Project ECSA GA4: Data Analysis and Quality Control

HUGO RADLEY\Student Number: 27377679

25 October, 2025

Contents

1 Introduction

The purpose of this report is to do a full data analysis and quality control on a data set received from the ECSA GA4 project. The data sets integrates data information on customers, products, and sales. The data analysis approach follows a clear structure of cleaning the data, processing the data, modeling and analyzing the models created. This help to improve the business and engineering decision available. All analysis was done using R and R markdown to ensure accurate calculations, a neat format, and professional documentation.

2 Part 1: Initial Data Analysis

Part 1 focuses on creating a data analysis of the csv files provided that is clean, reliable and informative. The data analysis consists of 3 components:
1.Data wrangling & preparation - verify data, correct misclassifications, ensure consistency in data.
2.Descriptive &exploratory analysis – summarize the core variables to identify trends in the data set.
3.Preliminary relationship assessment – Use visualizations and tables to detect patterns between the customers, products, and other metrics.
The data sets are then validated and understood with distributions. This provides a deeper understanding of the data using statistical methods.

2.1 Data Wrangling and Preparation

The data-wrangling steps ensured that the categories products, customers and sales from all the csv files was accurate. It used the product IDs prefixes, e.g. SOF for software, to combine the product, sales and customer data to do full analysis. This helped connecting different features and creating correlations between them.

2.2 Missing Value Check

As a key data quality step, the datasets were checked for missing (NA) values.

Table 1: Count of Missing (NA) Values per Dataset

Dataset	Missing_Values
Customers	0
Products	0
Sales	0

Analysis:
The missing value check found that 0 values was missing. Originally this could have been seen as positive, however in part 4 with the re-analysis it would be seen that this was however not the case. A missing value free data set is very valuable in data analysis due to the fact that it increases the accuracy of the analysis. According to this current data it seems no values were missing and that the data base was full and without missing values.

2.3 Customer Data Analysis

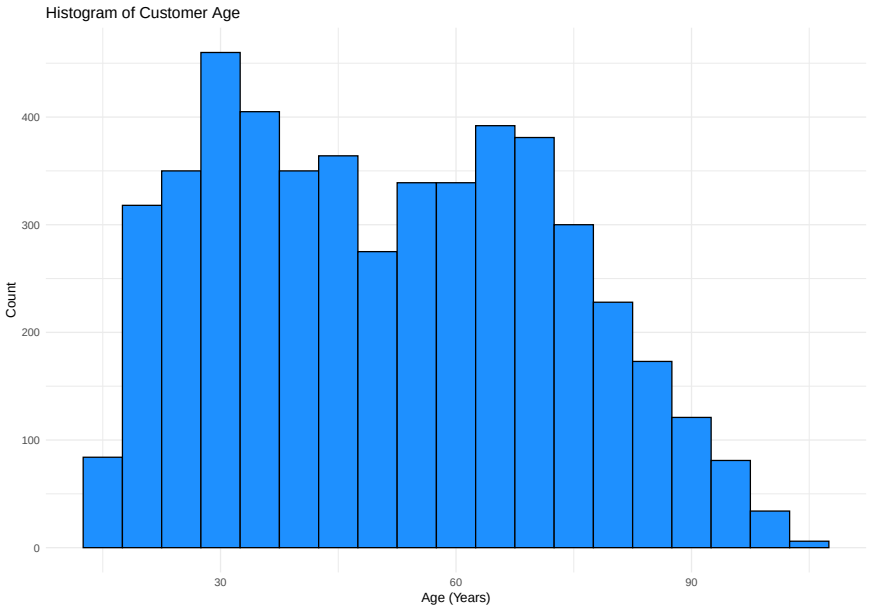
2.3.1 Descriptive Statistics

Table 2: Descriptive Statistics for Customer Age and Income

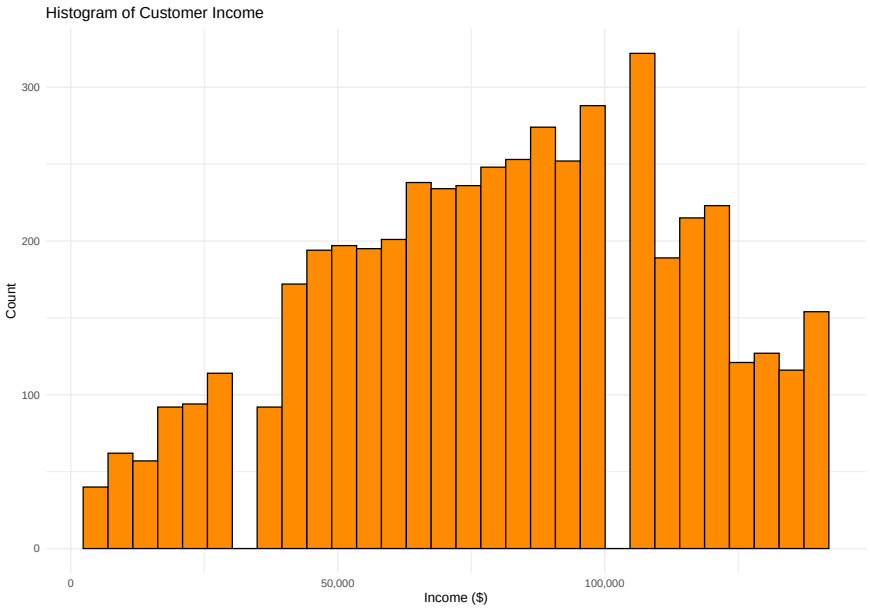
	N	Mean	Std. Dev.	Median	Min	Max	Range
Age	5000	51.55	21.22	51	16	105	89
Income	5000	80797.00	33150.11	85000	5000	140000	135000

Analysis: Descriptive Statistics (Age and Income)
The customers range from age 16 to 105 years old and has a mean of 52.55 years and a standard deviation of 21 years. This shows we are working with a wide variety of ages and not specific age groups that dominate. The diverse group of consumers earns a minimum of $5000 and maximum $140000 and a mean of $80797 which also suggests that we are dealing with a variety of income groups. The group is however skewed a bit to the left suggesting more individuals earn more. The customer base is overall an unbiased group of individuals ranging from all ages and all incomes.

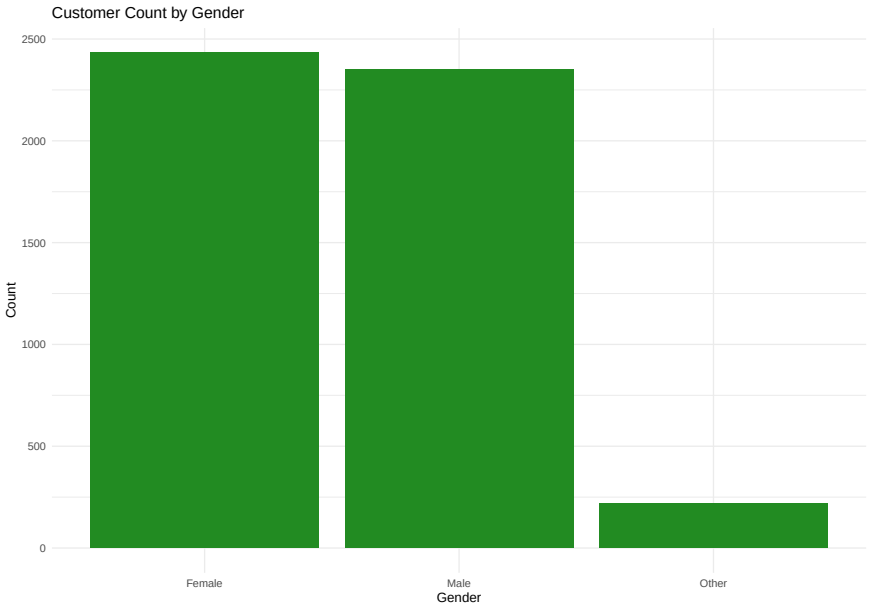
2.3.2 Visualizations



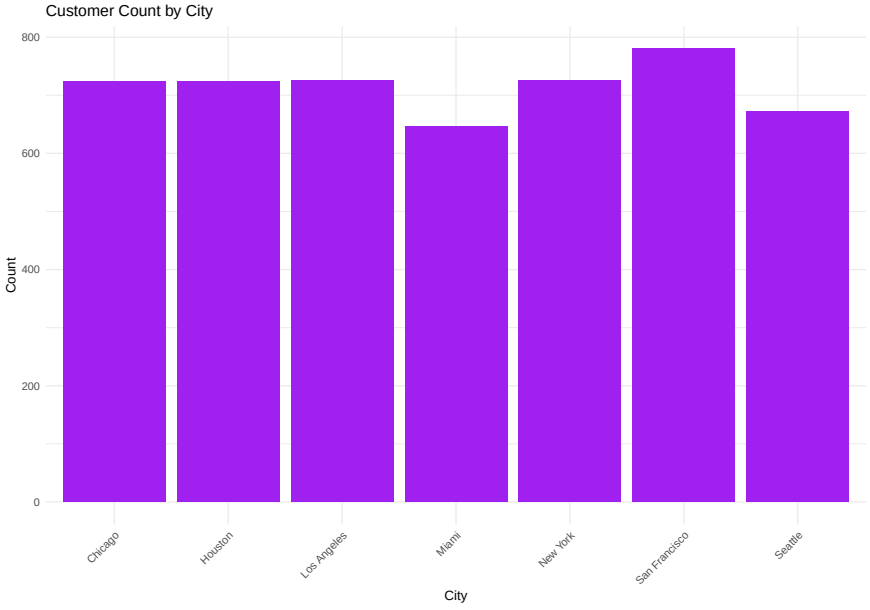
Analysis:Histogram of Customer Age
The bell curved histogram of age is skewed a bit to the right suggesting that the amount of older people gradually decrease while you younger people increase at a much faster tempo. It also suggests that fewer teenagers and elderly buys the products with the main group being the 30-year old group forming majority of the customers.



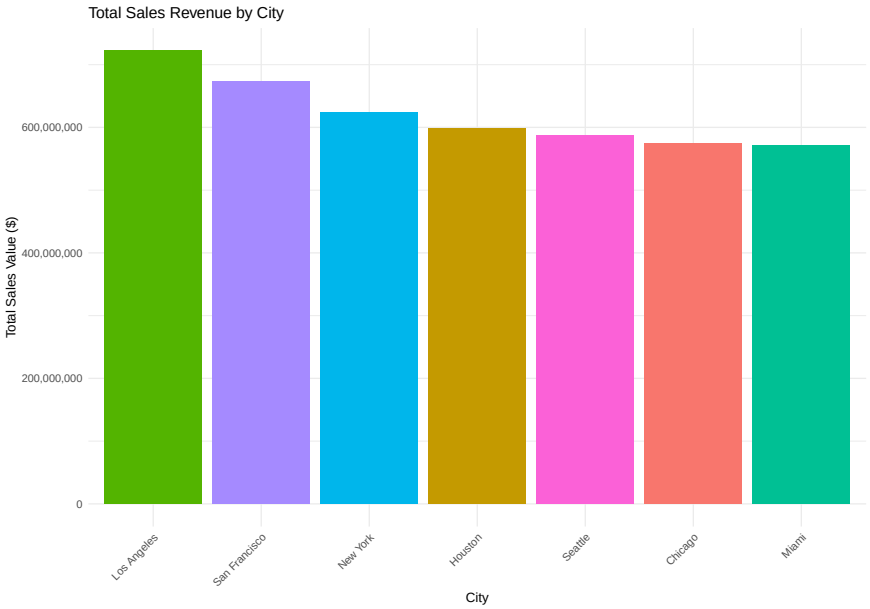
Analysis: Histogram of Customer Income
Ad acknowledged earlier the income count graph is skewed to the left indicating a slow increase in income and then a sudden drop. this indicates that the majority of people earn on the mid-up side and not a lot of customers earn very low or very high incomes. The $120000 category has the most customers.



Analysis: Customer Count by Gender
The gender graph shows that the male and females are relatively similar with rare cases of other. This shows that there are not much of a disparity in customers gender and the product is for all people. Other could be due to customers not wanting to explicitly state their gender or some identifying as something else. Over all the genders are relatively balanced.



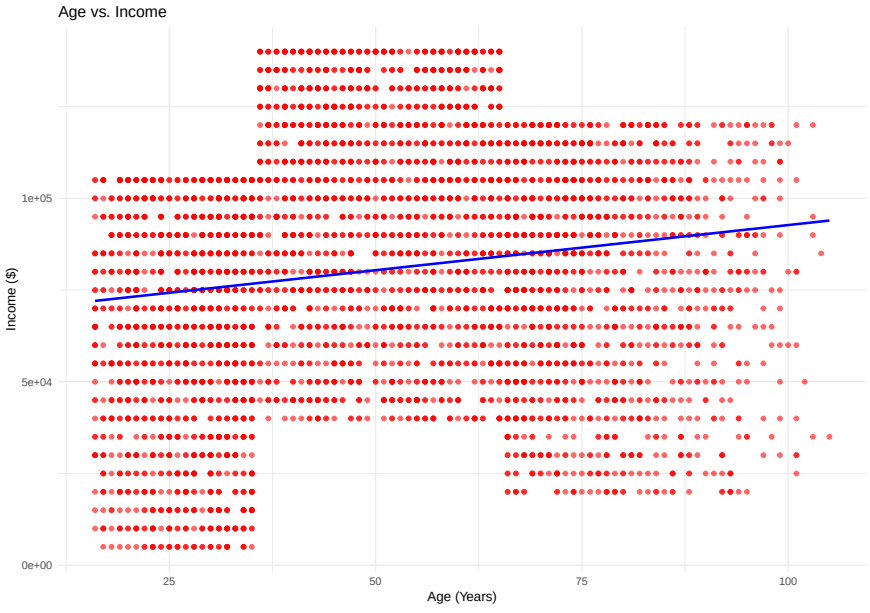
Analysis: Customer Count by City
Customer counts by city is also relatively evenly spread. San Francisco with the highest participation, followed by the cities, New York, Houston, Chicago and Los Angeles. Miami show smaller populations, however it is not far from the highest city. Thus the spread by city is also very even suggesting no bias to either city.



Analysis: Total Sales Revenue by City
Los Angeles leads the sales and Miami trails last. This correlates to the number of customer per city. Urban areas thus places higher in sales mainly due to the higher number of people and could thus contribute to higher ROI values.It is thus worth targeting these areas above rural areas.

Table 3: Pearson Correlation Coefficient for Age and Income

	Age	Income
Age	1.000	0.158
Income	0.158	1.000



Analysis:Age vs Income Scatter and Correlation
The scatterplot suggests no real correlation between age and income. The only ages that stands out is 38

to 63 have no records of low incomes and also have records of much higher incomes than younger and older people. This graph shows that income are influenced more by other factors than by age.

