Industrial Engineering
Stellenbosch University

27 October 2025

LOCHNER, A.S. [22992677]

Table of Contents

List Of Figures

Introduction:

Data Analysis is of vital importance in today's business climate. It can help companies optimize management, maximize profit, reduce waste and gain a competitive advantage. This report explores such analysis and provides the reader useful insight as to how we analyze data and what there is to gain from the analysis.

Part 1 – Descriptive Statistics

In this section of the report, we aim to use descriptive statistics to analyze the following datasets provided:

- Customer data
- Product data (normal and head office)
- Sales data (for both 2022 and 2023)

Descriptive statistics enable us to gain helpful insights into the structure (spread) of the datasets, identify key findings regarding the company's sales history and tendencies, and to promote company growth and development with data-driven strategies and advice.

This is achieved by, firstly, loading and merging the datasets. Working with one comprehensive dataset is better than having to contend with multiple different ones and enables us to do statistical analysis with more possible feature combinations, with the aim of gaining richer insights into the data and the companies performance. Data quality assessments can either be done before or after joining the datasets at hand. This is done to identify any missing or duplicate entries in our datasets that might hinder later analysis. Visualization and descriptive analysis of the data is done to better identify and understand trends, seasonality and relationships within the datasets, with the ultimate goal of gaining useful business insights into the company at hand.

Analysis

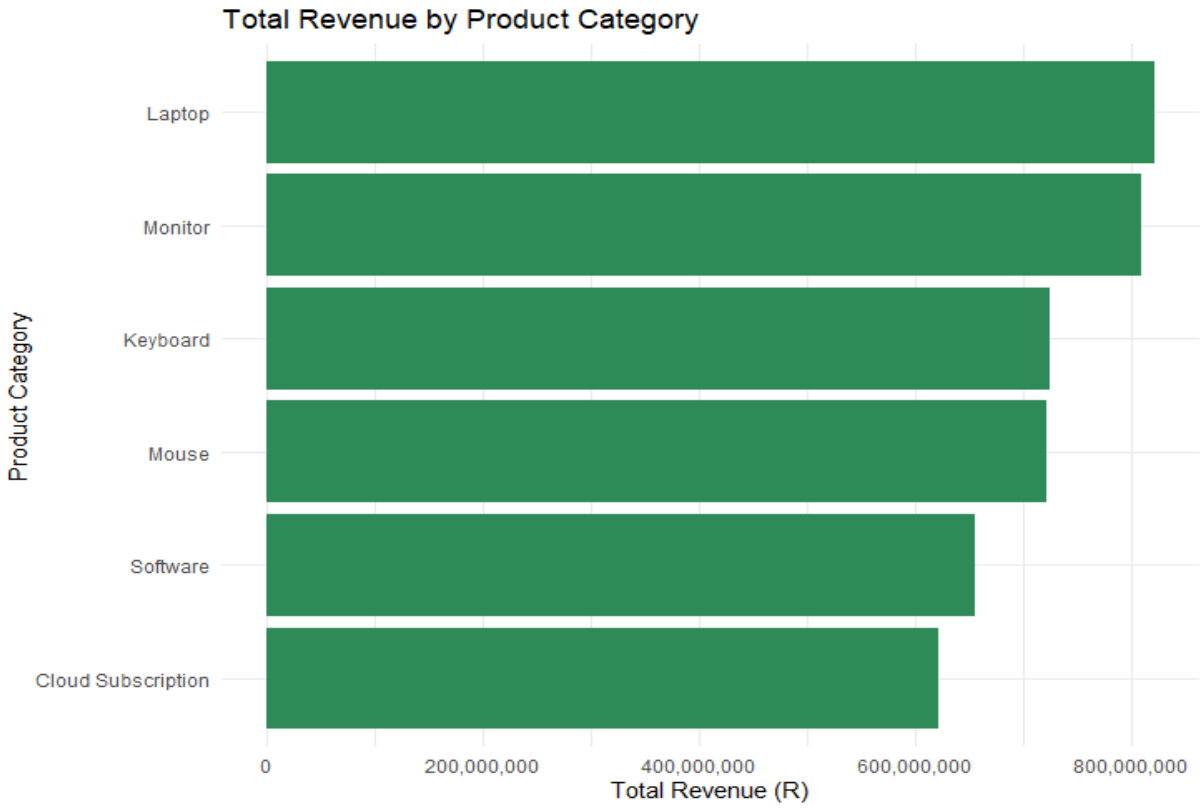

Figure 1 Total Revenue by Product Category

Figure 1 above is our first look at the data itself. We have plotted the total revenue per product category in order to identify top performing categories. Notice that the revenue is not the same for the different product categories. Top category performers include Laptop, Monitor and Keyboard, with Mouse, Software and Cloud Subscription contributing the least to the companies' revenue.

