OCTOBER 1, 2025

QUALITY ASSURANCE ECSA PROJECT

27094626

Table of Contents

Introduction

A few datasets containing data about customers, products, workers and sales are being investigated. The following report aims to apply Basic Data Analysis methods to the datasets to gain important financial insights and to be able to visualize the data. The goal is to be able to analyse trends and relationships of the respective datasets and also be able to see how they are interlinked.

PART 1

1.2

Data Loading and Inspection

Inspecting the datasets using built-in functions were insightful with respect to seeing what data we are working with. The following table could be coded so that each dataset's dimensions, structure and columns are visible.

	Dataset	Rows	Columns	Column_Names	Var_Types
1	customerdata	5000	5	CustomerID, Gender, Age, Income, City	character, character, numeric, numeric, character
2	productdata	60	5	ProductID, Category, Description, SellingPrice, Markup	character, character, character, numeric, numeric
3	products_Headoffice	360	5	ProductID, Category, Description, SellingPrice, Markup	character, character, character, numeric, numeric
4	sales	100000	9	CustomerID, ProductID, Quantity, orderTime, orderDa...	character, character, numeric, numeric, numeric, num...

It is important to note that the "productsdata" dataset is much smaller than the others when the dataset's metrics are compared. Seeing what type of variables are in the columns is also valuable since it indicates which columns might need to be converted to numeric values at a later stage.

Summary Statistics

Using dplyr::summarise() function, code was developed to get the mean and the standard deviation of the columns of the datasets. Comprehensive summaries were also retrieved using skimr::skim() as well as frequency counts for categorical features.

Summaries of the customer dataset led to the following conclusions:

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
Age	1	5000	51.55	21.22	51	50.88	26.69	16	105
Income	2	5000	80797.00	33150.11	85000	81665.00	37065.00	5000	140000

The target demographic can be identified as a middle-aged demographic, since the mean is 51. This can give the company incentive to focus and tailor marketing around this age group through television, newspapers and radio instead of social media, since social media is used less by this demographic.

The wide standard deviation in income is also proof that the customer base is varied and that the company should be able to offer budget and luxury services. Median income is higher than mean income which means the left-skewed distribution, this correlates with low income outliers. Therefore these customers need to be accommodated by tiered pricing.

Summaries of the products dataset led to the following conclusions:

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
SellingPrice	1	60	4493.59	6503.77	794.18	3189.25	525.72	350.45	19725.18
Markup	2	60	20.46	6.07	20.34	20.51	7.31	10.13	29.84

The average markup of 20% which indicates premium pricing. The large standard deviation shows that the prices vary, which is good since the company serves a diverse market. Customer segmentation is very important in this case.

Summaries of the head office products dataset led to the following conclusions:

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
SellingPrice	1	360	4410.96	6463.82	797.22	3054.23	515.75	290.52	22420.14
Markup	2	360	20.39	5.67	20.58	20.43	6.66	10.06	30.00

The mean price and markup of the products dataset and the head office products is similar which means that pricing discrepancies is not really a problem. The slightly lower standard deviation of the head office products' selling price could imply that local branches need to be standardized to promote consistency.

Summaries of the sales dataset led to the following conclusions:

	vars	n	mean	sd	median	trimmed	mad	min	max
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Quantity	1	1e+05	13.50	13.76	6.00	11.46	5.93	1.00	50.00
orderTime	2	1e+05	12.93	5.50	13.00	13.12	5.93	1.00	23.00
orderDay	3	1e+05	15.50	8.65	15.00	15.50	10.38	1.00	30.00
orderMonth	4	1e+05	6.45	3.28	6.00	6.45	4.45	1.00	12.00
orderYear	5	1e+05	2022.46	0.50	2022.00	2022.45	0.00	2022.00	2023.00
pickingHours	6	1e+05	14.70	10.39	14.05	13.54	6.92	0.43	45.06
deliveryHours	7	1e+05	17.48	10.00	19.55	17.78	8.90	0.28	38.05

n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
0	1	13.50347	13.7601316	1.0000000	3.000000	6.000	23.00000	50.0000	▇▁▁▁▁
0	1	12.93230	5.4951268	1.0000000	9.000000	13.000	17.00000	23.0000	▂▅▇▇▅
0	1	15.49683	8.6465055	1.0000000	8.000000	15.000	23.00000	30.0000	▇▇▇▇▇
0	1	6.44813	3.2834460	1.0000000	4.000000	6.000	9.00000	12.0000	▇▆▆▆▇
0	1	2022.46273	0.4986115	2022.0000000	2022.000000	2022.000	2023.00000	2023.0000	▇▁▁▁▇
0	1	14.69547	10.3873345	0.4258889	9.390833	14.055	18.72167	45.0575	▃▇▂▁▁
0	1	17.47646	9.9999440	0.2772000	11.546000	19.546	25.04400	38.0460	▅▃▇▇▂

Mean delivery hours are longer than picking hours which means that a there is a bottleneck at delivery that needs to be improved by better systems or management. The mean order time, day and month is also an important statistic for being able to make accurate forecasts. Midday, middle of the month and June and July are the most popular times and this needs to be taken into account with production planning.

Handling missing values

A quick analysis was performed to see if there are any missing values that can skew the results.

The summary statistics and the is.na() function confirmed that all the datasets does not contain any missing values so no imputation is needed to remove missing values and there no risk of having skewed results.

Data filtering and subsetting

As a data analyst that aims to make accurate business recommendations the following questions were considered:

Are male or female customers ordering more? The following table was obtained that shows that female customers order more, but the order quantity between male and female is more of less the same. This means that having male specific marketing could increase sales.

Gender	Total_Quantity	Total_Orders	Unique_Customers	Avg_Quantity_per_Order
<chr>	<dbl>	<int>	<int>	<dbl>
Female	679382	49288	2432	13.78392
Male	618251	46476	2350	13.30259
Other	52714	4236	218	12.44429

The second question that was considered what the sales performance was per city. The results showed that Los Angeles has the highest sales performance. These regions need to be prioritized with strategic resource allocation. Underperforming regions' offers might need to be localized.

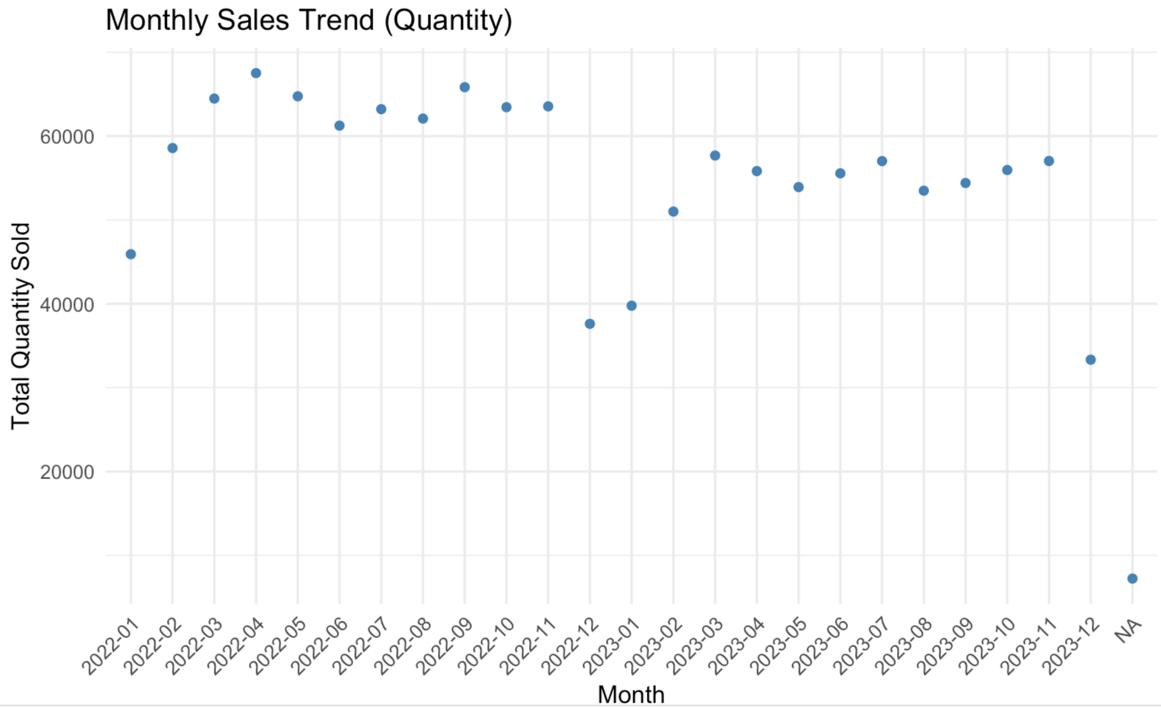
City	Total_Orders	Total_Quantity	Total_Revenue	Unique_Customers	Avg_Revenue_per_Order
<chr>	<int>	<dbl>	<dbl>	<int>	<dbl>
Los Angeles	14978	219425	722173478	726	48215.61
San Francisco	15803	216887	674061345	780	42654.01
New York	14423	188441	623940704	726	43260.12
Houston	14325	189672	598904598	724	41808.35
Seattle	13593	187569	587080997	673	43189.95
Chicago	14062	175784	574464154	724	40852.24
Miami	12816	172569	571962403	647	44628.78

