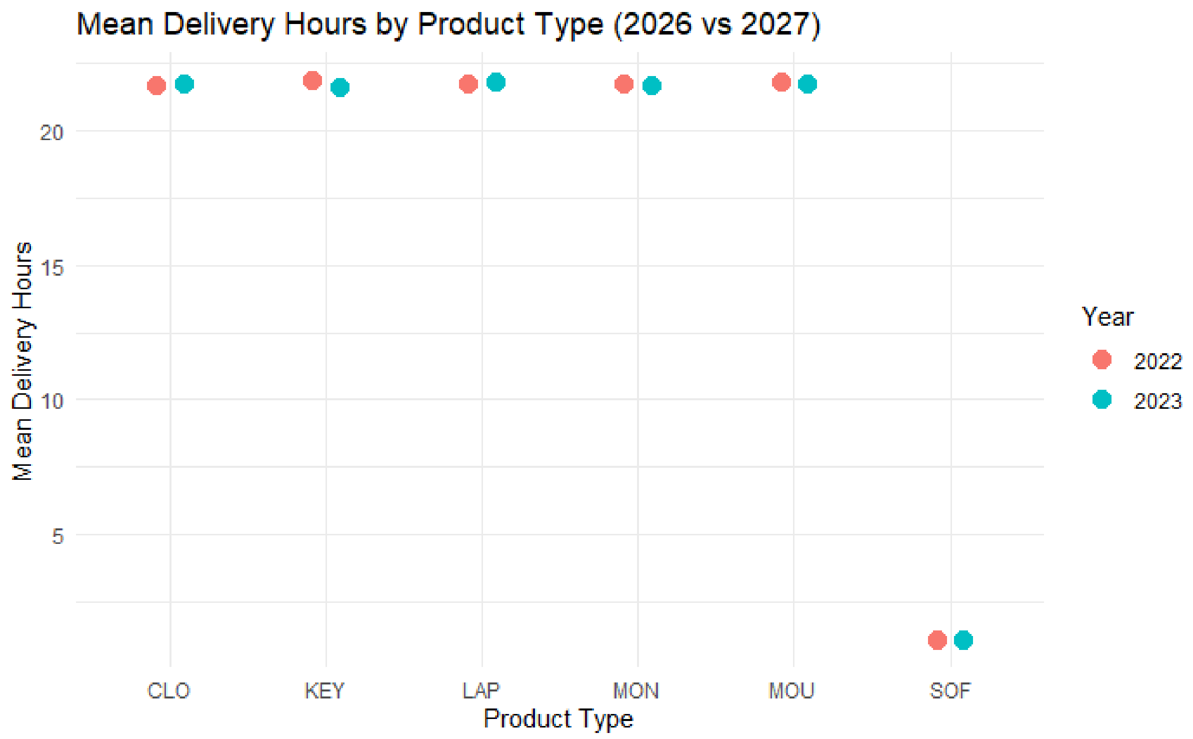


Figure 63

From this plot we can confirm the consistent rank orders across products with CLO at the highest and SOF at the lowest.