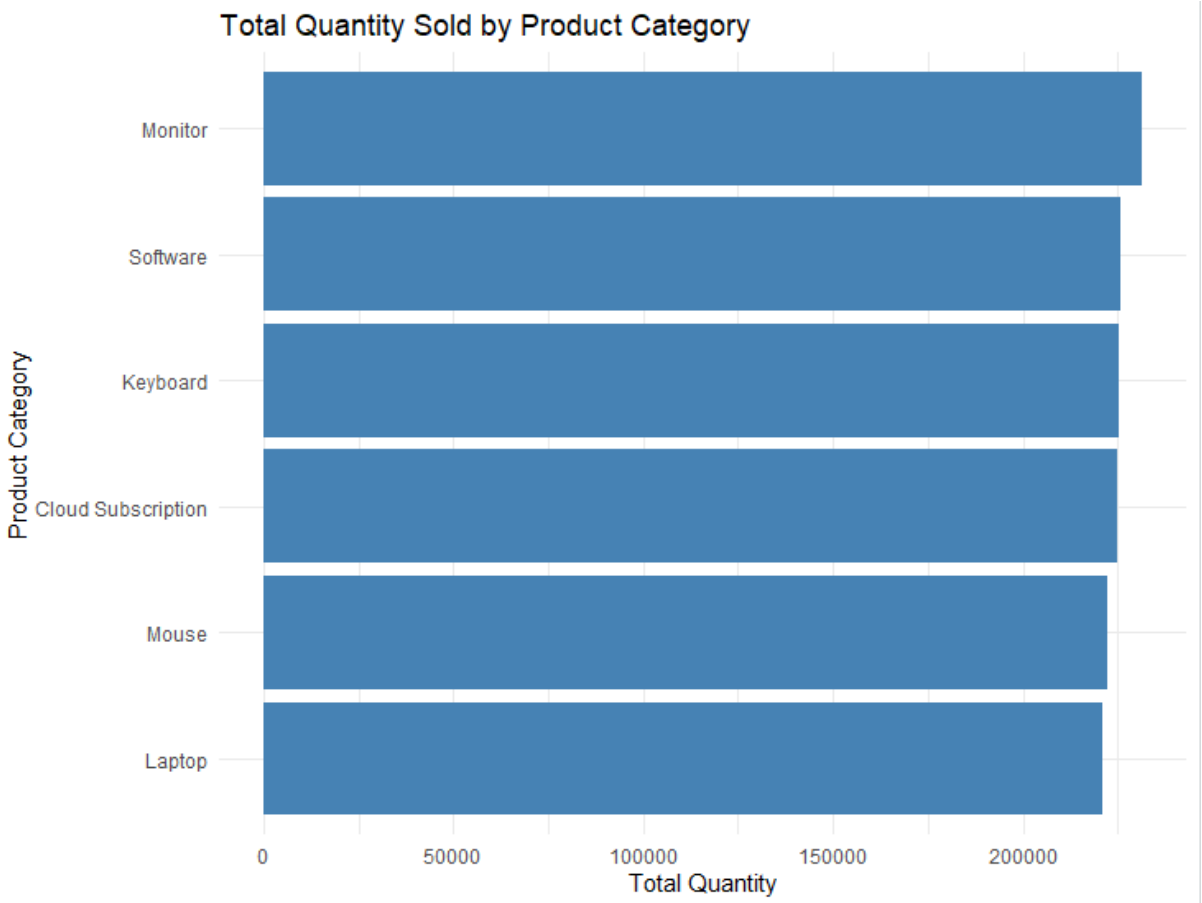

Figure 2 Total Quantity sold by Product Category

Comparing the findings from figure 1 to figure 2 above, we start to make some interesting discoveries. Although Laptop is the category that makes the highest contribution to revenue, it has the lowest quantity of sales across all 6 product categories, hinting to the idea that it sells at a good markup. Monitor has the highest quantity of sales and the second highest contribution to company revenue. It is also worth noting that quantity of product sales across the different categories vary very little, i.e., the company sells almost the same amount for all the respective product categories.