By merging the sales and product data the product which sells the most can also be identified. The product that sells the most is aliceblue marble keyboards. Monitors are also very popular. It is important to identify popular products so that more of them can be produced and so that these products can be promoted.

ProductID	Description	Category	Total_Units_Sold
<chr>	<chr>	<chr>	<dbl>
MOU059	aliceblue marble	Keyboard	29675
SOF001	coral matt	Software	29336
SOF004	blue silk	Monitor	29219
SOF010	chocolate sandpaper	Monitor	29168
MOU058	azure sandpaper	Monitor	28924
MOU054	coral marble	Mouse	28875
MOU052	azure matt	Monitor	28804
SOF007	black bright	Software	28517
MOU057	aliceblue marble	Laptop	28423
SOF005	aliceblue wood	Keyboard	28412

Data visualization

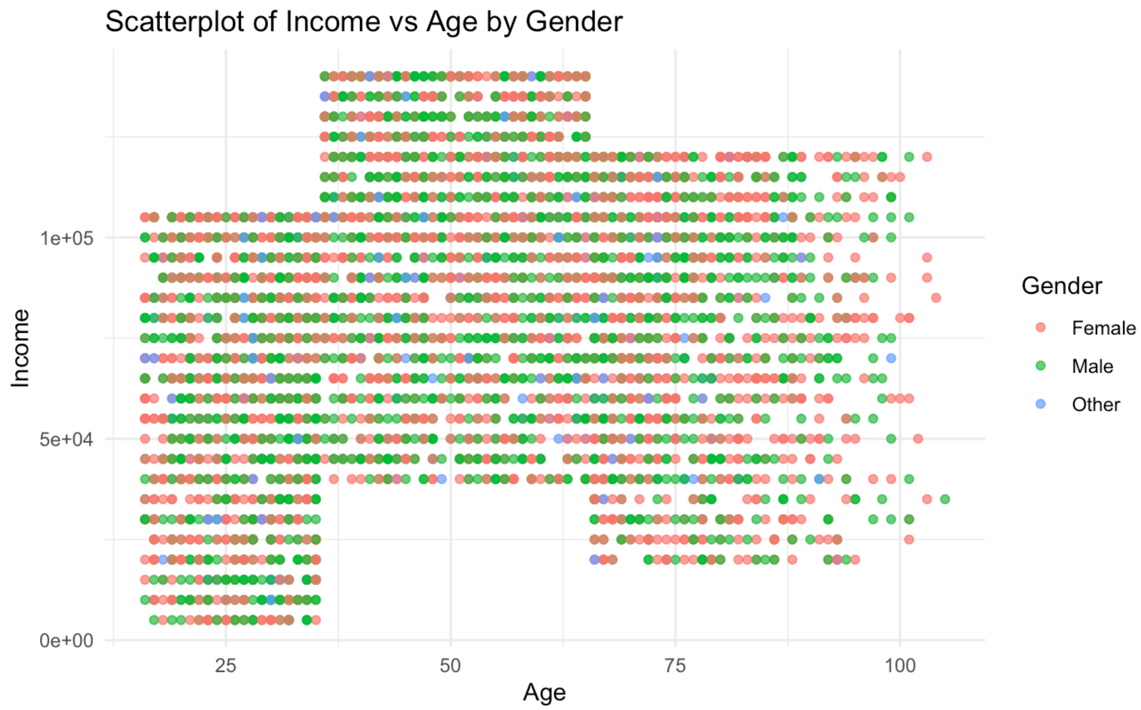
The sales trends over time can be visualized by adding a Date column.



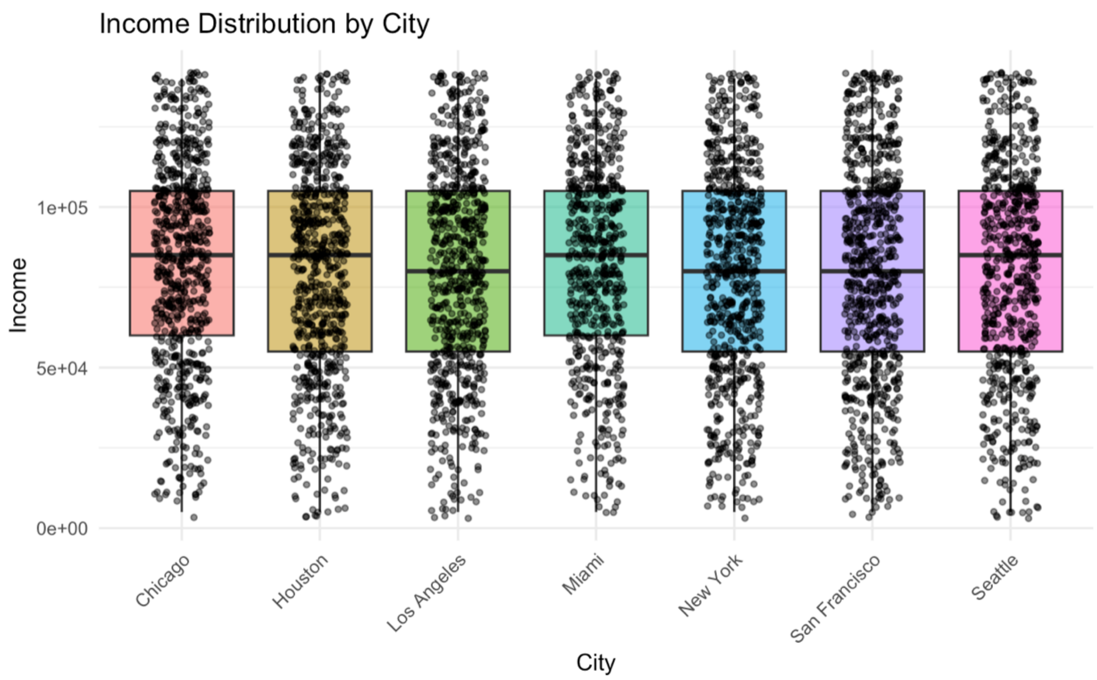
By analysing this graph it is clear that there is definite seasonality since it is visible in both years. In March and April the sales are at their highest in both years. The low quantity sold in January and December is apparent in both years and this must be addressed by having specials or promotions that are holiday specific. Upselling or personalised product recommendation using algorithms can boost sales. (Apsis.com, 2020) Bundling is also a possible solution.

It is also clear that the 2023 sales is slightly lower than 2022. This negative trend is not good. It would be recommended to focus on bestselling items and to make improvements to the supply chain in order to gain a competitive edge.

A scatterplot of the relationship between income and age is also very insightful. This scatterplot shows that the average income is much higher between about 40 and 60. This demographic needs to be targeted in this specific company by the types of products that are being sold and by marketing.