Customer Data Summary Conclusion:
Overall the customers are spread out over most age and income and centered in specific cities but relatively the same for all. the age-income has a low correlation and should not be used to conduct analysises.

2.4 Products Data Analysis

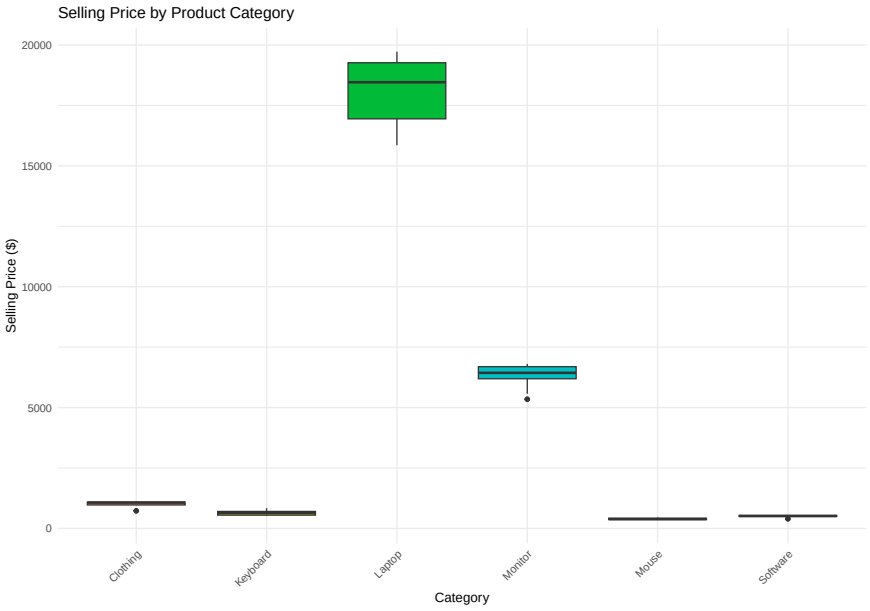
2.4.1 Summary Statistics

Table 4: Summary Statistics by Product Category

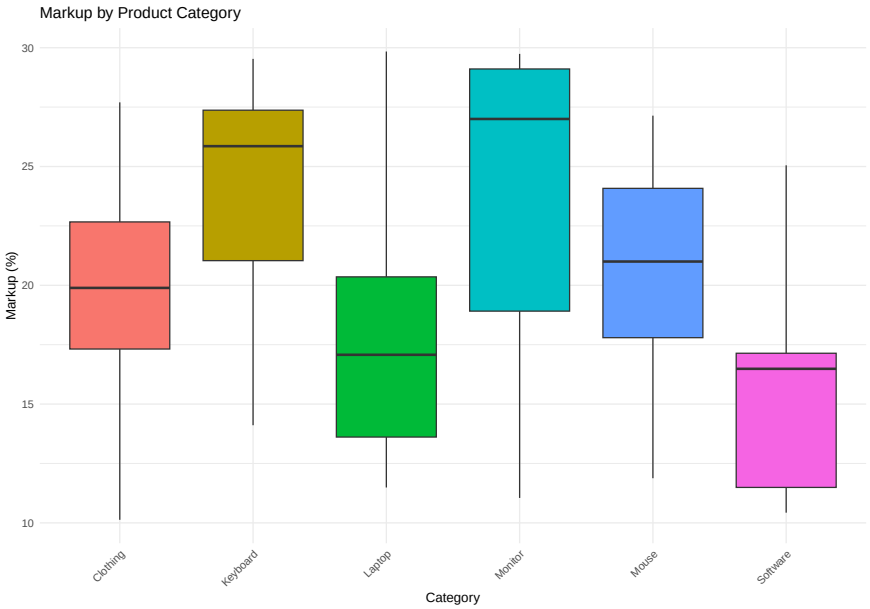
Category	N	Mean Price	SD Price	Mean Markup	SD Markup
Clothing	10	1019.06	118.32	19.96	4.98
Keyboard	10	644.66	107.23	23.98	5.12
Laptop	10	18086.43	1357.43	18.43	6.71
Monitor	10	6310.52	501.87	23.87	6.71
Mouse	10	394.70	33.84	20.50	4.63
Software	10	506.18	44.47	16.04	5.11

Analysis: Summary Statistics by Category
The product categories differs substantially by all price metrics. Laptops with the highest price ($18086.43) and mouse with the lowest ($394.7). Keyboards however has the largest markup of 23.98%. There is thus a difference in pricing structure and market for all the different products. The hardware components generate more proportional profits than software.

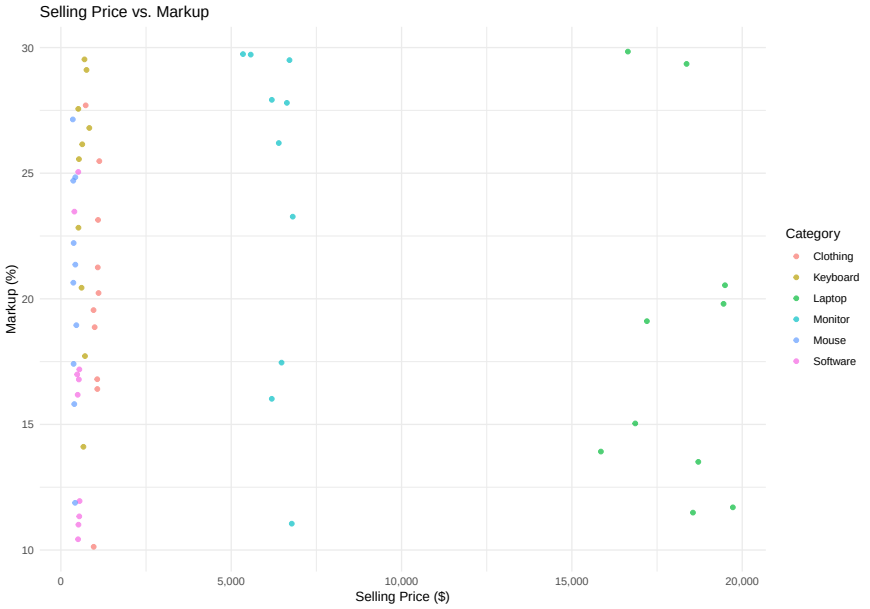
2.4.2 Visualizations



Analysis: Selling Price by Category (Boxplot)
Most products have a low selling price with laptops as the highest selling price, but an interesting observation is also that laptops have a higher variability while all the other products have a much more consistent pricing.



Analysis: Markup by Category (Boxplot)
Markups mostly range within the 15-25% area with the highest markup and highest variability comes from the monitor group. Keyboards might be a better seller as although it has a bit lower markup its minimum markup is higher than monitors minimum markup due to the less variability.



Analysis: Selling Price vs Markup (Scatter)
The scatterplot shows that there are no strong correlation between the markup % and the selling price. The graph also shows the variability in both price and markup % of laptops and monitors.

Analysis: Product Data Summary Conclusion
The product analysis therefor shows that the different products differ in most aspects ranging from price, markup % and variability in these features. Consistent markups are better for planning and pricing control. The portfolio of the company is in tact with high value low volume products in combination with low value high volume products.

2.5 Sales Data Analysis

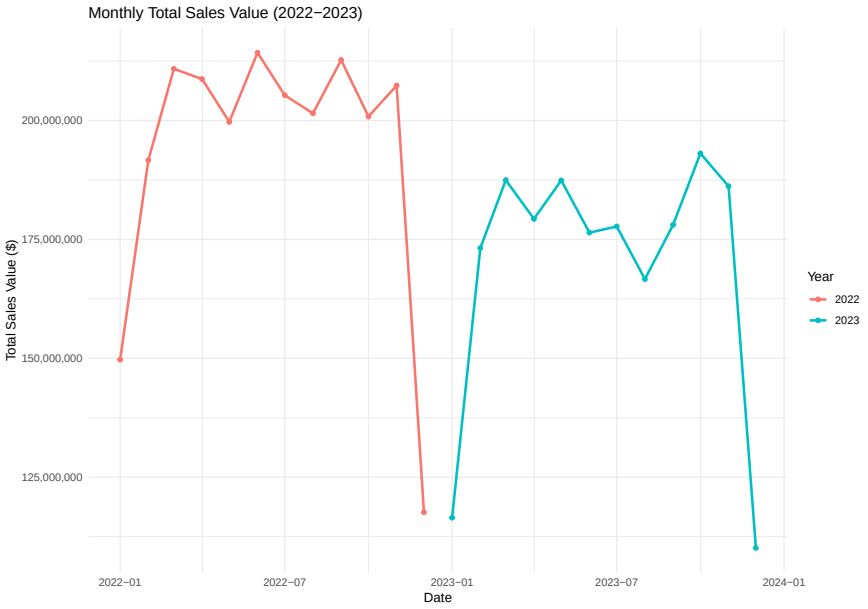
2.5.1 Summary Statistics

Table 5: Yearly Sales Summary

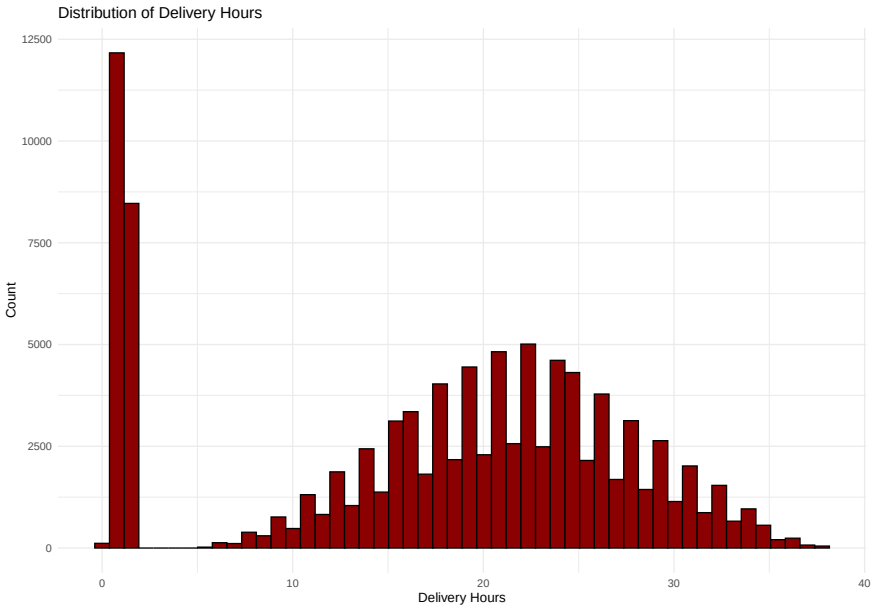
Year	Total Sales Value ($)	Total Quantity Sold	Mean Delivery (Hours)
2,022	2,320,410,018	722,141	17.51
2,023	2,032,177,660	628,206	17.44

Analysis: Yearly Sales Summary
The total sales decreased from 2022 to 2023 from $2.32 nillion to $2.03 billion, but the total sales dropped from 722141 to 628206. This suggest that prices were increased leading to a lower amount of products sold at a higher revenue due to the fact that revenue decrease is minimal compared to sales loss. delivery efficiency was maintained at relatively the same level despite the drop in sales volume.

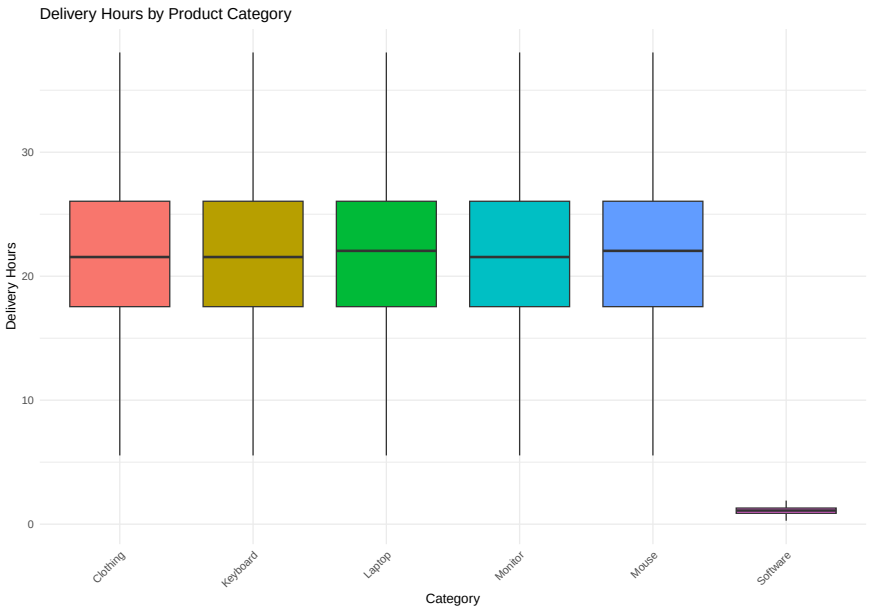
2.5.2 Visualizations



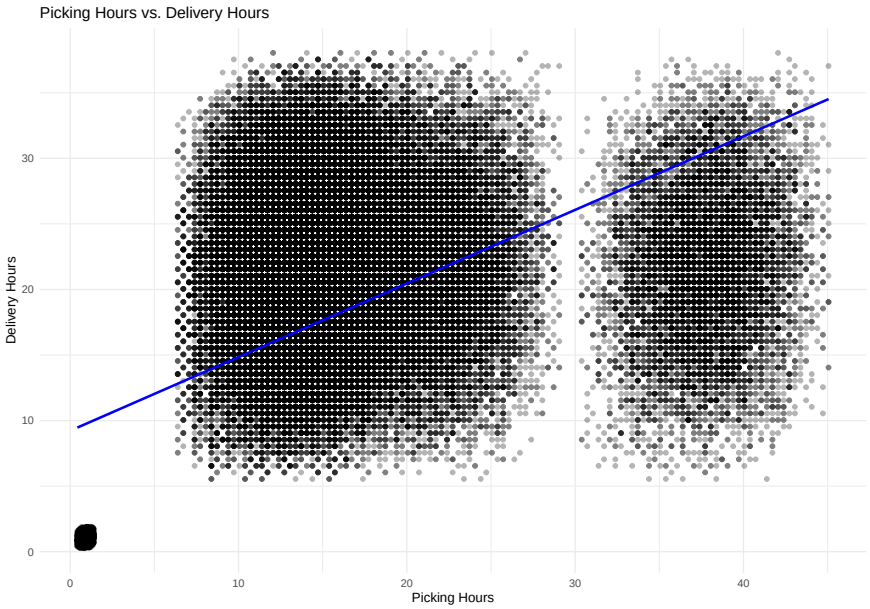
Analysis: Monthly Sales Value Trend
This graph shows that the sales have strong seasonal cycle with peaks and dips every 6 months. In 2023 the December month stood out while in 2022 there was a much more even spread in sales during the regular months. The market thus fluctuates and this could be due to promotions or holiday seasons.