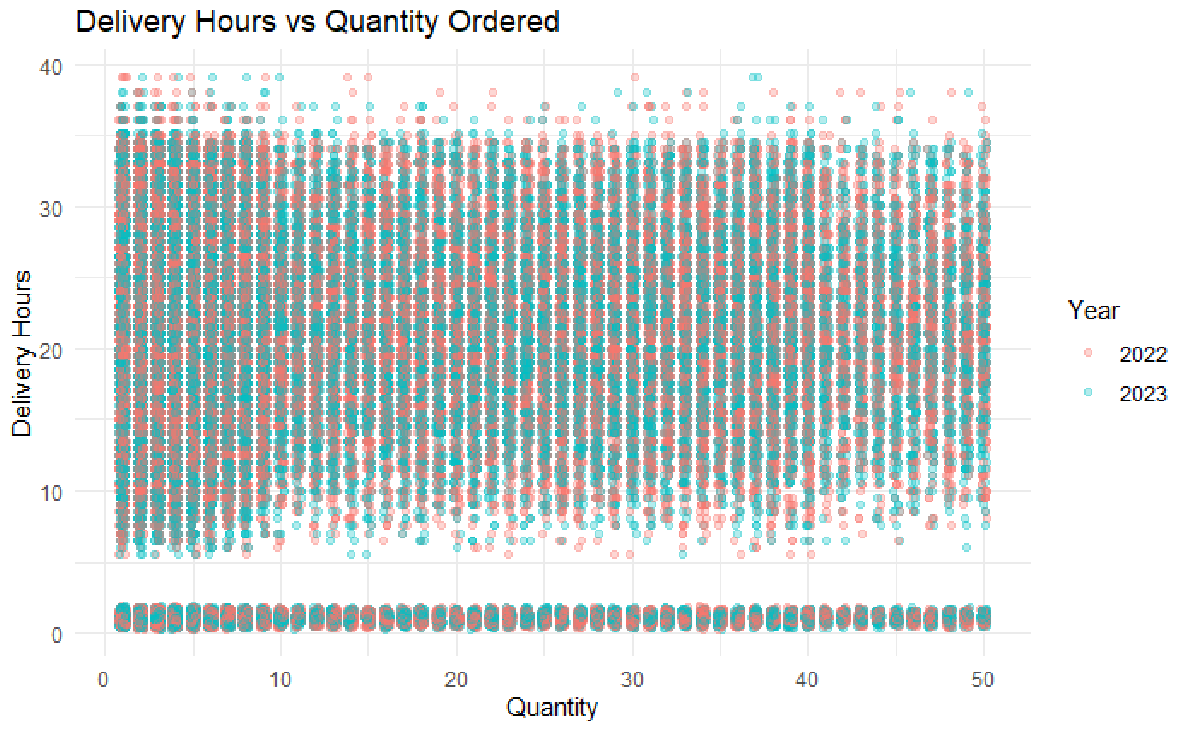


Figure 64

There is no visible relationship between order size and delivery hours which is an indication that delivery hours are independent of order size.

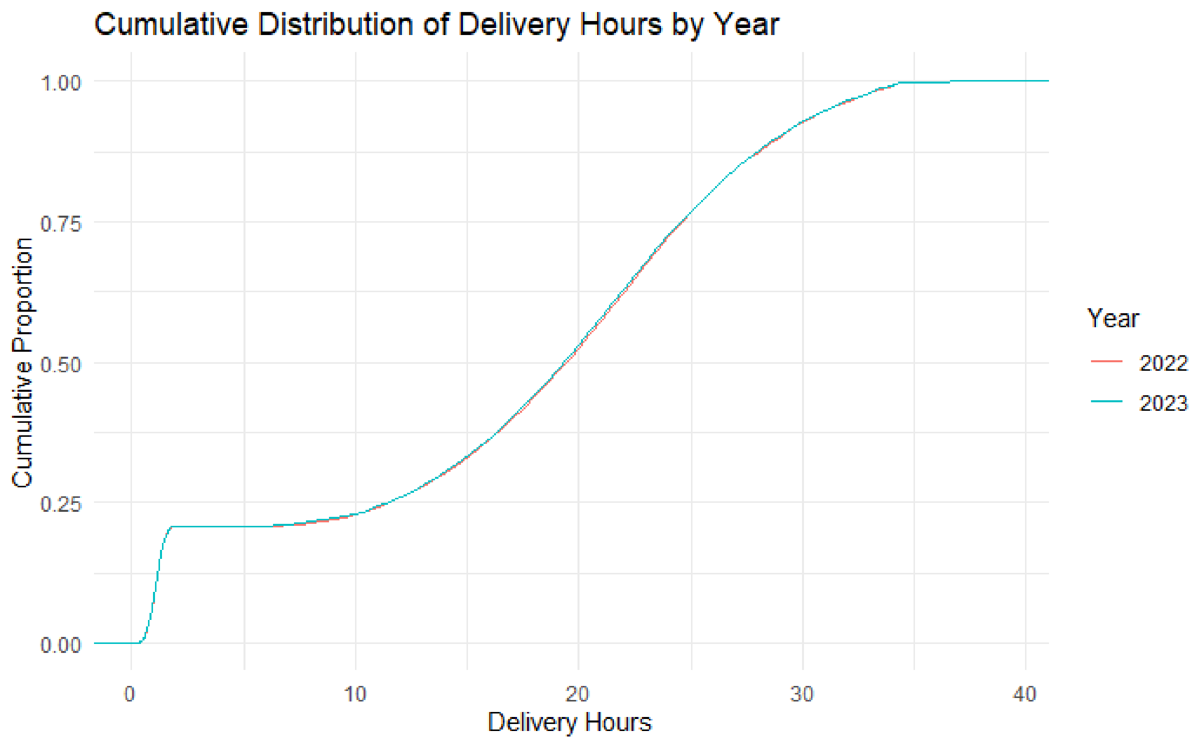


Figure 65

We can see that the ECDF's are not over lapping which confirms that the underlying process distribution remained consistent between the years.