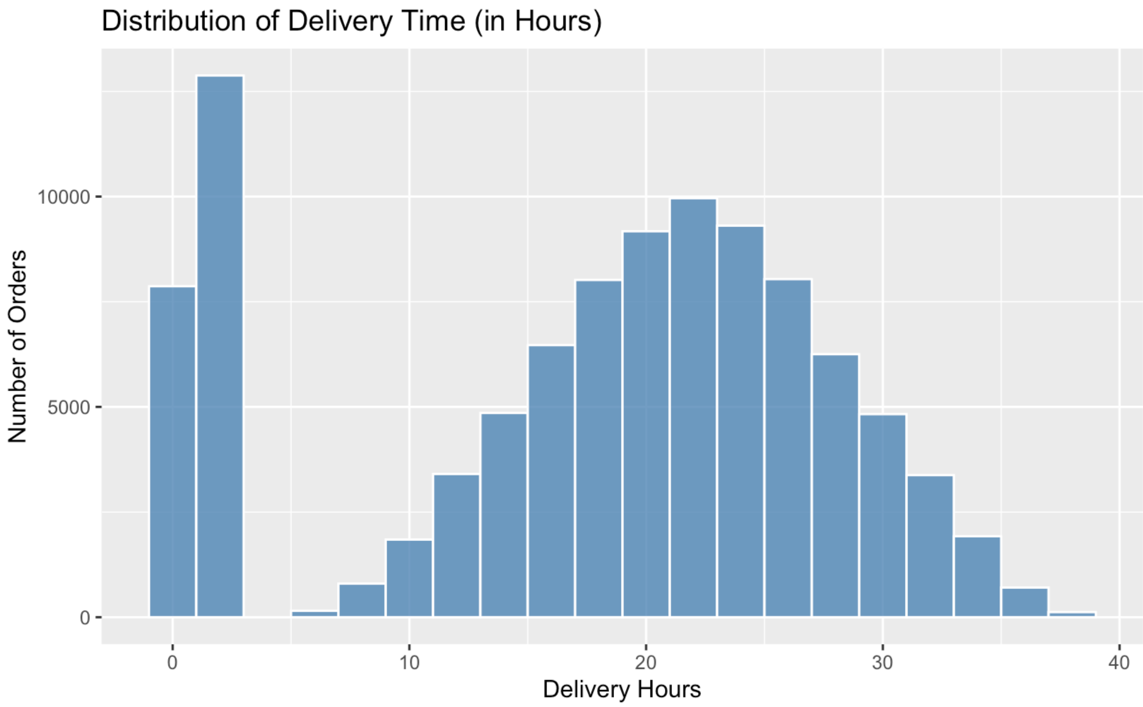


This boxplot shows income distribution per city. It shows that Miami has the least variability and has the highest median income. Miami is a good option for target



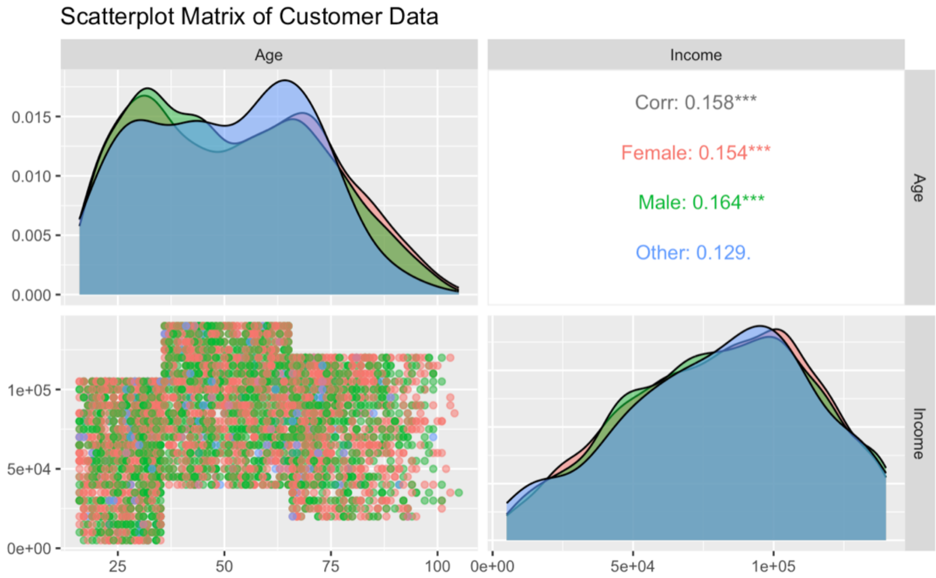
marketing and product positioning. It would be smart to allocate more resources towards cities that have large variability like Los Angeles and Houston. Diverse customer bases could benefit from tiered products and pricing strategies.

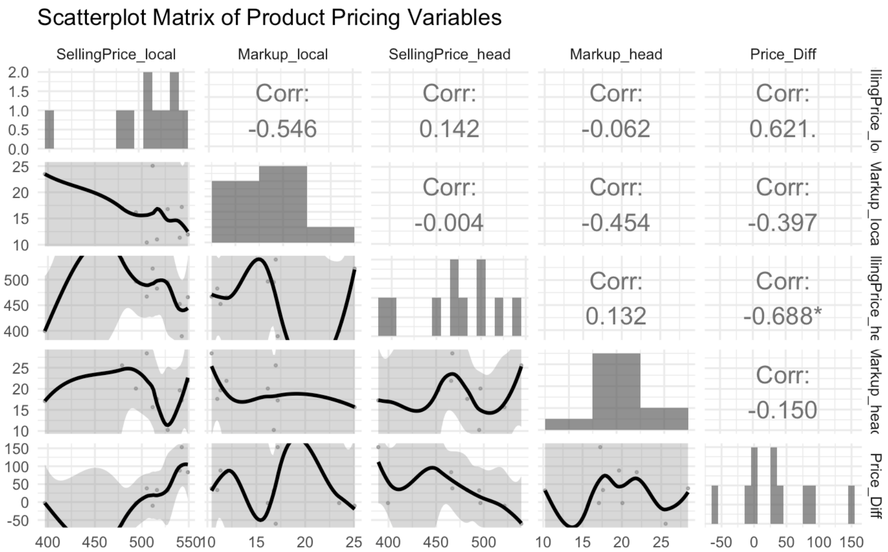
This histogram that shows the distribution of delivery times is also valuable. This shows that a large number of orders are deliverd instantly. The other peak is between 20-22 hours and the bell shaped curve shows consistent delivery performance. It would be good to create a target delivery time of 20 hours to be able to track performance. The deliveries that take longer than 30 hours need to be investigated and optimized by finding out which orders are taking so long.



Exploring relationships

This scatterplot matrix shows the relationship between age and income in the customer dataset. It shows that as age increases income increases regardless of gender. It also shows that there are more customers in the 30-60 age since diagonal density plots is right skewed. This plot can help tailor product offers and communication strategies.





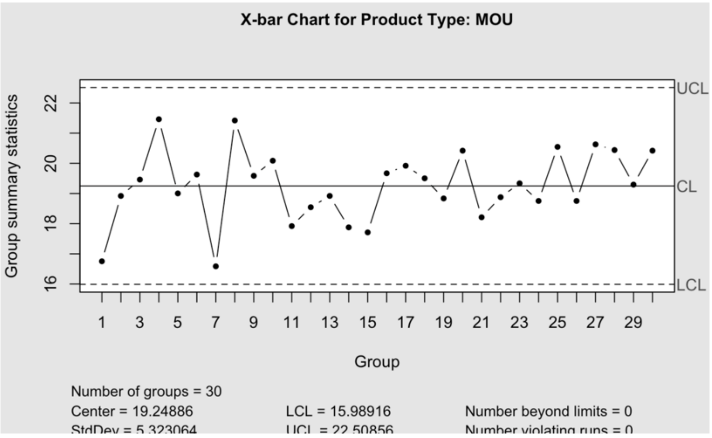
The second scatterplot matrix compares the local prices and the prices at head office. It shows that they are normally distributed and with some variation. An important note is that the as the local selling price increases the local markup tends to decrease. Markup at head office is not closely corelated with any of the other factors.