Analysis: Distribution of Delivery Hours
The delivery hours shows a strong normal distribution with a lot of values which are definitely outliers. This suggest that the delivery hours are not the most trustworthy data and that by clamping data manipulation methods it can be much more reliable. The delivery hours are centered around the 17-18 hours and a minimum of around 6 and maximum of 37. There is however not a lot of delays or early deliveries with most within the mid range.

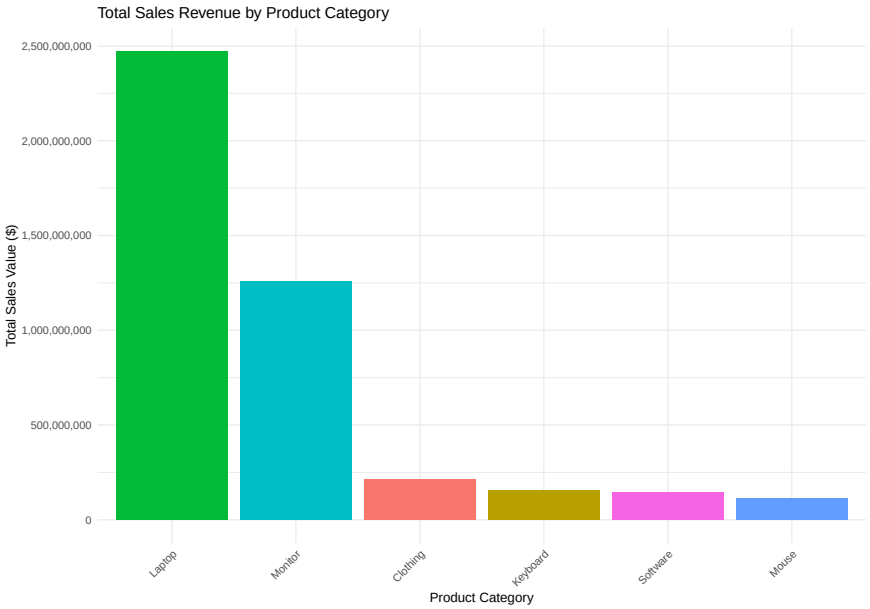


Analysis: Delivery Hours by Product Category (Boxplot)
The boxplots reveal that delivery hours and product catagories have no correlation at all and that all products are handeled exactly the same when it comes to delivery. This can be due to te company using the same delivery methods to all its products an its products being relatively teh same size. Software has a delivery hour of 0 which makes sense as it is not a physical product and is bought over the internet.



Analysis: Picking Hours vs Delivery Hours (Scatter)
Picking hours vs delivery hours does not necessarily have a high regression correlation, but the graph clearly illustrates that there are two clusters in the picking hours. The Delivery hours are spread out equally but picking hours have a cluster with average around 15h and another around 38h. The blue regression line is thus not a right conclusion and rather than a correlation between these two features the graphs found the interesting clustering.

Analysis: Picking Hours vs Selling Price (Scatter)
This scatter plot is very bound within certain areas due to the selling price differing substantially. This is mainly due to different products with different selling prices. There is however a correlation that the higher value products have higher picking hours. This can be due to workers being more cautious while handling these products vs lower value products is getting handled with less sensitivity and greater speeds. There are also a lot of low value products at the 0 picking hour zone suggesting that is the software.

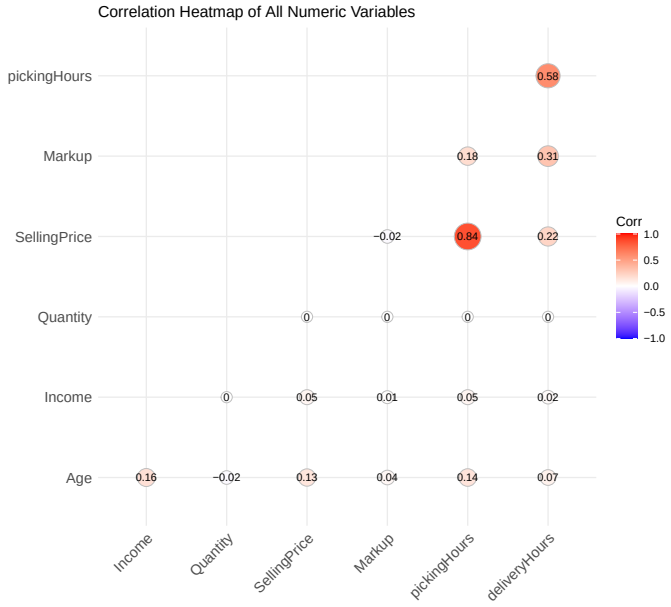


Analysis: Total Sales Revenue by Product Category (Bar Chart)
Laptops dominate the total sales with monitors far ahead in second place. This shows that al the other products are just some accessories people mainly buy with laptops. It is therefore safe to assume that the company is a laptop and monitor company.

Analysis: Sales Data Summary Conclusion:
The sales analysis therefore produced a few conclusions and insights to the seasonality in sales, reduction in sales, picking hours clusters, and type of company dealt with in the analysis.

2.6 Relationship Analysis (Heatmap)

To get a complete overview of all numeric variables, we can create a correlation heatmap. This plot shows the correlation (from -1 to 1) between every pair of variables.



Analysis: Relationship (Heatmap)
The only true correlation that can be used from this sales section is therefor the selling price vs picking hours with a correlation of 0.84 and in some extent picking hours and delivery hours.This just further enhance te argument that higher value products gets handled more carefully and also provides the insight that usually when orders are prepared it is delivered within that hour. This is however very intuitive.

2.7 Part 1 Conclusion

Part 1 provided a clear statistical and robust analysis of the customer, product and sales data of this company in 2022 and 2023. Despite some wrong interpretations like the cloud product seen as the clothing product and some errors within the data set, the analysis provided a deep understanding of the data and the causes of certain findings.

3 Part 3: Statistical Process Control (X-bar & S)

3.1 Process Capability Summary

The following table summarizes the Cp and Cpk (process capability indices) of the data for each product type of the company. It is calculated by using the first 1000 deliveries and with limits of LSL = 0h, USL = 32h. It is generally seen a capable if Cpk >= 1.33.

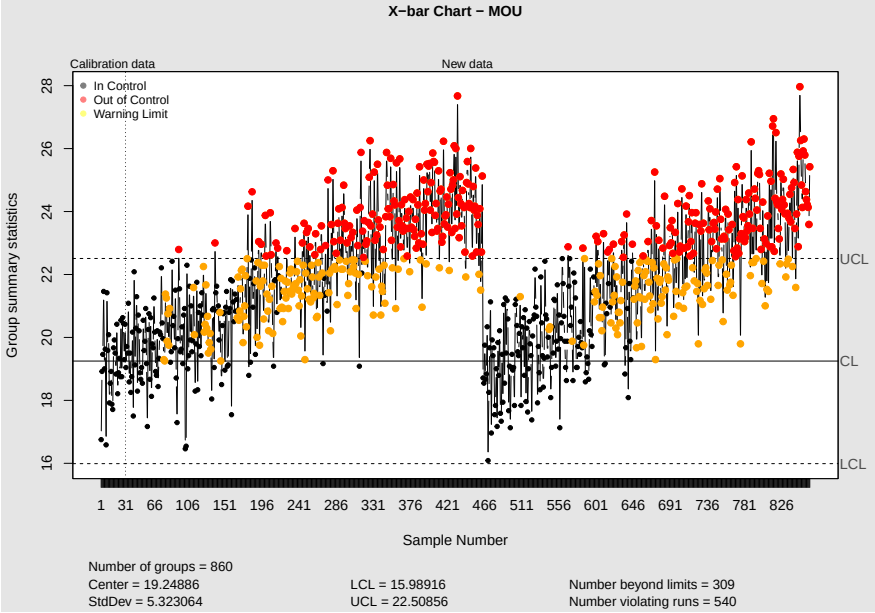
Table 6: Process Capability Summary (first 1000 deliveries, LSL=0, USL=32)

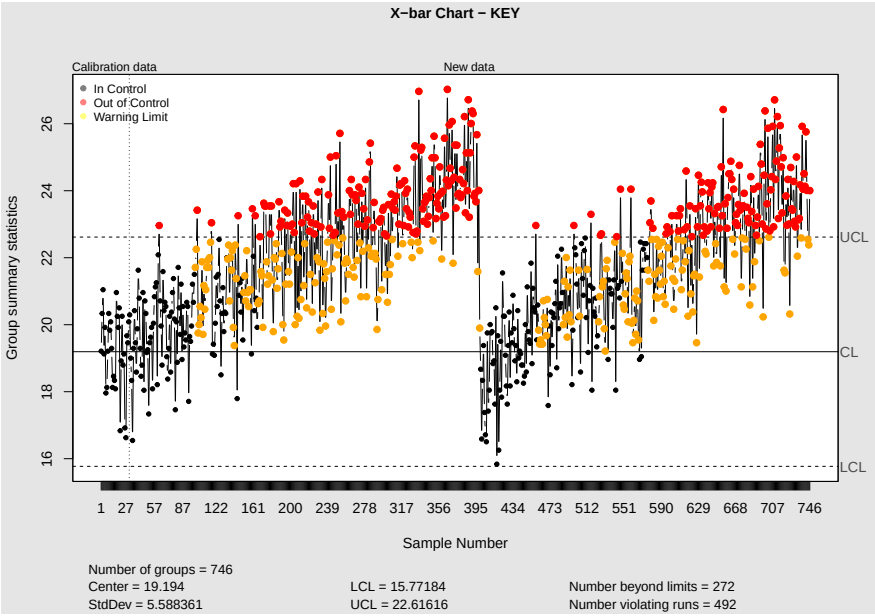
Product Type	N	Mean	Std Dev	Cp	Cpk	Capable
CLO	1000	19.226	5.941	0.898	0.717	No
KEY	1000	19.276	5.815	0.917	0.729	No
LAP	1000	19.606	5.934	0.899	0.696	No
MON	1000	19.410	5.999	0.889	0.700	No
MOU	1000	19.298	5.828	0.915	0.727	No
SOF	1000	0.955	0.294	18.135	1.083	No

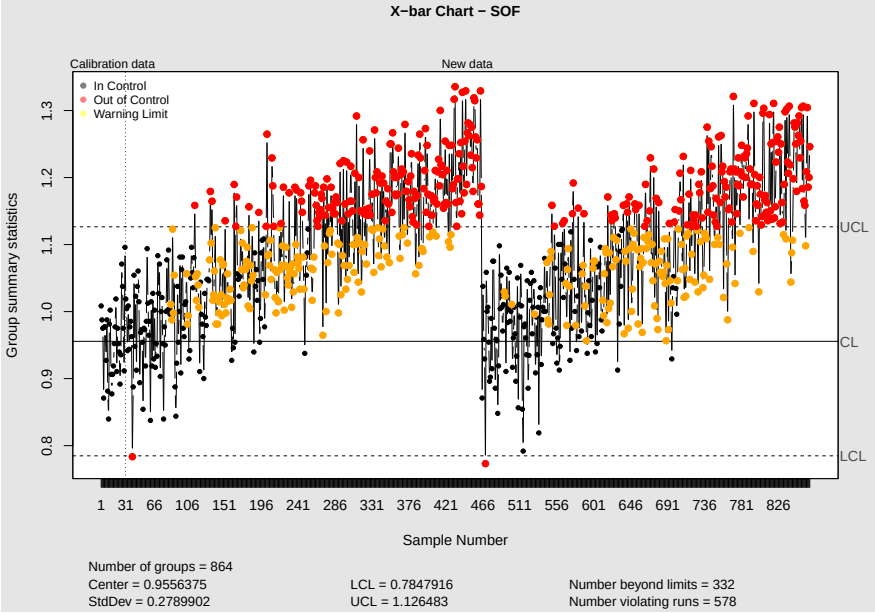
Analysis:Process Capability Summary:
This table shows that neither of the products is capable of meeting the specified voice of the customer (VOC) with the USL requirements of 32h. This is seen through all the Cpk's being lower than the 1.33 threshold. Products CLO, KEY, LAP, MON, MOU are extremely incapable with Cpk's even lower than 0.7 and a mean delivery hour in the range of 19h. Software (SOF) on the other hand have a low delivery hour due to not being a physical product, however this is also incapable with a Cpk of 1.083.