Outcome

After the analysis was done I could confirm the following:

- Delivery performance is slightly yet significantly impacted by the year, suggesting that overall process improvement will occur in 2027.
- Large and consistent variations in product types are a reflection of the processing timeframes or intrinsic logistical complexity of each product group.
- The absence of interaction implies consistent operational advancement in every category.

The DOE and ANOVA analyses show that year and product type have an impact on delivery performance, with moderate improvement across years and substantial evidence of product-specific time differences. Consistent efficiency gains across all categories are implied by the non-significant interaction. The delivery process is steady, predictable, and getting better, according to both statistical and graphic proof. SOF items have especially quick handling times. By validating significant multivariate differences in overall process performance, the MANOVA results support these findings.

Part 7

Expected number of reliable days per year

We see that the data given was taken over a 397 day period. We know that reliable days consist of days with 15 or 16 staff members.

Therefore the reliable proportion over the 397 day period is:

Reliable proportion = (96+270)/397 = 0.9219

And then reliable days per year will be:

Reliable days per year = 0.9219*365 = 337 day

Non reliable days per year = 365 – 337 = 28 days.

Profit optimization

Model setup

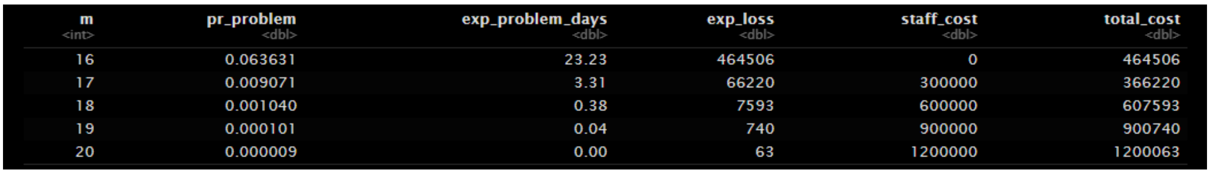
We let X ~ bin(m,p) be the number of workers present when m are employed and each show up independently with probability p.

From the chart given to us we can get the mean of employees on duty to be 15.58 and therefore p = 15.58/16 = 0.9740.

A problem day will occur if X<15. Then the expected annual cost at a staffing level of m will be:

C(m) = 365 * Pr(X < 15 | m,p) * 2 000 + max(0, m-6) * 300 000

See the model's result from R below:



```
    m        pr_problem        exp_problem_days     exp_loss     staff_cost     total_cost
<int>             <dbl>                   <dbl>        <dbl>          <dbl>          <dbl>
   16          0.063631                   23.23       464506              0         464506
   17          0.009071                    3.31        66220         300000         366220
   18          0.001040                    0.38         7593         600000         607593
   19          0.000101                    0.04          740         900000         900740
   20          0.000009                    0.00           63        1200000        1200063
```

Figure 66

Final recommendation

Based on the model's results, I give the following recommendations:

- An additional worker should be hired. So the company should go from 16 to 17 workers. This will minimize the total annual cost from R 464 506 to R 366 220.
- It will not benefit the company to hire more than 1 additional worker as the staff cost dominate the small additional reliability gain.

References

- Datacamp.com. (2023). *Data Analysis Cheat Sheets | Learn & Grow Your Data Analysis Skills*. [online] Available at: https://www.datacamp.com/cheat-sheet/category/data-analysis [Accessed 05 Oct. 2025].

- Bright Data. (n.d.). Guide to Data Wrangling. [online] Available at: https://www.google.com/aclk?sa=l&ai=DChcSEwjjmayu8d36AhVY4O0KHZ3aA9UYABAAGg JkZw&sig=AOD64_1V2lRa1YeLvVsGuqTFFKCt1mDa5Q&q&adurl&ved=2ahUKEwj5uKau8d 36AhXObMAKHcUgApAQ0Qx6BAgHEAE [Accessed 07 Oct. 2025].

- www.sthda.com. (n.d.). *MANOVA Test in R: Multivariate Analysis of Variance - Easy Guides - Wiki - STHDA*. [online] Available at: http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance.

- Shen, S. (2021) 7 steps to ensure and sustain data quality, Medium. Towards Data Science. Available at: https://towardsdatascience.com/7-steps-to-ensure-and-sustain-data-quality3c0040591366 (Accessed: October 15, 2025).

- Trochim, P.W.M.K. (2022) Descriptive statistics, Research Methods Knowledge. Base. Conjointly. Available at: https://conjointly.com/kb/descriptive-statistics/ (Accessed: October 06, 2025).