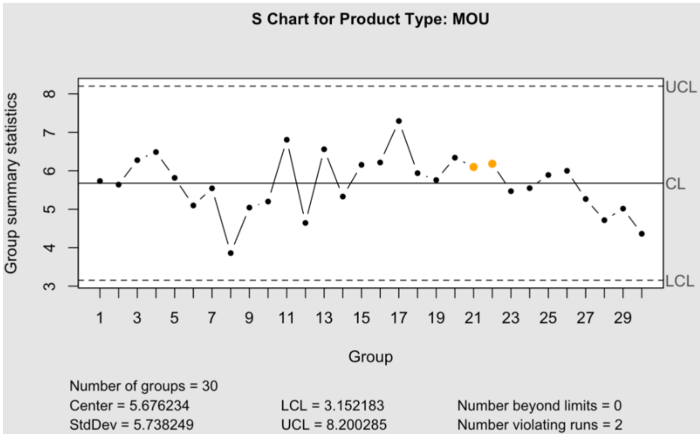
PART 3

3.1

By using statistical process control using the delivery times for each of the six products we can learn about the mean delivery time and variability from the xbar charts and the S Charts. Since SPC charts identify problems early it is crucial to analyse them.

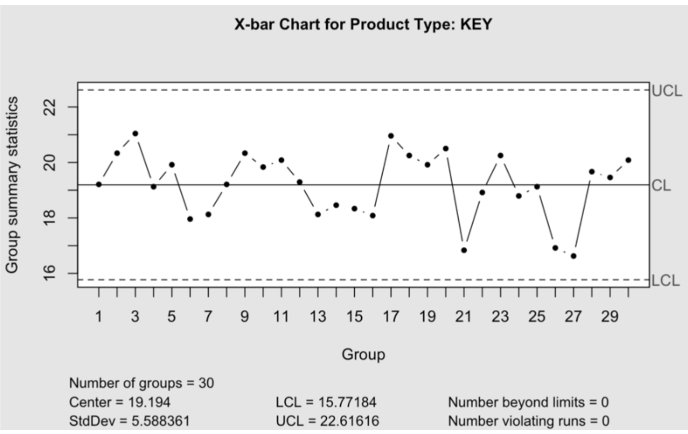
For Product type MOU:

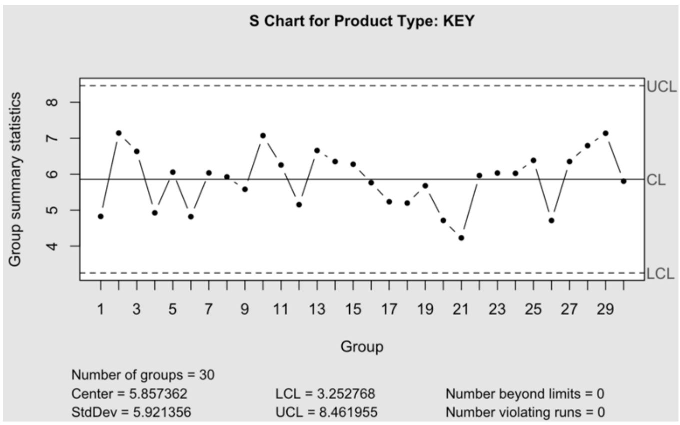




All the points are within the control limits for the xbar chart, so the process in statistically in control. There is no unusual patterns and there is consistent delivery performance thus far. On the S chart the orange dots serve as warning since it is violating runs. The last six samples that are showing a decreasing trend could point to process tightening.

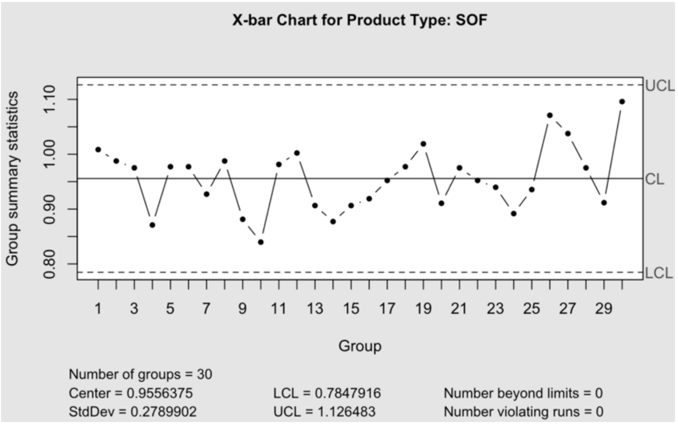
For product type KEY

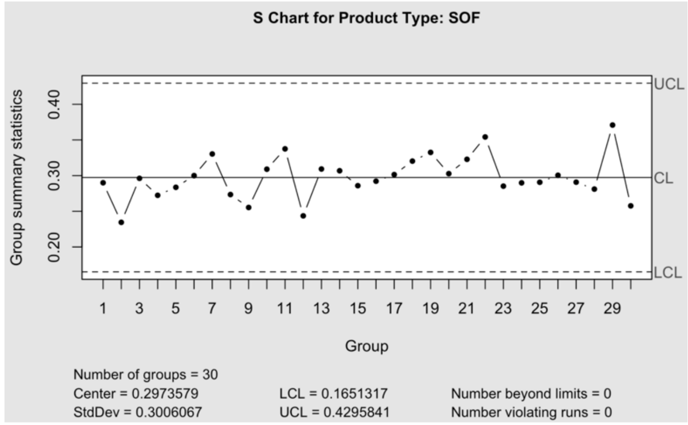




Since some of the delivery times hit 2 sigma it could be worthwhile to investigate causes like inventory control and supplier delays.

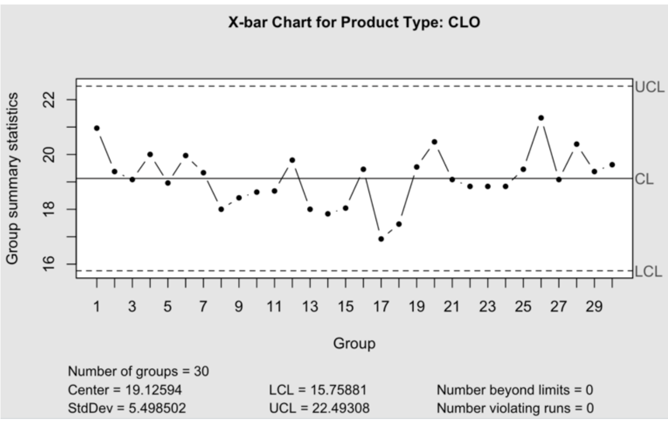
For product type SOF:

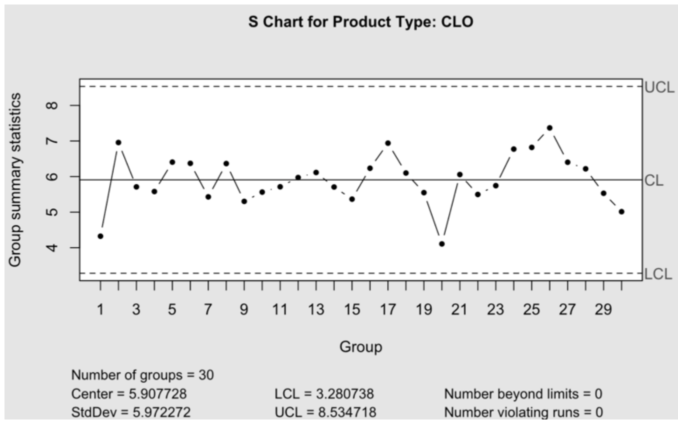




Software is an online transaction which justifies the short delivery times. This product therefore has the lowest variability and although the average delivery times are shifting a bit over time, but the average is rather consistent.