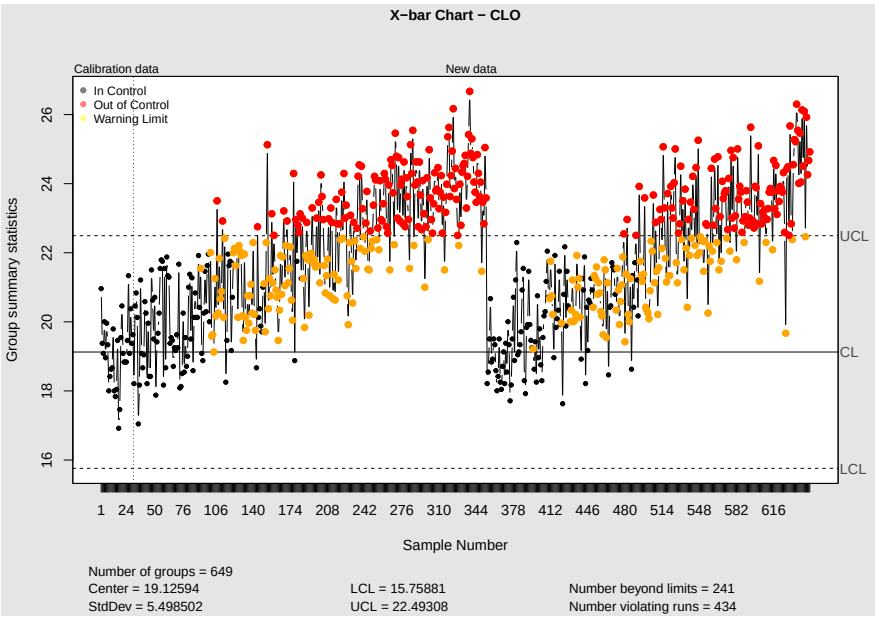
3.2 Control Charts (Phase I & II)

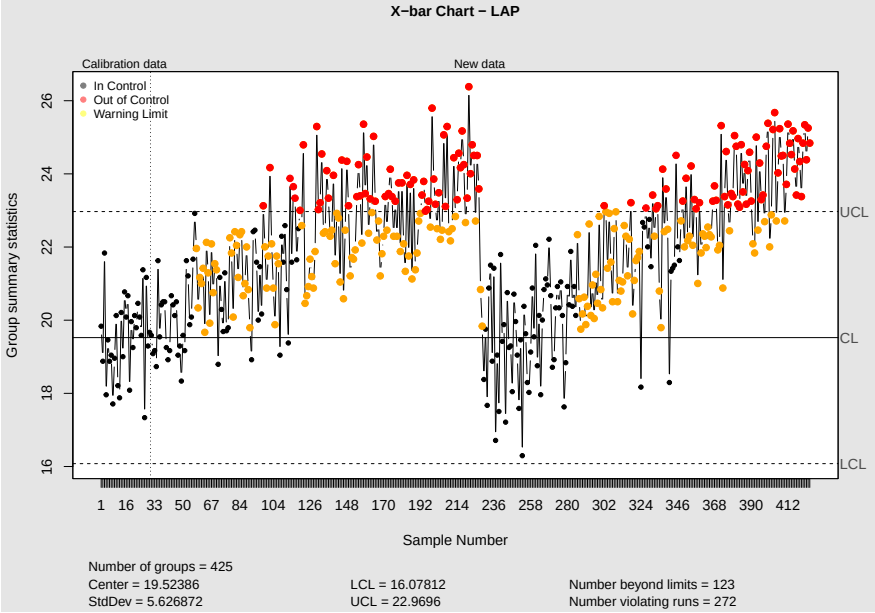
X-bar Control Charts

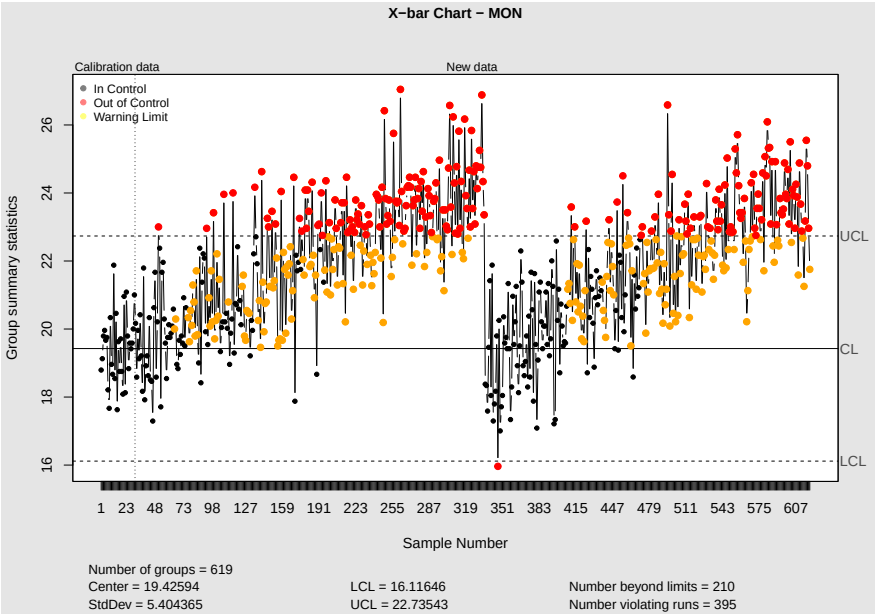

X−bar Chart − MOU
Calibration data
New data
In Control
Out of Control
Warning Limit
Group summary statistics
16
18
20
22
24
26
28
UCL
CL
LCL
1 31 66 106 151 196 241 286 331 376 421 466 511 556 601 646 691 736 781 826
Sample Number
Number of groups = 860
Center = 19.24886
StdDev = 5.323064
LCL = 15.98916
UCL = 22.50856
Number beyond limits = 309
Number violating runs = 540


X−bar Chart − KEY
Calibration data
New data
In Control
Out of Control
Warning Limit
Group summary statistics
16
18
20
22
24
26
UCL
CL
LCL
1 27 57 87 122 161 200 239 278 317 356 395 434 473 512 551 590 629 668 707 746
Sample Number
Number of groups = 746
Center = 19.194
StdDev = 5.588361
LCL = 15.77184
UCL = 22.61616
Number beyond limits = 272
Number violating runs = 492

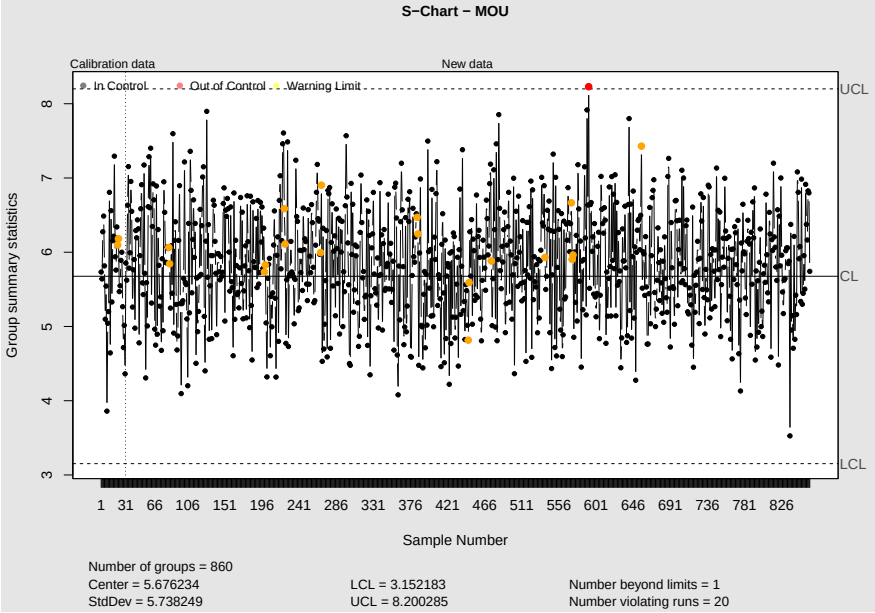

X−bar Chart − SOF
Calibration data
New data
In Control
Out of Control
Warning Limit
Group summary statistics
0.8
0.9
1.0
1.1
1.2
1.3
UCL
CL
LCL
1 31 66 106 151 196 241 286 331 376 421 466 511 556 601 646 691 736 781 826
Sample Number
Number of groups = 864
Center = 0.9556375
StdDev = 0.2789902
LCL = 0.7847916
UCL = 1.126483
Number beyond limits = 332
Number violating runs = 578

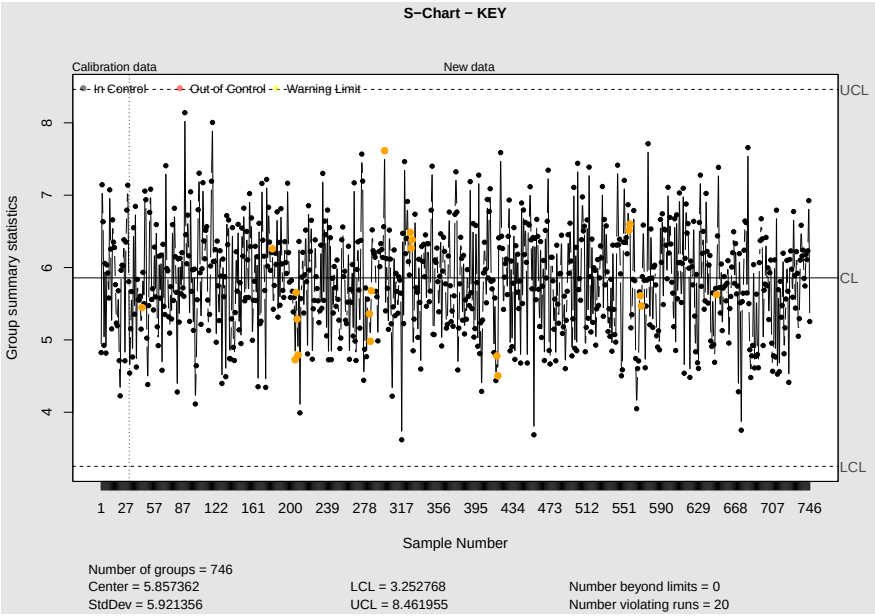

X−bar Chart − CLO
Calibration data
New data
In Control
Out of Control
Warning Limit
Group summary statistics
16
18
20
22
24
26
UCL
CL
LCL
1 24 50 76 106 140 174 208 242 276 310 344 378 412 446 480 514 548 582 616
Sample Number
Number of groups = 649
Center = 19.12594
StdDev = 5.498502
LCL = 15.75881
UCL = 22.49308
Number beyond limits = 241
Number violating runs = 434

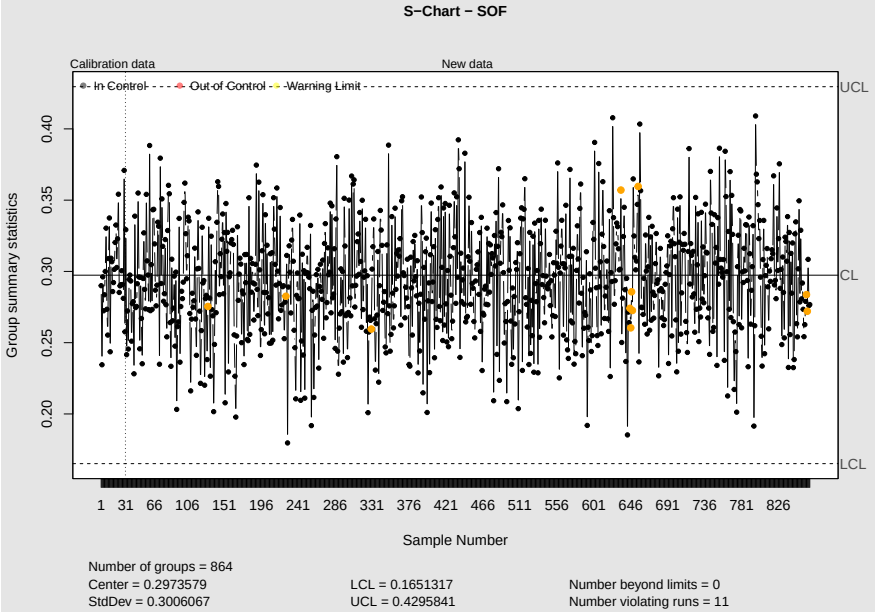


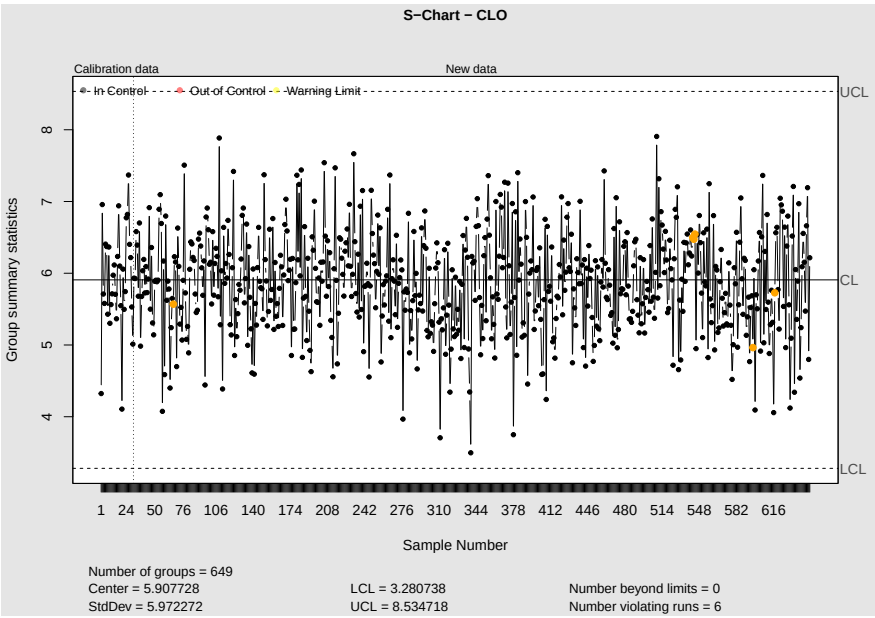


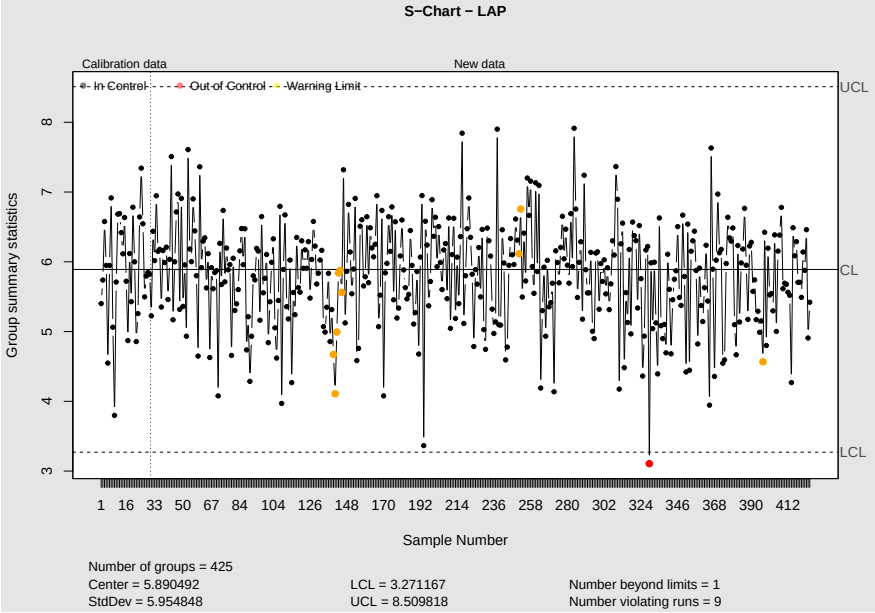
S-Control Charts

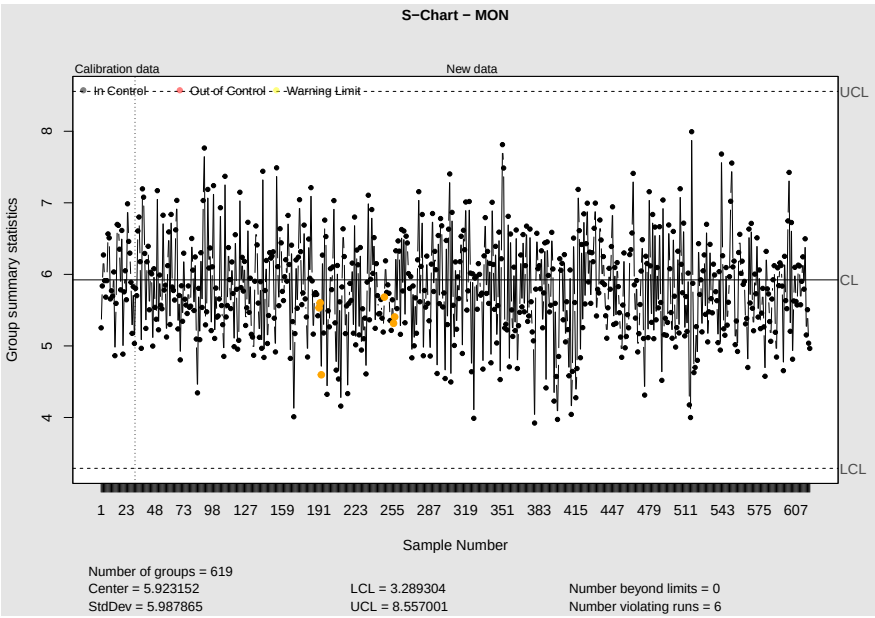

S−Chart − MOU
Calibration data
New data
In Control
Out of Control
Warning Limit
UCL
CL
LCL
Group summary statistics
3
4
5
6
7
8
1 31 66 106 151 196 241 286 331 376 421 466 511 556 601 646 691 736 781 826
Sample Number
Number of groups = 860
Center = 5.676234
StdDev = 5.738249
LCL = 3.152183
UCL = 8.200285
Number beyond limits = 1
Number violating runs = 20