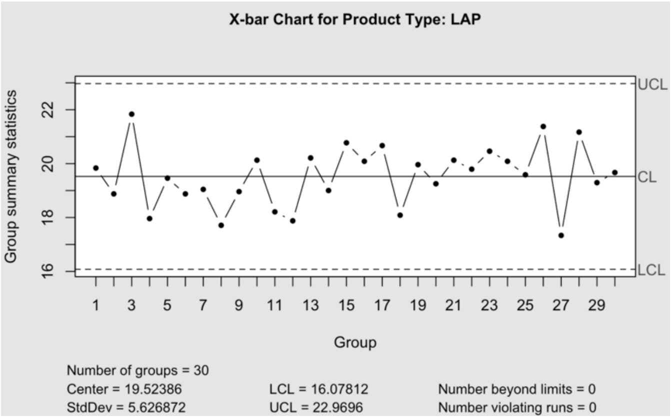
For product type CLO:

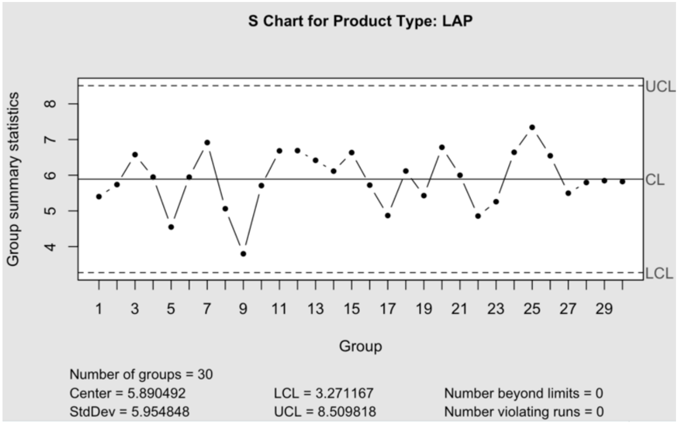




For product CLO the process is statistically under control since nothing is beyond the limits. This product’s xbar chart also has the least number of fluctuations. This correlates to being able to predict delivery time accurately, which is valuable for the customers.

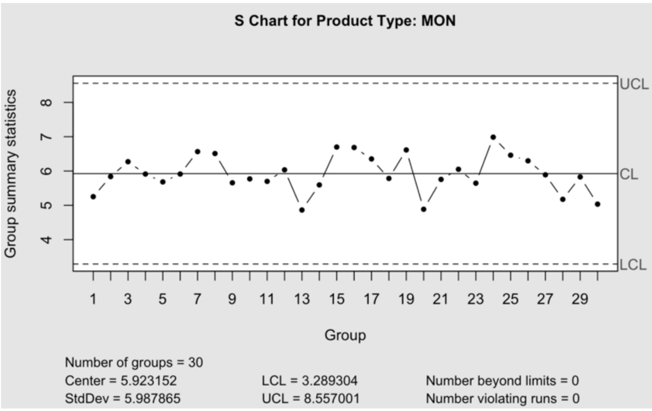
For product type LAP:

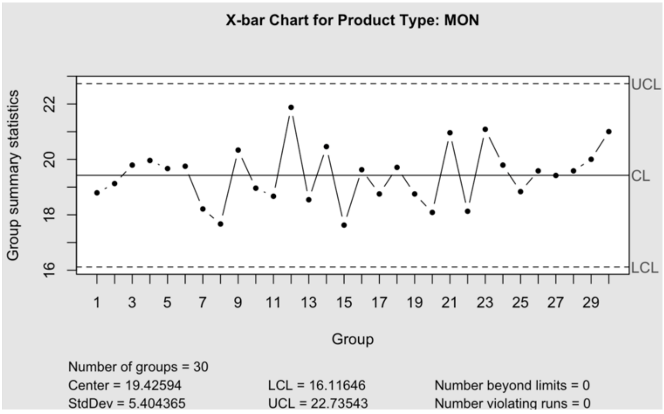




The S chart shows that the consistency of delivery times does not change a lot. There is some slight variation between 1 and 2 sigma, but still controlled. This product has the highest average delivery time. This can be addressed by identifying bottlenecks or speeding up the picking process with better movement in the warehouse.

For product type MON:





This product has the shortest average delivery time. This is crucial to keep in mind when managing inventory and delivery. There is quite large variability, but still not outside the limits. This could indicate that process redesign is needed.

3.2

For product type MOU:

It can be seen that the process remains under control and a slight downward trend can be seen in the xbar chart which means that the average delivery time is being decreased. This can be due to better process control over time like training or automation. If this shift is permanent it might be worth updating the limits. There is more variation, therefore processes need to be standardized and equipment maintenance is important.

For product type KEY:

The graphs does not exhibit definite changes. Therefore the processes are still the same.

For product type SOF:

A decreasing trend can be seen in the xbar chart and one sample point is on the lower limit. This is a warning that should be investigated. Since software is an online transaction it could indicate that the technology was updated to be more responsive.

For product type CLO:

The graphs exhibit no noteworthy changes that needs to be flagged. The trends and patterns are relatively consistent.

For product type LAP:

A smaller variation can be noticed. This is positive since delivery time can be predicted accurately for customers which leads to higher customer satisfaction. The average delivery time stays the same.

For product type MON:

The x bar graph exhibit stays consistent. The trends and patterns are relatively consistent. Larger variation can be witnessed and this could indicate that the process needs to be controlled so that predicted delivery times are more accurate.

3.3

	ProductType <chr>	Cp <dbl>	Cpl <dbl>	Cpu <dbl>	Cpk <dbl>	Capable <lgl>
1	MOU	0.915	1.104	0.727	0.727	FALSE
2	KEY	0.917	1.105	0.729	0.729	FALSE
4	CLO	0.898	1.079	0.717	0.717	FALSE
5	LAP	0.899	1.101	0.696	0.696	FALSE
6	MON	0.889	1.079	0.700	0.700	FALSE

None of the products are capable for Voice of customer since all the Cp values are below 1.0 indicating wide process variation and it will not stay between the specification limits of 0 to 32 hours. The target Cp is 1.33, since it is a well-controlled process that will lead to customer satisfaction. It is not ideal that customers receive delivery times beyond the limits frequently.

The Cpk values are all under 1.0 and is off-centered. CPU values are particularly low. This correlates to a slow process that is too variable near the upper limit.

All products have a lower Cpu than Cpl which indicates that the deliveries are too slow for VOC.

Deliveries needs to happen faster. A few recommendations to attain this would be to reduce variation by standardizing delivery processes and continuous flow (Amper.co, 2018). The company should aim to improve reliability by improving scheduling. Training needs to be improved so that staff can work faster and more efficiently. Implementing 5S workplace organization could significantly improve quality and efficiency. (Amper.co, 2018) The LEAN principles can be applied to identify wastes like motion, transport and waiting (Skhmot, 2017). These can be adressed by automation, balancing workload or redesigning layout of product flow. A just-in-time inventory model can also be beneficial.

3.4

What we can learn from implementing rule A, B, and C is the following:

4.2

Rule A showed that none of the samples were outside the upper +3 sigma control limits which means that there is not excessive variation and that the process is rather stable with regards to delivery hours.

Rule B revealed that all the products have a run of 30 consecutive samples. All the runs start at 1 which means there is early control. We learn that the processes are predictable and well-controlled.

Rule C had no output which correlates to consistent and cantered delivery times.

From this we can learn that the processes are currently of high quality and that they are consistent. There are no signs of outliers or a trend shift. We can see an opportunity for scaling of operations by increasing order volumes.

PART 4

4.1

The output that was given looks as follows.:

```
Rule A Type I Error Probability: 0.0027
Rule B Type I Error Probability: 0.02196
Rule C Type I Error Probability: 2.702336e-07
```

Rule A is highly unlikely to have a false alarm. This rule is reliable and strict.

Rule B's probability for a Type I error is not too rare. This rule could correlate with good control and does not really happen by chance.

Rule C has minimal chance for a Type I error, therefore when this rule is triggered there is a possibility that the process has shifted. It will not trigger by chance.

4.2

The Type II Error Probability (beta) is 0.84118. From this we can learn the chance that a process shift goes undetected. Since the probability is high it is likely that a shift will be missed. This can lead to issues with quality, which will badly impact customer satisfaction. It would be necessary to adjust the control limits to gain sensitivity of the SPC control charts. It would also be recommended to monitor the processes very finely in order to detect changes earlier.

4.3

After addressing the discrepancies the products_Headoffice.2025.csv and products_data2025.csv where created. Basic data analysis techniques from part 1 can be applied. The following changes in the results were noticed. Data loading and inspection can be condensed into the following summary table.

Dataset <chr>	Rows <int>	Columns <int>	Column_Names <chr>	Var_Types <chr>
products_data2025	60	5	ProductID, Category, Description, SellingPrice, Markup	character, character, character, numeric, numeric
products_Headoffice2025	300	5	ProductID, Category, Description, SellingPrice, Markup	character, character, character, numeric, numeric

Summary statistics

For products_Headoffice2025

Mean_SellingPrice <dbl>	SD_SellingPrice <dbl>	Median_SellingPrice <dbl>	Mean_Markup <dbl>	SD_Markup <dbl>	Median_Markup <dbl>
4493.593	6460.12	794.185	20.46167	6.031842	20.335

For products_data2025

Mean_SellingPrice <dbl>	SD_SellingPrice <dbl>	Median_SellingPrice <dbl>	Mean_Markup <dbl>	SD_Markup <dbl>	Median_Markup <dbl>
4493.593	6503.77	794.185	20.46167	6.072598	20.335

It is evident that the cleaned up datasets are more accurate when comparing it with the initial results. The Mean selling price is identical and the standard deviation of the markup and of the selling price is much less different, which indicates that the discrepancies were correctly handled.

The total sales per type using the updated prices is summarised in this table.

Category <chr>	TotalSalesValue <dbl>
LAP	2470814376
MON	1258942847
CLO	214110418
KEY	155002210
SOF	142527355

The products_Headoffice2025 and products_data2025 datasets contain no missing values, which means that there is not a chance for skewed results.

The data filtering and subsetting revealed that the gender that buys the most is still female.

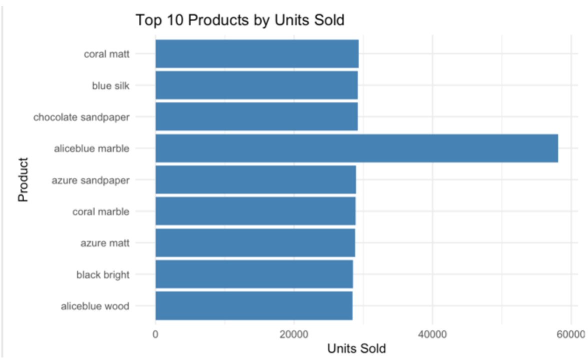
Gender <chr>	Total_Quantity <dbl>	Total_Orders <int>	Unique_Customers <int>	Avg_Quantity_per_Order <dbl>
Female	537298	39039	2432	13.76311
Male	489612	36896	2350	13.27006
Other	42137	3403	218	12.38231

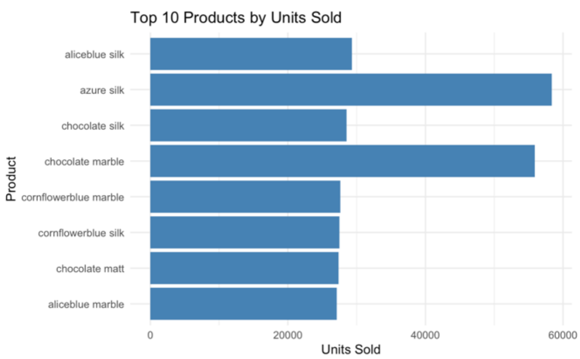
Comparing the product summary with the Part 1's answers some noticeable changes are evident. Even though the city that generates the most revenue is still Los Angeles, the average revenue per order has increased. The fixed dataset also causes a slightly

City <chr>	Total_Orders <int>	Total_Quantity <dbl>	Total_Revenue <dbl>	Unique_Customers <int>	Avg_Revenue_per_Order <dbl>
Los Angeles	11908	173643	704087676	726	59127.28
San Francisco	12460	171413	656036338	780	52651.39
New York	11471	149817	608642651	726	53059.25
Houston	11376	149250	582912268	724	51240.53
Seattle	10819	149261	571969602	673	52867.14
Chicago	11190	138685	559837583	724	50030.17
Miami	10114	136978	557911088	647	55162.26

lower total quantity and total revenue. It is good that these discrepancies were sorted out so that data analysis can be more accurate.

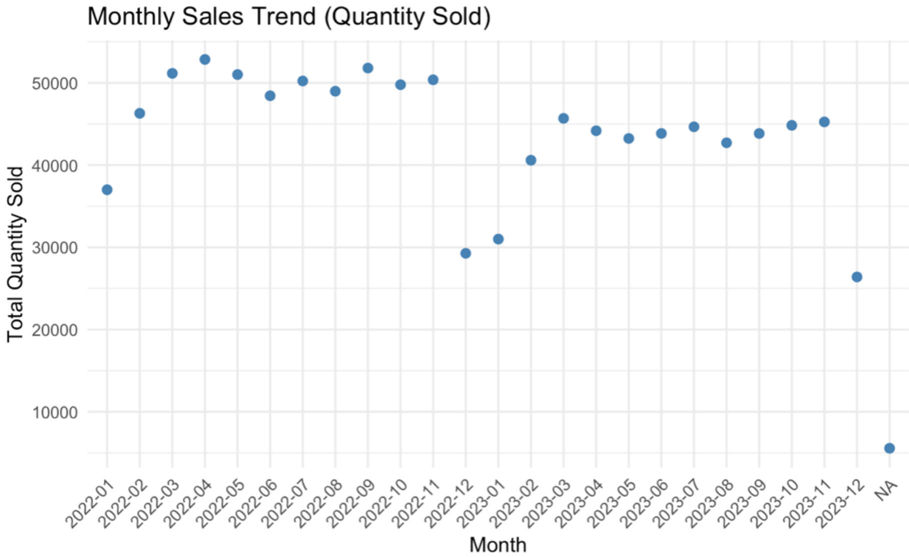
The graphs before fixing the data and after fixing the data shows significant changes in the top products. Previously the top product was aliceblue marble, but in the new graph it can be seen that azure silk and chocolate marble software generates the most revenue. This information is insightful and the business should put a focus on producing and updating these software products.



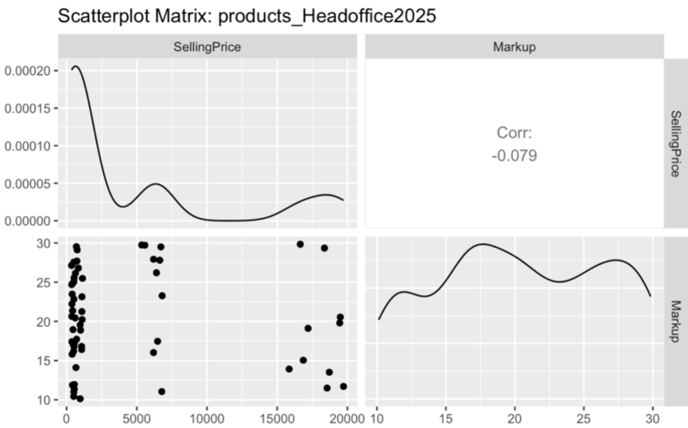


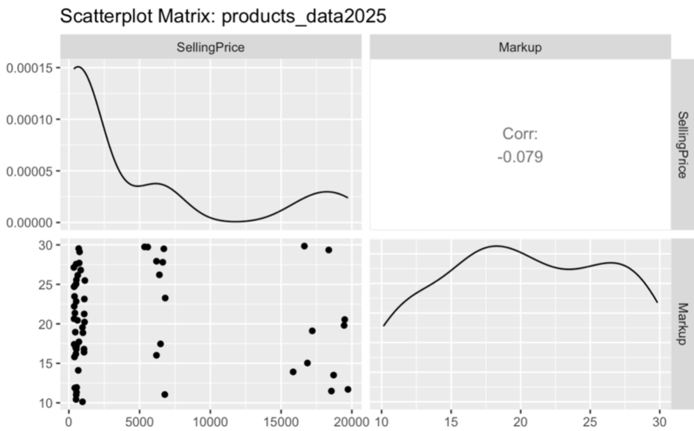
Data visualization

The visualization shows that the trend does not change too much, but the values are much lower than previously determined.



Data exploration





The scatterplot matrices show two identical scatterplots, which means that the damage control was worthwhile and the discrepancies were successfully straightened out.