S−Chart − KEY
Calibration data
New data
In Control
Out of Control
Warning Limit
UCL
CL
LCL
Group summary statistics
4
5
6
7
8
1 27 57 87 122 161 200 239 278 317 356 395 434 473 512 551 590 629 668 707 746
Sample Number
Number of groups = 746
Center = 5.857362
StdDev = 5.921356
LCL = 3.252768
UCL = 8.461955
Number beyond limits = 0
Number violating runs = 20


S−Chart − SOF
Calibration data
New data
In Control
Out of Control
Warning Limit
UCL
CL
LCL
Group summary statistics
0.20
0.25
0.30
0.35
0.40
1 31 66 106 151 196 241 286 331 376 421 466 511 556 601 646 691 736 781 826
Sample Number
Number of groups = 864
Center = 0.2973579
StdDev = 0.3006067
LCL = 0.1651317
UCL = 0.4295841
Number beyond limits = 0
Number violating runs = 11


S−Chart − CLO
Calibration data
New data
In Control
Out of Control
Warning Limit
UCL
CL
LCL
Group summary statistics
4
5
6
7
8
1 24 50 76 106 140 174 208 242 276 310 344 378 412 446 480 514 548 582 616
Sample Number
Number of groups = 649
Center = 5.907728
StdDev = 5.972272
LCL = 3.280738
UCL = 8.534718
Number beyond limits = 0
Number violating runs = 6





Analysis:

X-Bars:

X-bar Charts monitors the process mean / average. This was totally out of control with a lot of values out of control and a large amount on the warning limit. These charts used the first 30 calibration data samples to set the limits and then added the rest of the samples. The newer data suggests an immediate very strong upward trend. These upward red and yellow dots are violating runs and in total there are 405 for SOF. One can derive that all the products average delivery time has increased systematically and is no longer in the stable center.

S-Charts:

S-Charts is relatively similar, however it measures the standard deviations. All the products showed a strong tendency to stay within the bounds except for some individual outliers. This finding suggest that the standard deviation stayed consistent and predictable for all the products across all the samples.

3.3 Process Control Rule Violations Summary

The following tables summarizes out of control sample counts for 3 different rules. during phase 2 of control monitoring for each of the product types.

3.3.1 Rule A: S-Chart Points Above +3 Sigma

Rule A checks for points in the S-chart that are exceeding the UCL with a sigma +3.

Table 7: Rule A: S samples outside +3 (UCL)

Product Type	Samples Monitored (Phase II)	Violating s-samples (> +3)
CLO	619	None (0)
KEY	716	None (0)
LAP	395	None (0)
MON	589	None (0)
MOU	830	1 violations
SOF	834	None (0)

Analysis: (Rule A):
There is only one point which is marginal and confirms the S-Charts consistency.

3.3.2 Rule B: Longest Run of S-Chart Points Between +/- 1 Sigma

Rule B looks for instances where points are within 1 sigma above or bellow the middle. This also confirms the S-charts good control as once again its just one value outside of these bounds.

Table 8: Rule B: Longest consecutive S samples inside ±1

Product Type	Samples Monitored (Phase II)	Longest Run inside ±1 (s)
CLO	619	35
KEY	716	15
LAP	395	19
MON	589	34
MOU	830	16
SOF	834	21

Analysis: (Rule B):
This also confirms the S-charts good control as once again as the highest sample count is 35 which is still marginal in comparison to the total samples of 619 for CLO.

3.3.3 Rule C: 4 Consecutive X-Bar Points Above +2 Sigma

Rule C looks for values +2 sigma UCL on the X-Bar.

Table 9: Rule C: Runs of 4 consecutive X samples above +2

Product Type	Samples Monitored (Phase II)	Violating Xbar runs (4 above +2)

CLO	619	280 violations
KEY	716	317 violations
LAP	395	171 violations
MON	589	259 violations
MOU	830	360 violations
SOF	834	359 violations

Analysis: (Rule C):
Here is a dramatic amount of the values with the highest being 359 violations from 834 samples. This confirms that there was an upward shift in the mean delivery time as seen on the X-Bar charts.

3.4 Part 3 Conclusion

The SPC (Statistical Process Control) analysis provided us with the insights that the mean time of delivery has dramatically increased over time where the standard deviation of delivery time has stayed very consistent. It was also found that none of the products were able to fulfill the customer demands and should thus be looked into.

Part 6 will look deeper into the situation to find the root cause of the non-random upward trend in delivery times.

4 Part 4: Risk, Data Correction and Optimising

Thus section estimates the statistical risks with the SPC rules that was used at part 3 and also addresses the data quality issues of the product data sets.

4.1 Type I Error Estimation ( )

A type I error occurs when a process is when we concluded that a process is out of control while it is actually is within control. The theoretical probability of this will be calculated here within this section.

Assumptions: We assume the process statistic (sample mean $\bar{x}$ or sample standard deviation s) follows the expected distribution (approximately normal for $\bar{x}$ due to CLT, related to chi-squared for s) when the process is truly in control and centered on the established CL.

- **Rule A (1 S-sample > +3 sigma):** For standard control charts, the 3-sigma limits are designed such that the probability of a single point falling *outside* these limits purely by chance (when the process is in control) is very small. Assuming approximate normality for the sample statistic:
  - $P(\text{point} > UCL \text{ or point} < LCL) \approx P(Z > 3 \text{ or } Z < -3)$
  - This equals $2 \times P(Z < -3)$.
  - The probability for *Rule A specifically* (only > +3 sigma) is $P(\text{point} > UCL) \approx P(Z > 3)$.
- **Rule B (Longest run of S between +/- 1 sigma):** This rule measures stability *within* control limits and doesn't directly signal an out-of-control state. Therefore, the concept of a Type I error (false *alarm*) isn't directly applicable to this specific rule as defined. We report the probability of a point falling *within* the +/- 1 sigma zone by chance, assuming normality:
  - $P(-1 < Z < 1) \approx 0.6827$
- **Rule C (>= 4 consecutive X-bar > +2 sigma):** We need the probability (p) of a single $\bar{x}$ point falling above the +2 sigma limit by chance, assuming normality and the process being in control.

- $p = P(\bar{x} > CL + 2\sigma_{\bar{x}}) = P(Z > 2)$ The probability of getting *exactly* 4 consecutive points above +2 sigma by chance is p^4. The probability of *at least* 4 is more complex to calculate precisely for a continuous monitoring scenario but is dominated by the p^4 term. We calculate p^4 as an estimate for the likelihood of this specific pattern occurring purely by chance in any given sequence of 4 points.

Table 10: Theoretical Type I Error ( ) / Event Probability Estimates for SPC Rules

Rule	Relevant Probability (p)	Approx. Type I Error ( ) / Event Likelihood	Interpretation
Rule A (S > +3)	0.001350	0.00135	Prob. of one poin
Rule B Zone (S within ±1)	0.682689	NA	Prob. of one poin
Rule C (4 X > +2)	| 0.02275	| 0.0000	|Approx. prob. o

Analysis:
The type I error for rule A is relatively small with a value of 0.00135. This suggest that there is only a 0.135% chance for a false alarm with rule A. Rule B does not have a Type I error however we can see that we should expect around 68% of all the samples to be within 1 sigma of the middle. Rule C has basically 0% probability of a false alarm and is therefore the most reliable.

Also to clarify any unit has a 50% chance to be above the centerline.

4.2 Type II Error Estimation ( )

A type II error occurs when a process is deemed to not be out of control while it actually is. This is a much worse case. A scenario was given to work with. Scenario:

Original Process: 0=25.05 L (CL)

Original Control Limits: LCL = 25.011 L, UCL = 25.089 L

Shifted Process: 1=25.028 L, x =0.017 L (Std. dev. of sample means after shift)

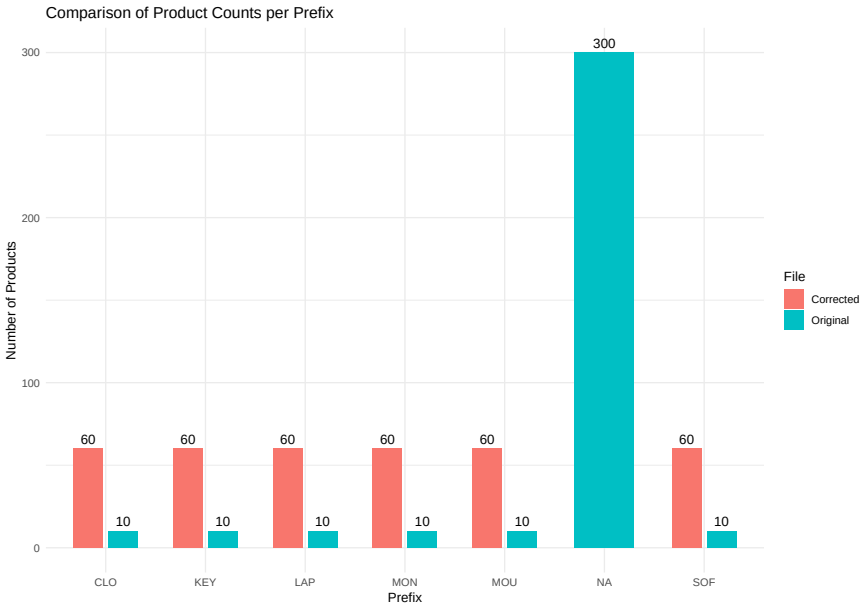
We need to calculate the probability that a sample mean (x) from the shifted process falls within the original control limits. This probability is .

Table 11: Type II Error ( ) Calculation for Shifted Bottle Filling Process

Parameter	Value
Original CL (0)	25.0500
Original LCL	25.0110
Original UCL	25.0890
Shifted Mean (1)	25.0280
Shifted Std.Dev ('x̄)	| 0.017
Z-score (LCL)	-1.0000
Z-score (UCL)	3.5882
Type II Error Prob ( )	0.8412
Power (1-)	0.1588

Analysis:
The calculated is 0.8412. Thus there is an 84.1% chance that this sample falls within the limits while we classified that it doesn't.

4.3 Data Correction and Re-Analysis

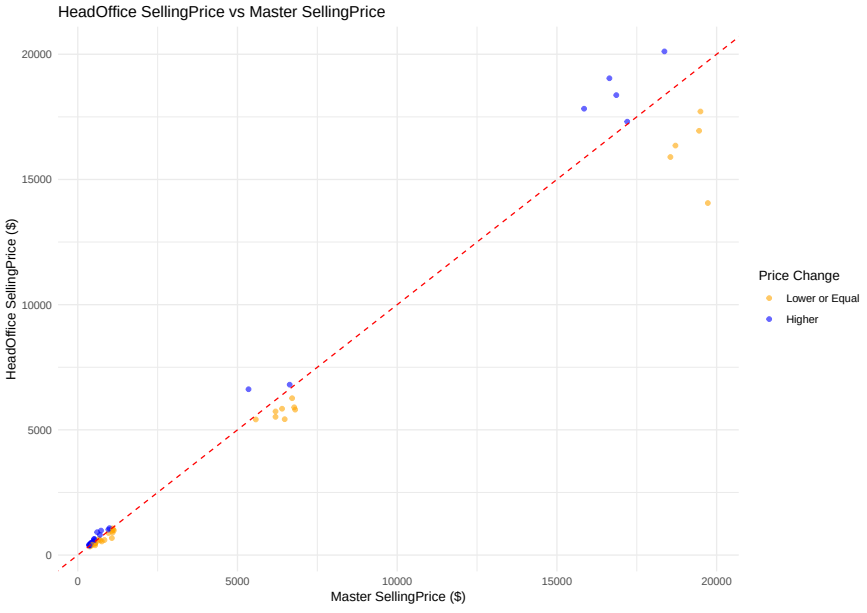


Analysis: Comparison of Product Counts per Prefix

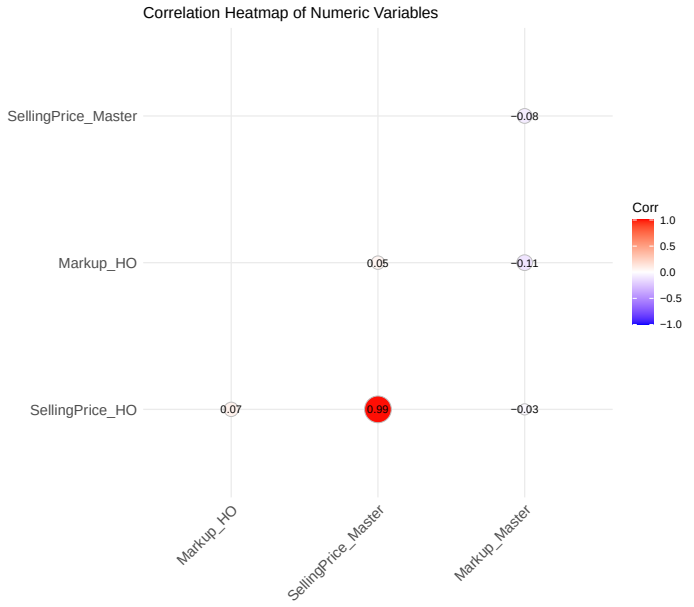
This graph shows that the data set was indeed wrong. The original set classified most of the data as NA while it was all spread across all the products. Thus there were not just 109 products of each but 60.