PART 5

5. Optimising Shop 1 and Shop 2 led to the following results:

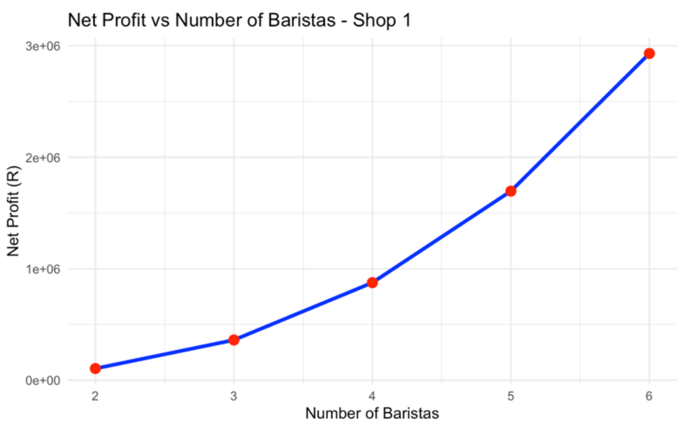
The optimal number of baristas for both shops is 6

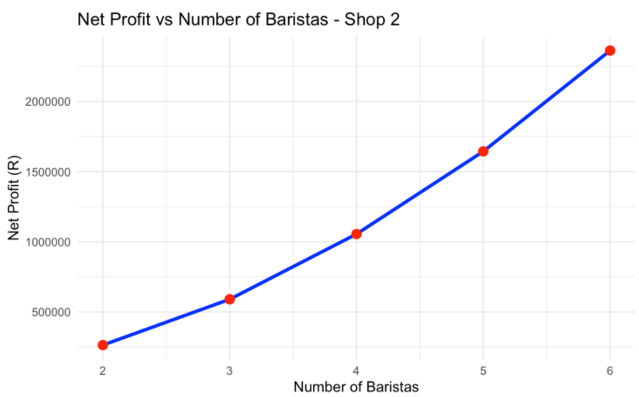
Reliable Service is 100% for both shops.

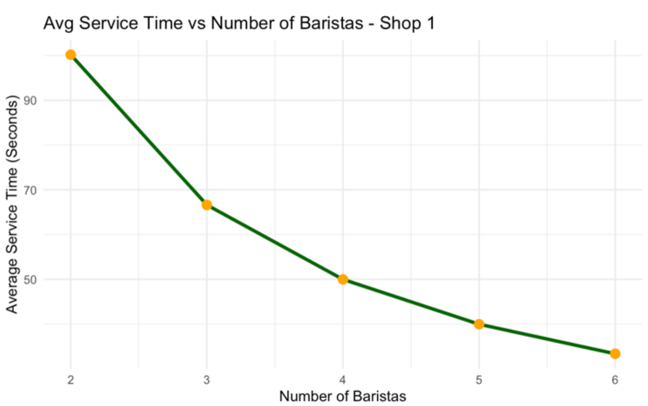
Max profit for Shop 1: R2930850

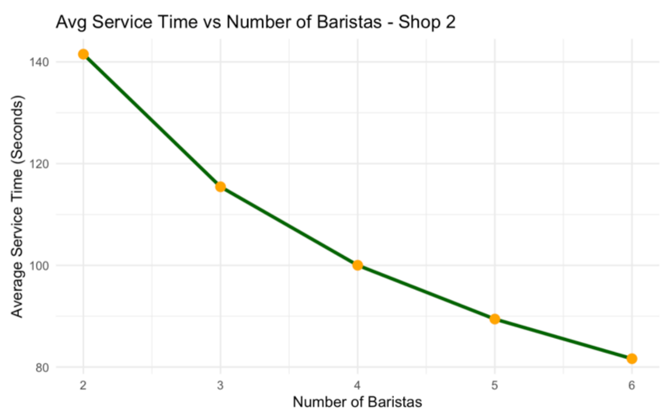
Max profit for Shop 2: R2361900

The problem was constrained to have at least two people that are on duty. A service threshold of 60 seconds was chosen, because that is the longest time that customers expect to wait at a quality coffee shop. Challenges can be addressed by making the workspaces compact and organized so that the workers have easy access to all tools and can work efficiently. Continuous monitoring to identify bottlenecks within the process will enable dynamic management of the employees and operating procedures. It can be seen that the average service time decreases as the number of baristas increases.









PART 6

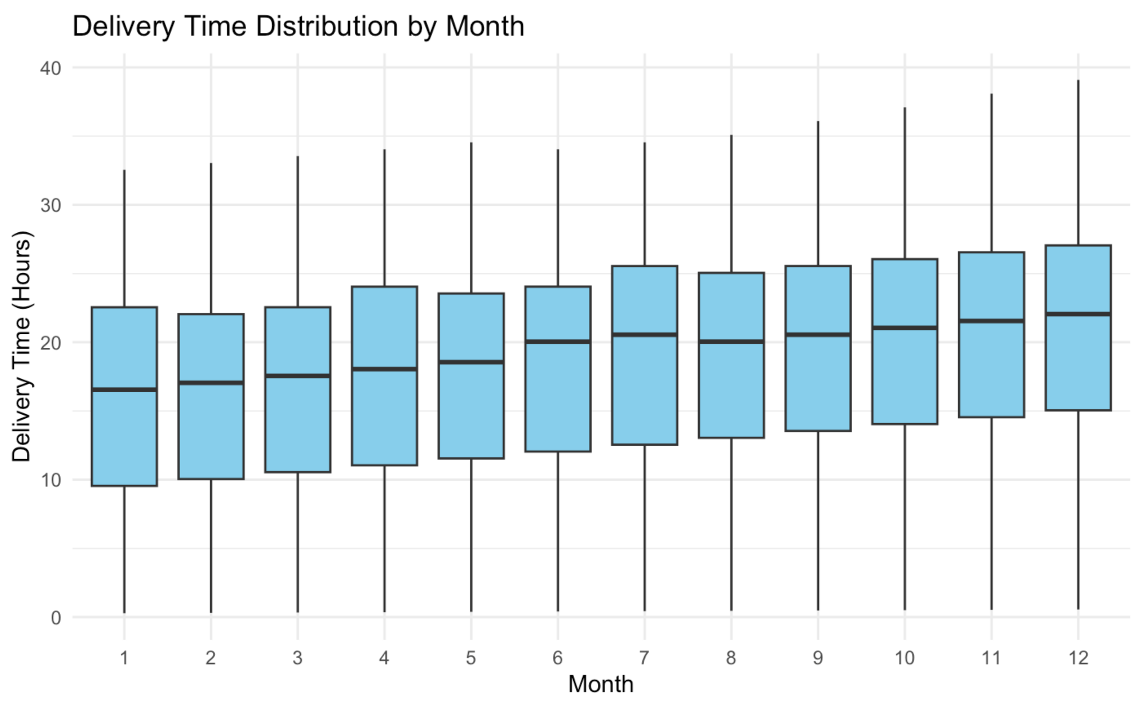
6. DOE and ANOVA

Is delivery times the same across different months?

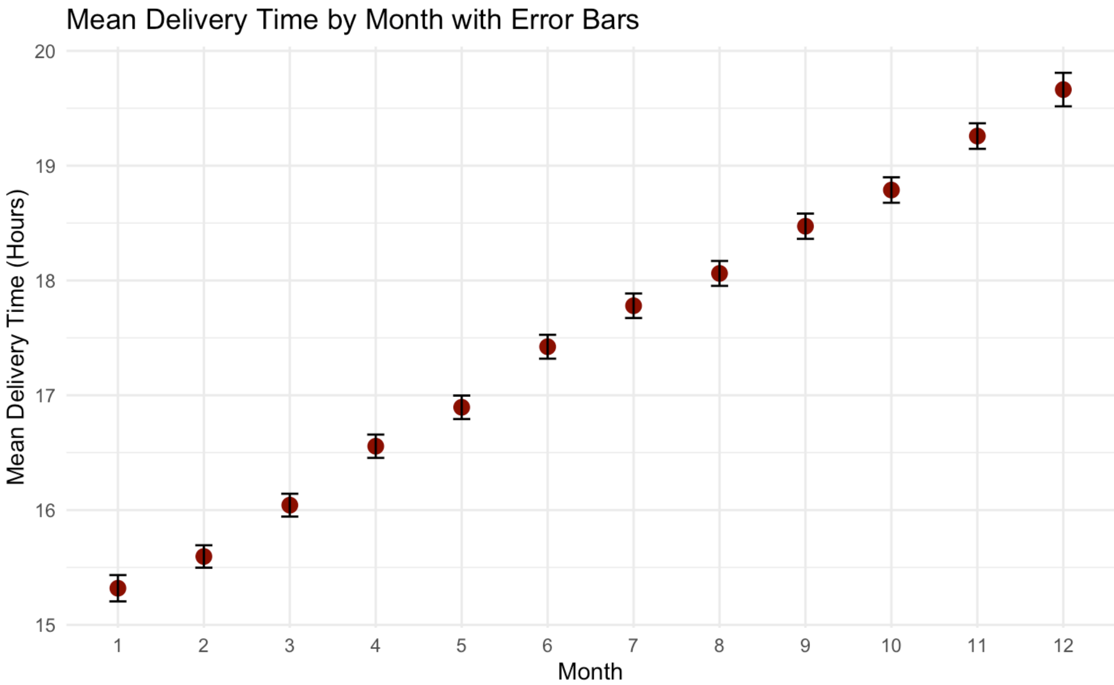
```
'Results in a standard ANOVA table. p-values => Type I errors"
[,1]        [,2]         [,3]    [,4]      [,5]     [,6]
"Source"    "SS"         "DoF"   "MS"      "fo"     "P-value"
"Treatment" "125804.85"  "11"    "11436.8" "116.61" "0"
"Error"     "6241367.87" "63636" "98.08"   "---"    "---"
"Total"     "6367172.72" "63647" "---"     "---"    "---"

'LSD value is: 0.3163"
```

A P value of zero allows is to reject the null hypothesis which assumes that all months have the same average time. The conclusion is that the difference in delivery across months is statistically extremely significant.

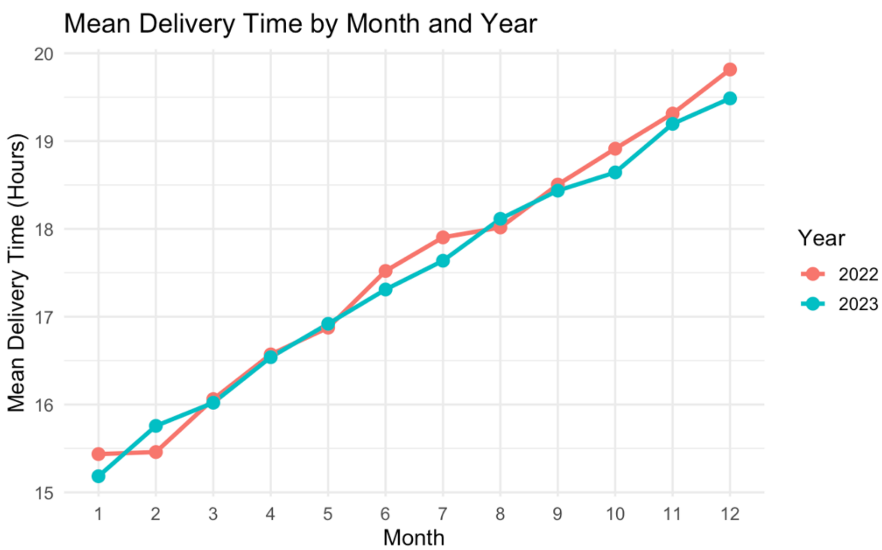


After setting up an ANOVA, this graph was obtained. We can witness an increasing trend in the median delivery times over the year. This can be from a few different causes, but it points out that there are operational inefficiencies. The delivery performance seems to be worsening as the year progresses. This needs to be addressed by making sure that staff fatigue is combatted and that the business processes can handle a rising demand due to seasonal reasons. The interquartile range also seems to be slightly wider towards the end of the year. This could suggest that there is more variability and inconsistency in the delivery times in the later months. This is also not desirable, since customers value consistent and accurate delivery promises. The company may have to



look at expanding their staffing reallocating resources.

This graph displays the upward trend very clearly and the small error bars are indicative of low variability across the months. The increase is not random, but rather consistent and systematic.



This graph provides a way to compare the years and to analyse consistency of the trend. The rising delivery trend towards the end of the year is an emerging problem that needs to be addressed by transport efficiency and better demand forecasting. It can also be seen that 2023 is below 2022, which indicates a slight improvement. Holiday peaks need to be planned for and bottlenecks need to be sorted out.

PART 7

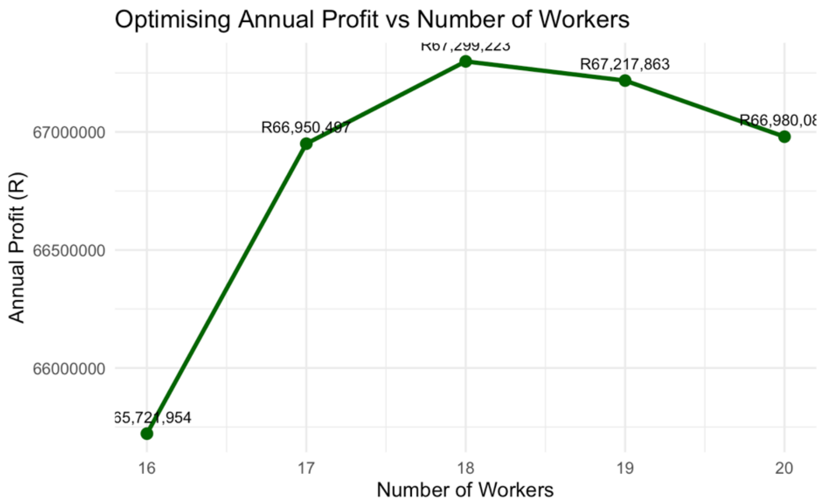
7.1

The total number of days that was observed was 397 days. To make an estimation, a predefined threshold of 15 workers on a given day was set for reliable service, since it allows responsiveness and efficiency. The probability of reliable service was calculated as 92,2%. Which means in a standard year of 365 days only 29 days will have unreliable service.

7.2

To find the optimal profit firstly p was estimated and expected reliability was calculated. After computing annual losses and simulating extra staff levels the annual profit for each case can be computed. The estimated attendance probability is 92.5% and the probability of having 15 or more workers on duty is 66.05%. The expected loss is

R2 478 046 per year. At 18 workers the reliability becomes 95.88% and an optimal profit of 67 299 223 can be reached. Adding workers to the 16 original staff members will minimise unreliable days and maximise profit. The problematic days is reduced from 124 days to 15 days.



This graph shows sweet spot between loss and extra cost for hiring new employees. Therefore the final recommendation would be to hire two additional workers.

Conclusion

In conclusion, the data revealed valuable insights. It allowed for business analysis and understanding after implementing descriptive statistics. An analysis on the business in part 1, 3 and 4 shows that the ages of 30-60 is the demographic that contributes the most to sales. The business strategy and focus can be adapted to accommodate this tendency, by utilising communication channels that will reach these age groups and having age specific promotions. The products that are most popular is the aliceblue marble keyboard, the coral matt software and the blue silk monitor. To increase profitability the company needs capitalise on these products. By making improvements to the supply chain processes the cost can be reduced significantly. A recommendation would be to look into where wastage happens in these operations. The drop in sales from 2022 to 2023 needs to be addressed by operational improvements. We were also able to identify seasonal trends. This is helpful to in forecasting and predicting future demand. Identifying the average delivery and pick up times also helps to be able to provide accurate lead times to customers and increase responsiveness so that the company can gain a competitive edge. By assessing overall market potential more aggressive sales can be promoted. (Efficy.com, 2024). The last two profit optimization questions also involves calculating the trade-off between service loss and extra cost.

References

Apsis.com. (2020). *6 Ways to Beat January Sales Slumps | Apsis*. [online] Available at: https://www.apsis.com/blog/6-e-commerce-tips-survive-dark-days-january [Accessed 1 Oct. 2025].

Efficy.com. (2024). *Set Your Sales Goals Using 4 Steps to Help Your SME Grow Quickly | Efficy*. [online] Available at: https://www.efficy.com/sales-goals/ [Accessed 1 Oct. 2025].

Amper.co. (2018). *9 real-life Lean Manufacturing examples*. [online] Available at: https://www.amper.co/post/lean-manufacturing-examples.

Skhmot, N. (2017). *The 8 Wastes of Lean*. [online] The Lean Way. Available at: https://theleanway.net/The-8-Wastes-of-Lean.