Table 12: Summary Comparison of HeadOffice vs. Master Product Prices

Mean HeadOffice Price ($)	Mean Master Price ($)	Mean Difference ($)	Number of Products
4,298.06	4,493.59	-195.54	60

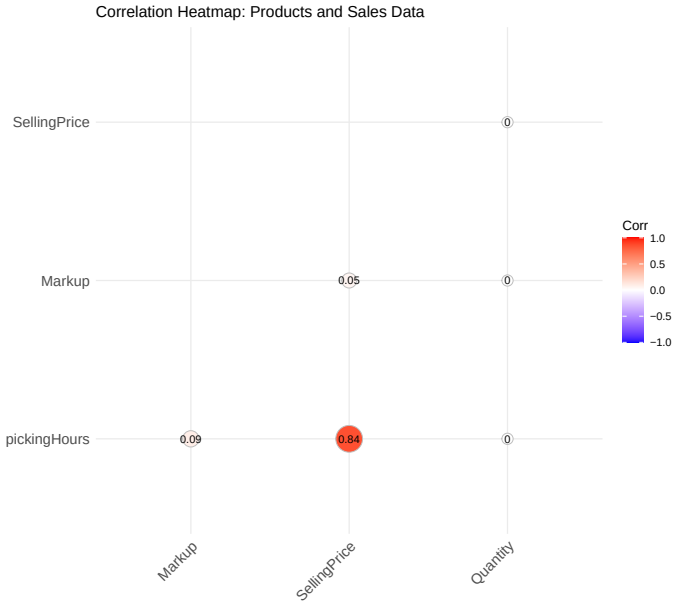


Analysis: HeadOffice Selling Price vs Master Selling Price
This graph compares the old sales prices to the new ones. The red line would be if both sets was the same. However one can see that the higher value products differs much more from each other that the lower value process. All in all the selling prices are relatively similar after the correction of the data.



Analysis: Correlation Heatmap (HeadOffice/Master)

This heatmap suggests that the selling price of the old and corrected data is basically similar enhancing the point that it was not the sales prices that was wrong in the data sets.

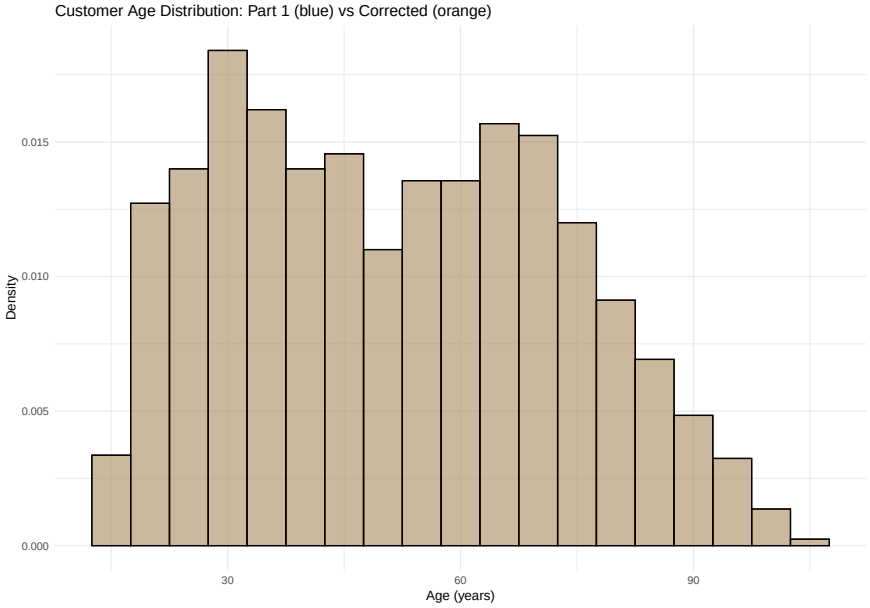


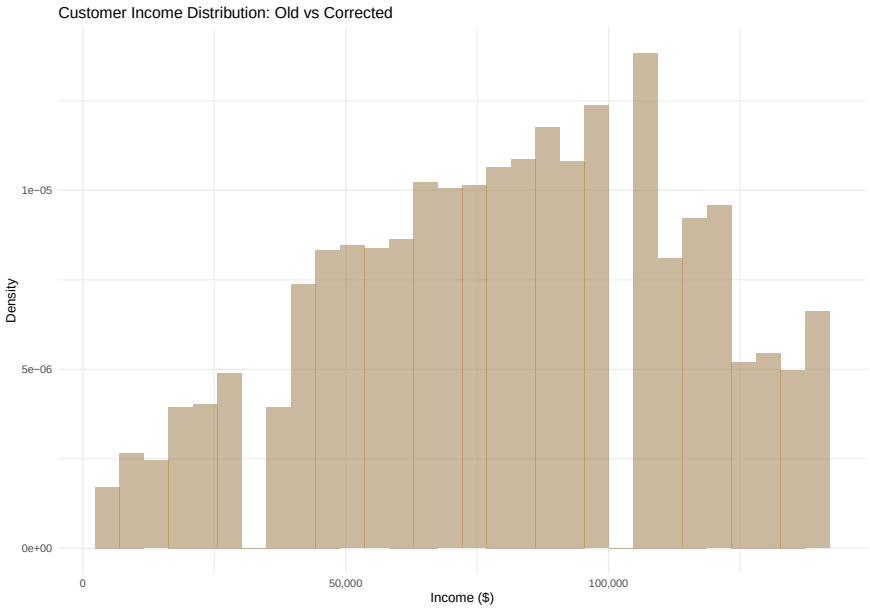
Analysis: Correlation Heatmap (Products and Sales Data)

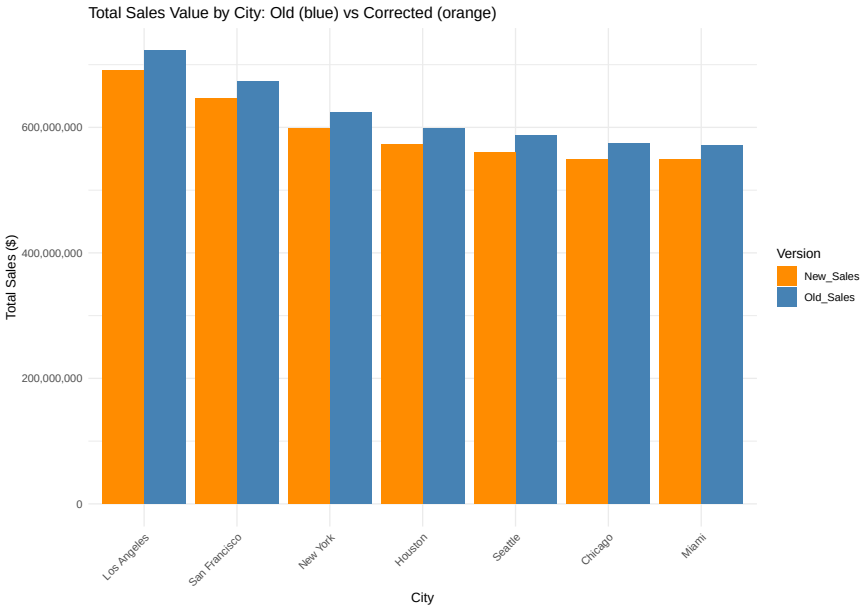
Selling price and picking hours still have a high corelation. No change here.

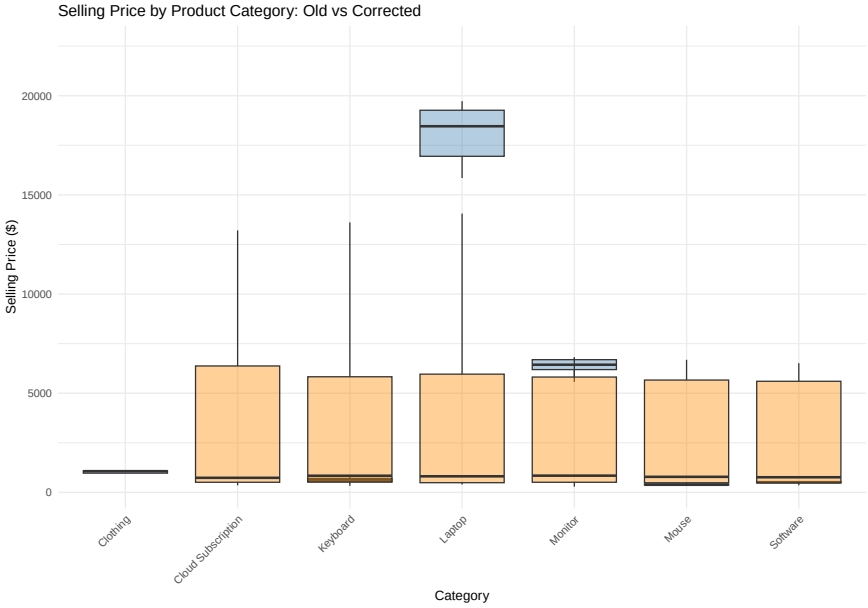
"Old vs. Corrected" Comparison Plots

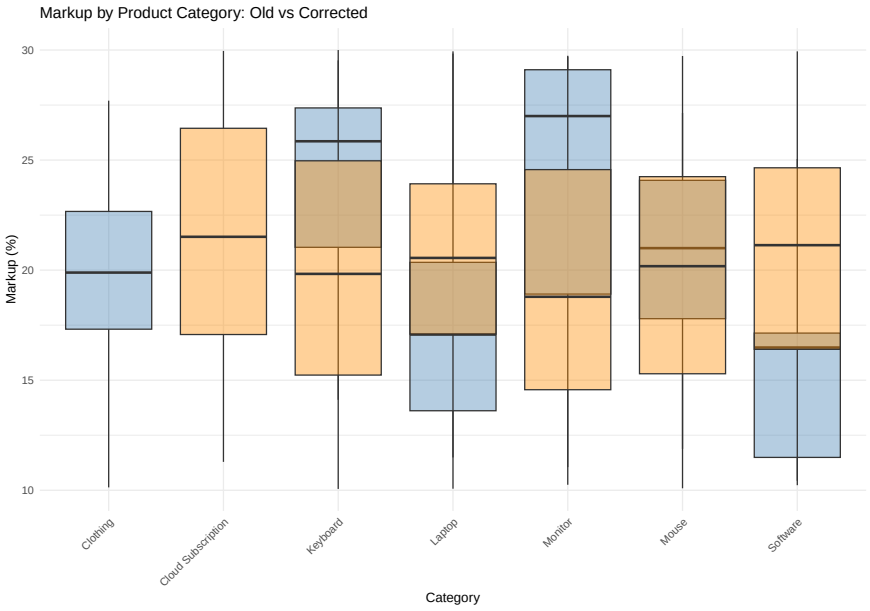
This series of graphs (density plots, bar charts, boxplots) overlays the original data (labeled "Old_Sales" or blue) with the newly corrected data (labeled "New_Sales" or orange) to visualize the impact of the correction.

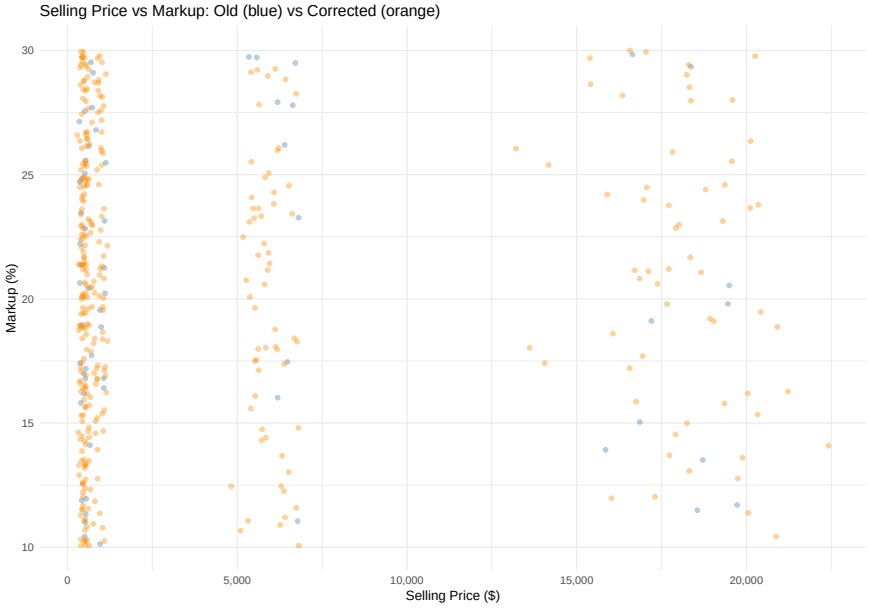





Total Sales Value by City: Old (blue) vs Corrected (orange)
Total Sales ($)
600,000,000
400,000,000
200,000,000
0
Los Angeles
San Francisco
New York
Houston
Seattle
Chicago
Miami
City
Version
New_Sales
Old_Sales


Selling Price by Product Category: Old vs Corrected
Selling Price ($)
20000
15000
10000
5000
0
Clothing
Cloud Subscripton
Keyboard
Laptop
Monitor
Mouse
Software
Category


Markup by Product Category: Old vs Corrected
Markup (%)
30
25
20
15
10
Clothing
Cloud Subscription
Keyboard
Laptop
Monitor
Mouse
Software
Category


Selling Price vs Markup: Old (blue) vs Corrected (orange)
Markup (%)
30
25
20
15
10
0
5,000
10,000
15,000
20,000
Selling Price ($)

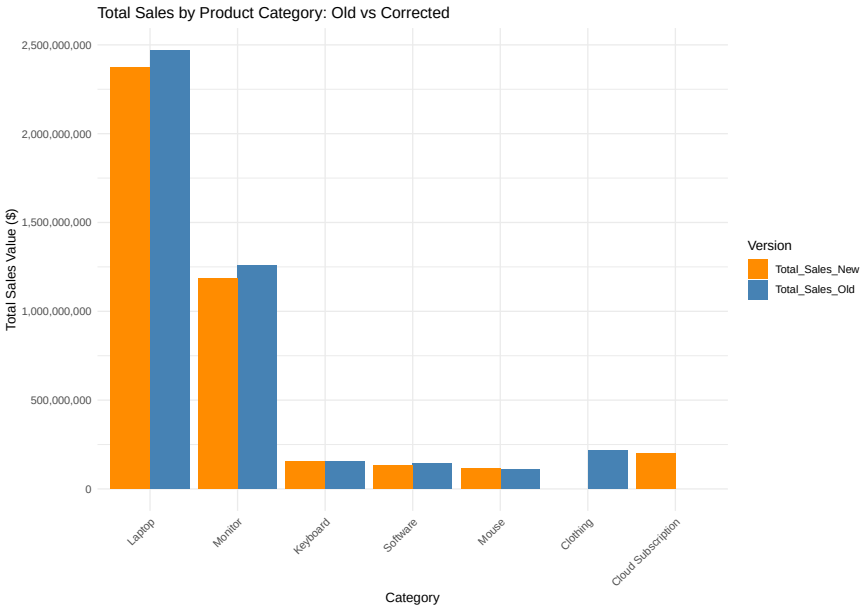
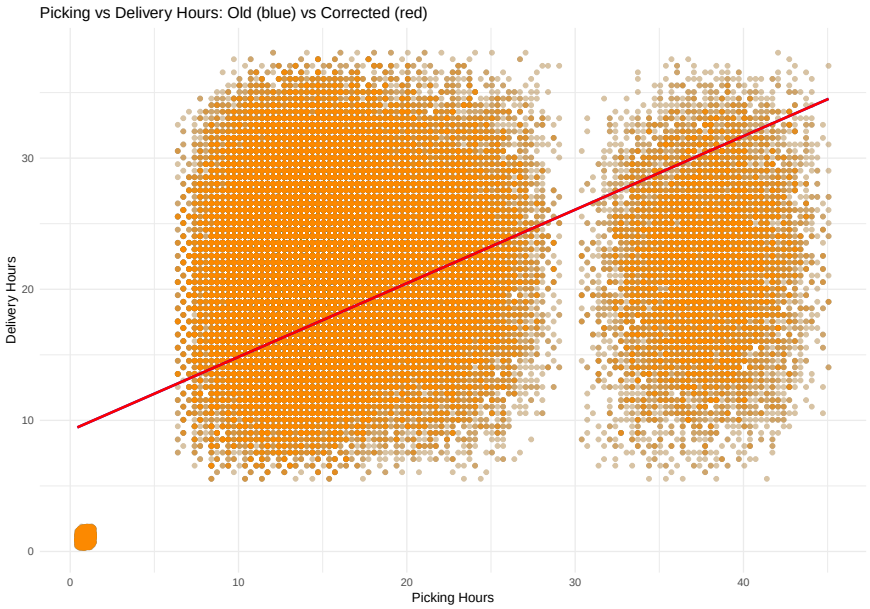
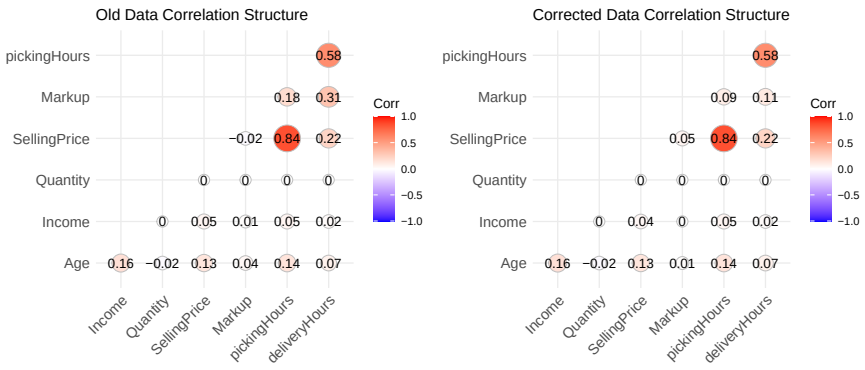


Table 13: Change in Total Sales Value by Product Category (Old vs Corrected)

Category	Total_Sales_Old	Total_Sales_New	Diff	Pct_Change
Clothing	214110418	NA	NA	NA
Keyboard	155002210	157138505	2136295	1.38
Laptop	2470814376	2374726735	-96087641	-3.89
Monitor	1258942847	1184192159	-74750688	-5.94
Mouse	111190471	116556049	5365578	4.83
Software	142527355	132497284	-10030072	-7.04
Cloud Subscription	NA	201266954	NA	NA





Analysis: Comments on Differences
The "Old vs. Corrected" Comparison Plots are used top analyze the impact of the improved data set. The density plots, bar charts and box plots are overlays of the old vs new data sets. The plots for Customer Age , Customer Income , and Total Sales Value by City are nearly identical showing that none of this data changed with the new set.

The boxplots for Selling Price and Markup has a clear difference. It can also be seen that the previous analysis mistaken Cloud subscriptions for Clothing which is now eliminated. The markup buy product changed notably for most of the products.The total sales by product category also mostly decreased for all products and some even with up to 7%.

The heatmaps provided us with the information that no relationships changed except slightly for the selling price vs markup, picking hours vs markup, delivery hours vs markup, income vs selling price, age vs markup, income vs markup. This suggest that only the markup and selling price actually adapted due to more data available.

4.4 Part 4 Conclusion

This section quantified the risk of the SPC A,B,C rules, Type I and Type II risks and finally the data correction analysis. All in all the data that was corrected did not have a major impact due to the through part 1 and assumptions guessed relatively right.

5 Part 5: Optimise Profit

This part focusses on the number on baristas vs the service time and ultimately amount of baristas needed to maximize profit of the coffee shop.

5.1 Data Loading and Reliability Analysis

Tghe fisrt step is to load the data sets for shop 1 and shop 2 and calculating the reliability of the past year for both shops.

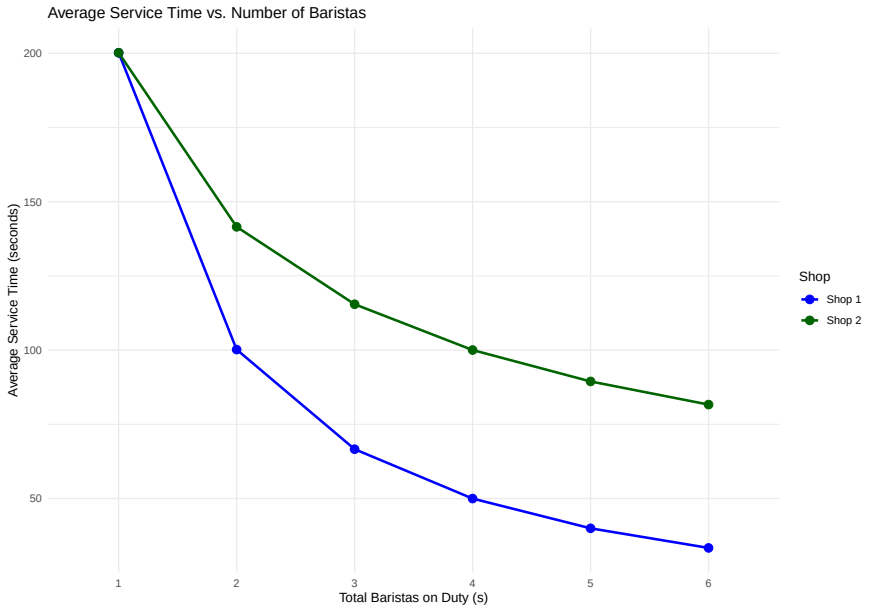
Table 14: Observed Customer Service Reliability (1 Year)

Shop	Total Customers (1 Year)	Reliable Customers (>=2 Baristas)	Reliability %
Shop 1	2e+05	199583	99.8%
Shop 2	2e+05	197804	98.9%

If less than 2 baristas are present and service is not reliable. The data shows that that rarely happened.

5.2 Service Time vs. Number of Baristas

Here we analyze the service time in seconds vs number of baristas in.



Analysis: Average Service Time vs. Number of Baristas (Line):
This graph illustrate that for both shops the average service time decreases per customer if the number of baristas increase. Unfortunately 6 baristas is the maximum amount according to the brief. The rate flattens between 4,5 and six which suggests a deeper analysis is still needed to understand how many is needed for a maximized profit. Shop 1 has an overall lower service time. This can suggest that the equipment or space is more efficient.

5.3 Profit Optimisation Model

Here we will build a model that estimates the daily profit for the different number of baristas present.

Model:

Average Service Time, T(s): We calculate the average service time (in seconds) from the data for each number of baristas (s).

Total Customers Served per Day, C(s): We estimate the total number of customers a shop can serve in a day. We assume an. 8-hour (28,800 seconds) workday where all s baristas work in parallel.

- Total Daily Capacity = (Total Seconds per Day / T(s)) × s.

Total Daily Profit, P(s): We use the costs and profits from the brief:

- Revenue = C(s)×R30 (profit/customer)
- Cost = s × R1,000 (cost/barista/day)
- Profit = Revenue - Cost

5.4 Optimisation Results

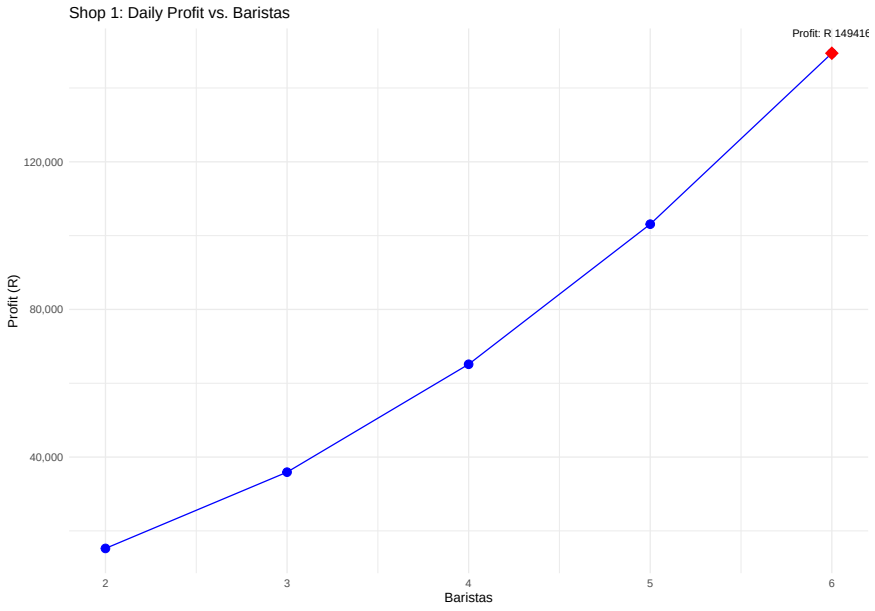
Table 15: Profit Optimization - Shop 1 (8-hour Day)

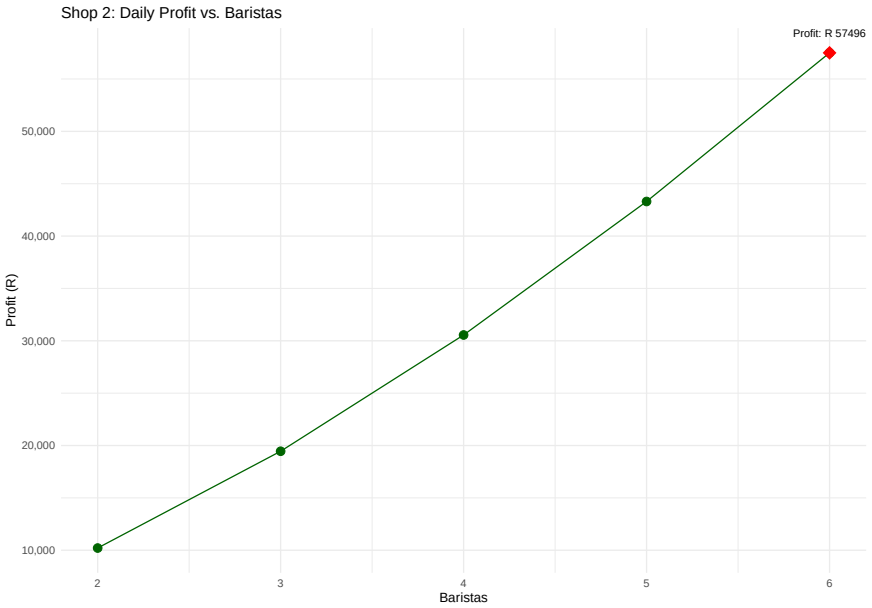
Shop	Baristas (s)	Avg Service (s)	Customers/Day	Revenue/Day	Cost/Day	Profit/Day
Shop 1	2	100.17	575.02	17250.51	2000	15250.51
Shop 1	3	66.61	1297.07	38912.06	3000	35912.06
Shop 1	4	49.98	2304.90	69147.14	4000	65147.14
Shop 1	5	39.96	3603.44	108103.14	5000	103103.14
Shop 1	**6**	**33.36**	**5180.53**	**155415.97**	**6000**	**149415.97**

Table 16: Profit Optimization - Shop 2 (8-hour Day)

Shop	Baristas (s)	Avg Service (s)	Customers/Day	Revenue/Day	Cost/Day	Profit/Day
Shop 2	2	141.51	407.03	12210.75	2000	10210.75
Shop 2	3	115.44	748.43	22453.04	3000	19453.04
Shop 2	4	100.02	1151.82	34554.72	4000	30554.72
Shop 2	5	89.44	1610.09	48302.71	5000	43302.71
Shop 2	**6**	**81.64**	**2116.54**	**63496.17**	**6000**	**57496.17**

5.5 Profit vs. Baristas Plots





Analysis:
Shop 1: It is thus seen that the profit is dramatically optimized for 6 baristas. for both shop 1 and 2. This is due to the fact that additional baristas increase the total efficiency and thus the service time each barista takes and additional baristas also means the total available time is more due to an entire extra shift. It thus has a double effect decreasing the costs per customer dramatically.

5.6 Part 5 Conclusion

Thus the profit optimization for the coffee shop showed 6 baristas is needed for both shops. The model assumed a 8h work day, R30 profit a customer and an additional cost of R1000 for each barista.

The profit optimization analysis for the coffee shops was conclusive. Based on the provided data, the relationship between staffing and service time follows a classic diminishing returns curve (Shop 1 is inherently faster than Shop 2).

6 Part 6: DOE and ANOVA

This section uses data analysis of variance (ANOVA) to to analyze the difference in the delivery times. This is based on different factors namely monthly and yearly for each product. This analysis builds on part 3 that focused on the SOF (software) product. The goal is to determine whether the delivery performance over 2026/2027 varies a lot.

6.1 Data Preparation for ANOVA

The `ordData` object created in Part 3 are used. It is filtered for SOF products and the time variables, year and month are treated as categorical factors.

6.2 ANOVA: Comparing Delivery Hours by Year

Hypothesis:

H0: The mean delivery time for “SOF” products is the same across the years present in the data (e.g., 2026= 2027).

Ha: The mean delivery time for “SOF” products is different for at least one year.

We perform a one-way ANOVA comparing deliveryHours across the Year factor.

— Debug: Trying to print Year ANOVA Table —

Table 17: ANOVA Results: Year - SOF (Basic)

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Signif
Year	1	0.017	0.017	0.179	0.672	
Residuals	20747	1965.723	0.095	0.179	0.672	

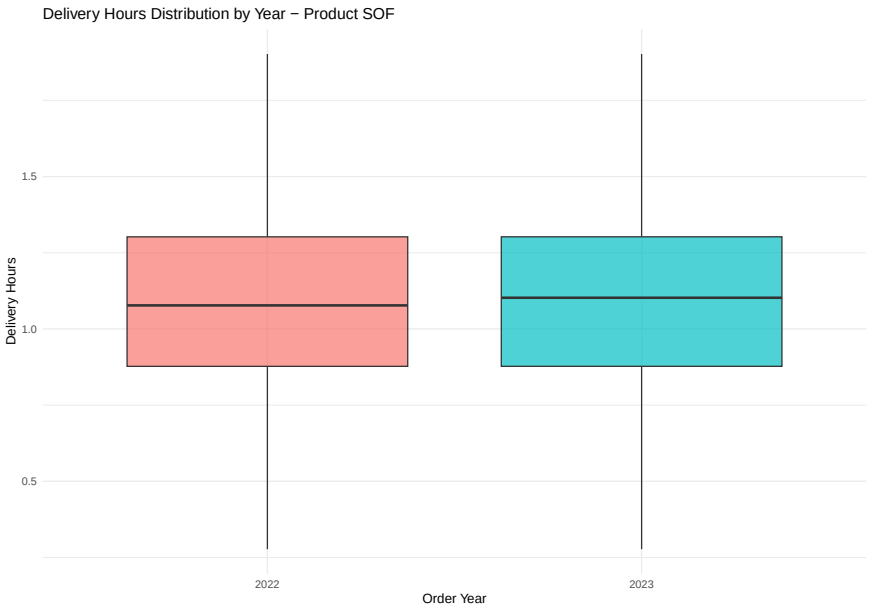


Figure 1: Boxplot of Delivery Hours by Year

Analysis: Delivery Hours Distribution by Year (Boxplot):
This box plot compares the 2022,2023 delivery hours which is nearly identical. the mean, 1st and 3rd quantile and whiskers are exactly the same in the image which suggest no change in performance.

6.3 ANOVA: Comparing Delivery Hours by Month

Hypothesis:

H0: The mean delivery time for “SOF” products is the same across all months (e.g., Jan= Feb=...= Dec).

Ha: The mean delivery time for “SOF” products is different for at least one month.

We perform a one-way ANOVA comparing deliveryHours across the Month factor.

— Debug: Trying to print Month ANOVA Table —

Table 18: ANOVA Results: Month - SOF (Basic)

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Signif
Month	11	138.191	12.563	142.549	0	***
Residuals	20737	1827.549	0.088	142.549	0	***

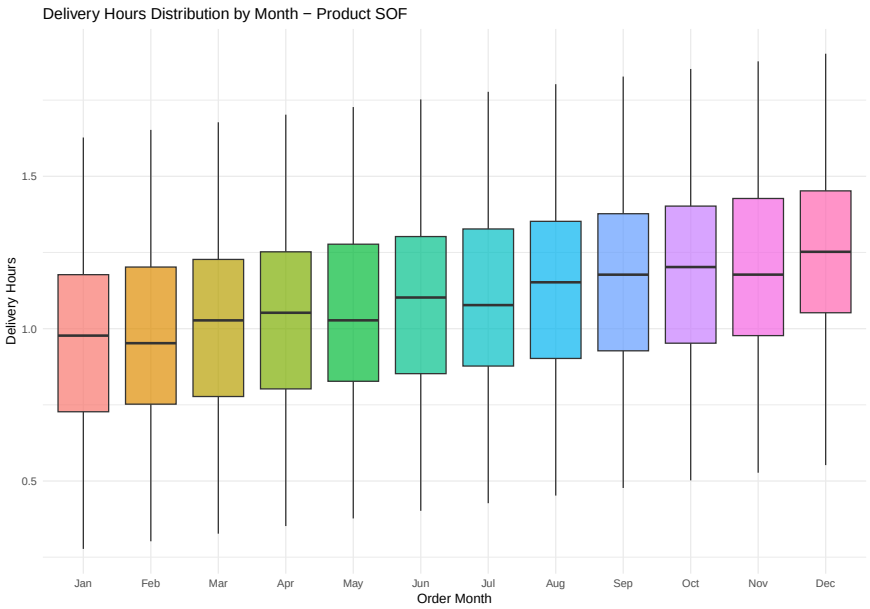


Figure 2: Boxplot of Delivery Hours by Month

Analysis: Delivery Hours Distribution by Month (Boxplot):
This plot shows the difference in monthly delivery times. This shows a clear upward trend in time from January to December. This is a strong visual pattern showing the effect of seasonality.

— Debug: Trying to print Tukey Table —

Tukey HSD Results ($p < 0.05$):

Table 19: Significant Month Differences - SOF (Basic)

Comparison	diff	lwr	upr	p adj
Apr-Jan	0.0815	0.0461	0.1170	0.0000
May-Jan	0.0976	0.0620	0.1331	0.0000
Jun-Jan	0.1288	0.0929	0.1648	0.0000
Jul-Jan	0.1392	0.1037	0.1747	0.0000
Aug-Jan	0.1667	0.1312	0.2023	0.0000
Sep-Jan	0.2058	0.1703	0.2414	0.0000
Oct-Jan	0.2228	0.1872	0.2584	0.0000
Nov-Jan	0.2375	0.2018	0.2732	0.0000
Dec-Jan	0.2840	0.2435	0.3245	0.0000
Apr-Feb	0.0617	0.0301	0.0934	0.0000
May-Feb	0.0778	0.0460	0.1095	0.0000
Jun-Feb	0.1090	0.0768	0.1413	0.0000
Jul-Feb	0.1194	0.0877	0.1511	0.0000

Comparison	diff	lwr	upr	p adj
Aug-Feb	0.1469	0.1152	0.1787	0.0000
Sep-Feb	0.1860	0.1542	0.2178	0.0000
Oct-Feb	0.2030	0.1712	0.2348	0.0000
Nov-Feb	0.2177	0.1858	0.2496	0.0000
Dec-Feb	0.2642	0.2270	0.3014	0.0000
Jun-Mar	0.0818	0.0495	0.1141	0.0000
Jul-Mar	0.0922	0.0604	0.1240	0.0000
Aug-Mar	0.1197	0.0879	0.1515	0.0000
Sep-Mar	0.1588	0.1270	0.1907	0.0000
Oct-Mar	0.1758	0.1439	0.2077	0.0000
Nov-Mar	0.1905	0.1585	0.2225	0.0000
Dec-Mar	0.2370	0.1997	0.2743	0.0000
Aug-Apr	0.0852	0.0535	0.1170	0.0000
Sep-Apr	0.1243	0.0925	0.1561	0.0000
Oct-Apr	0.1413	0.1094	0.1731	0.0000
Nov-Apr	0.1560	0.1241	0.1879	0.0000
Dec-Apr	0.2025	0.1652	0.2397	0.0000
Aug-May	0.0692	0.0373	0.1010	0.0000
Sep-May	0.1083	0.0763	0.1402	0.0000
Oct-May	0.1252	0.0933	0.1572	0.0000
Nov-May	0.1399	0.1079	0.1720	0.0000
Dec-May	0.1864	0.1491	0.2238	0.0000
Sep-Jun	0.0770	0.0446	0.1094	0.0000
Oct-Jun	0.0939	0.0615	0.1264	0.0000
Nov-Jun	0.1087	0.0761	0.1412	0.0000
Dec-Jun	0.1551	0.1174	0.1929	0.0000
Sep-Jul	0.0667	0.0348	0.0985	0.0000
Oct-Jul	0.0836	0.0517	0.1155	0.0000
Nov-Jul	0.0983	0.0663	0.1304	0.0000
Dec-Jul	0.1448	0.1075	0.1821	0.0000
Nov-Aug	0.0708	0.0387	0.1028	0.0000
Dec-Aug	0.1173	0.0799	0.1546	0.0000
Dec-Sep	0.0782	0.0408	0.1155	0.0000
Jul-Apr	0.0576	0.0259	0.0894	0.0000
Oct-Aug	0.0560	0.0241	0.0880	0.0000
Dec-Oct	0.0612	0.0238	0.0986	0.0000
May-Mar	0.0506	0.0187	0.0824	0.0000
Jun-Apr	0.0473	0.0151	0.0796	0.0001
Mar-Jan	0.0470	0.0115	0.0825	0.0009
Jul-May	0.0416	0.0097	0.0735	0.0012
Dec-Nov	0.0465	0.0090	0.0839	0.0029
Sep-Aug	0.0391	0.0072	0.0710	0.0036
Aug-Jun	0.0379	0.0055	0.0702	0.0072
Apr-Mar	0.0345	0.0028	0.0662	0.0193

Analysis:
In the main ANOVA table, a difference in mean delivery times through the months is shown as significant if P<=0.05.

The box plots was used to visualize the monthly distributions.

6.4 Part 6 Conclusion

The ANOVA (Analysis of variance) was therefore successful to identify the reason for the instability in the part 3 SPC charts. The first ANOVE compared the delivery hours by year and found a p-value of 0.672 and thus that it is not the reason. The second ANOVA compared the monthly delivery times and yielded a p-value of 0 indicating that it is definitely the reason for the SPC instability.

The Analysis of Variance (ANOVA) successfully identified the root cause of the instability seen in the Part 3 SPC charts.

The "out-of-control" upward trend in data points in part 3 is due to the upward trend in monthly delivery times and not just random noise. This is a predictable seasonal effect that repeated from 2022 to 2023. It is thus unable due to the fact that the seasonality is not accounted for and thus VOC is not met.

7 Part 7: Reliability of Service

Part 7 analyze the reliability of the service of a car rental company depending of the number of agents working. The goal is to optimize the staffing in order to optimize the cost and thus the profit.

7.1 Estimated Service Reliability

The graph given shows a distribution of workers present during a time period of 397 days. At miniumum 15 people is needed.

Table 20: Summary of Observed Service Reliability (Basic)

Metric	Value
Total Days Observed	397
Days with Reliable Service (>=15 Staff)	366
Proportion Reliable	92.2%
Estimated Reliable Days per Year	336.5

Therefore, we estimate that reliable service should be expected on approximately 92.2% days per year.

7.2 Profit Optimisation (Binomial Model)

We need to find the optimal number of personnel to assign (S) to minimize total daily costs. Problems occur (costing R20,000/day) if fewer than 15 workers are present (X<15). Personnel cost R25,000/month each.

Model: X B(S,p), where S is assigned personnel, p is probability present.

Assumptions:

p=0.90 (assumed constant).

Costs (daily): Cprob= R20,000; Cperson = R r round(25000 / (365.25/12), 2).

Objective: MINIMIZE E[Cost(S)]: E[Cost(S)]=(S×Cperson)+(P(X<15 S,p)×Cprob) where P(X<15 S,p)=P(X 14 S,p), calculated using pbinom(14, size=S, prob=p).

Table 21: Cost Optimization vs. Assigned Personnel (S), p=0.9 (Basic)

Assigned (S)	P(Staff < 15)	E[Problem Cost]	Staffing Cost	E[Total Cost]
14	1.00	20000.00	11498.97	31498.97
15	0.79	15882.18	12320.33	28202.51
16	0.49	9705.44	13141.68	22847.13
17	0.24	4764.06	13963.04	18727.10
18	0.10	1963.94	14784.39	16748.33
19	0.04	703.88	15605.75	16309.63
20	0.01	225.06	16427.10	16652.17
21	0.00	65.46	17248.46	17313.92
22	0.00	17.57	18069.82	18087.39

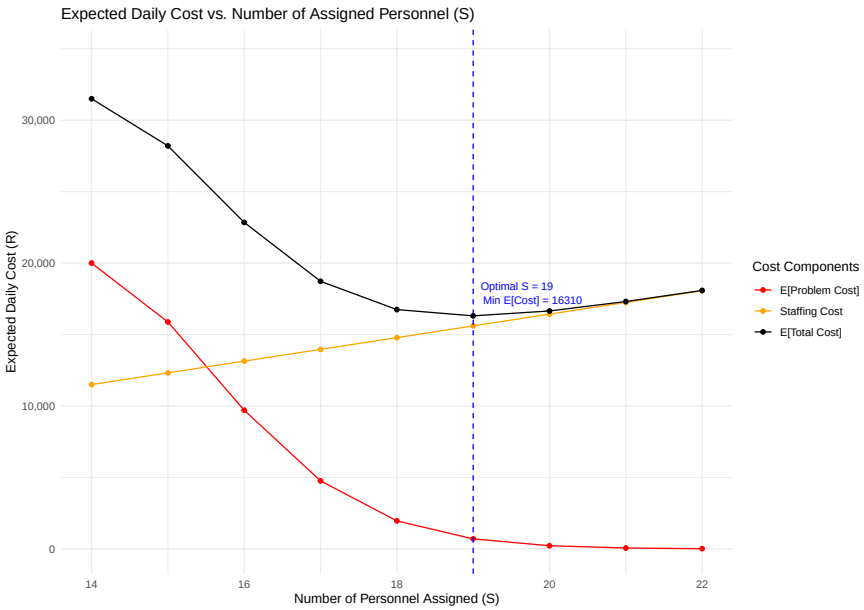


Figure 3: Cost Optimization Plot

Analysis: Expected Daily Cost vs. Number of Assigned Personnel (Line):
The expected daily cost vs the number of assigned personal plot illustrates that the optimal amount of workers is 19. This is the point where the additional staffing cost and lessened problem costs is in equilibrium. A minimum cost of R16,310 is expected at 19 workers.

Based on the binomial model…
* The minimum total expected daily cost occurs when assigning **19** personnel.
* The minimum total expected daily cost is approximately **R16310**.

7.3 Part 7 Conclusion

Firstly part seven concludes that 92.2% of the days a quality service was expected. It is further found that in orde4r to optimize profit by reducing costs the trade off between service issues cost and workers cost is at 19 workers.

8 Project Conclusion

This project fulfilled the ECSA GA4 by executing a comprehensive data analysis on different topics.

Part one established a clean data analysis on the product, customer and sales data of a company. Part two was not given. Part 3 focussed on SPC graphs to determine the delivery reliability of a company to fulfill the VOC. Part analyzed the risk involved in using rules A,B and C used in part 3 as well as the Type I and Type II risk involved in the SPC analysis done in part 3. Part 5 was an optimization model to maximize the profit of 2 coffee shops by determining the optimal number of baristas. Part 6 proved that the upward trend of delivery times in part 3 was due to a seasonal upward monthly delivery time. Part seven built a model to determine that the optimal employees needed at a car rental company was 19 to minimize the total cost trade-off between service issues cost and employee cost.

Ultimately this project tested a variety of data analysis logics to show a competence in Quality Assurance 344.

9 References & Acknowledgment

All data and problem statements were provided by the course instructor.