Let us see if plotting the quantity of sales (across all categories) versus the city in which it is bought can give us any useful insights as to whether there are some cities with higher buying power than others. This is done to potentially identify important focus areas for the company as a city with high buying power could be the exact one they should focus on if the business is expanding, building new distribution centers, looking to expand to new markets/clientele, etc.

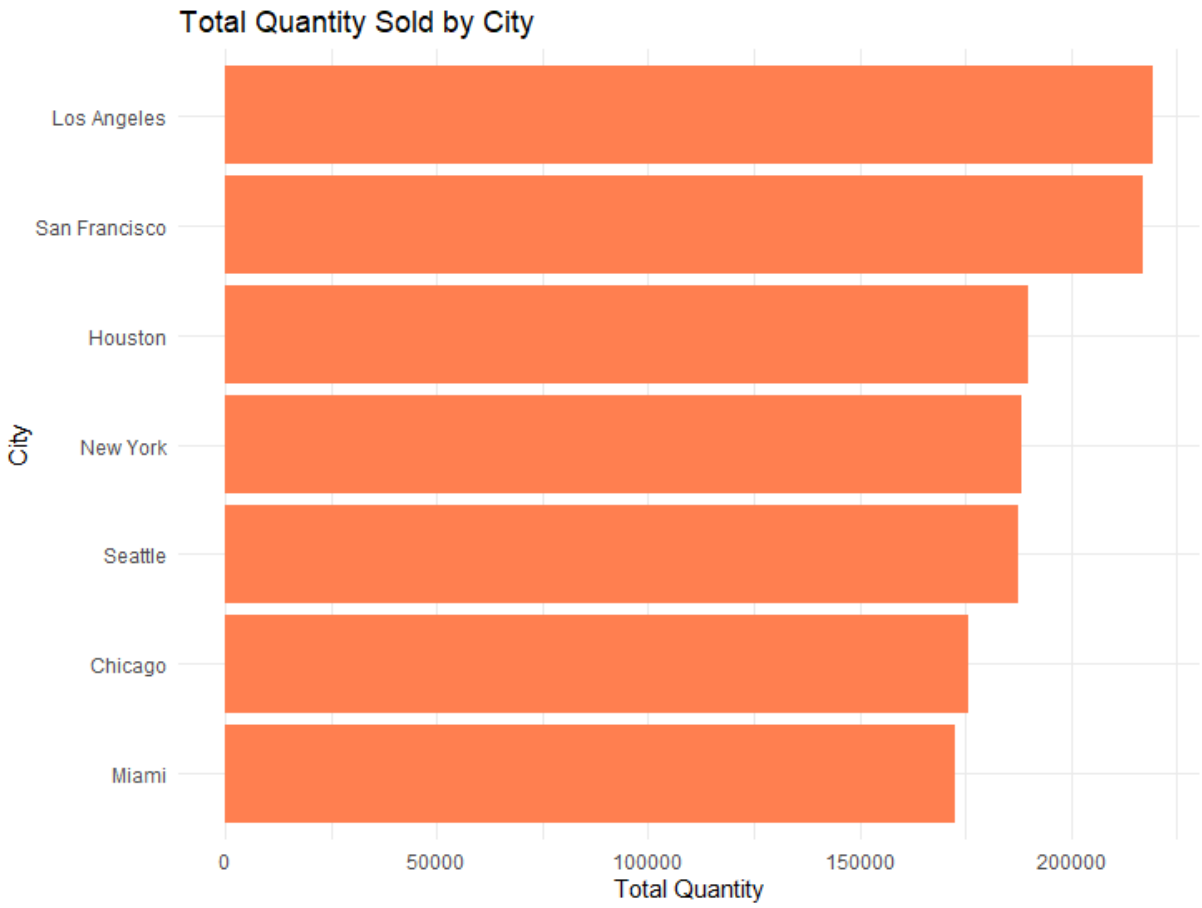

Figure 3 Total Quantity Sold by City

Figure 3 above shows us exactly that: Los Angeles and San Francisco are strong candidates for company expansion logging the highest quantity of sales over the years 2022 and 2023. Houston, New York and Seattle practically tie in third place and Chicago and Miami reported the lowest quantity. Although some customer demographics are included in the customer dataset (such as age, income, etc.) we do not have any information on the population of the cities. This is important to take into account as naturally cities with higher populations can be expected to have higher sales and vice versa.

Operational Efficiency

Operations within the company include product collection, product delivery, and order volumes. Let us take a closer look at what the data tells us about the companies' operational efficiency.

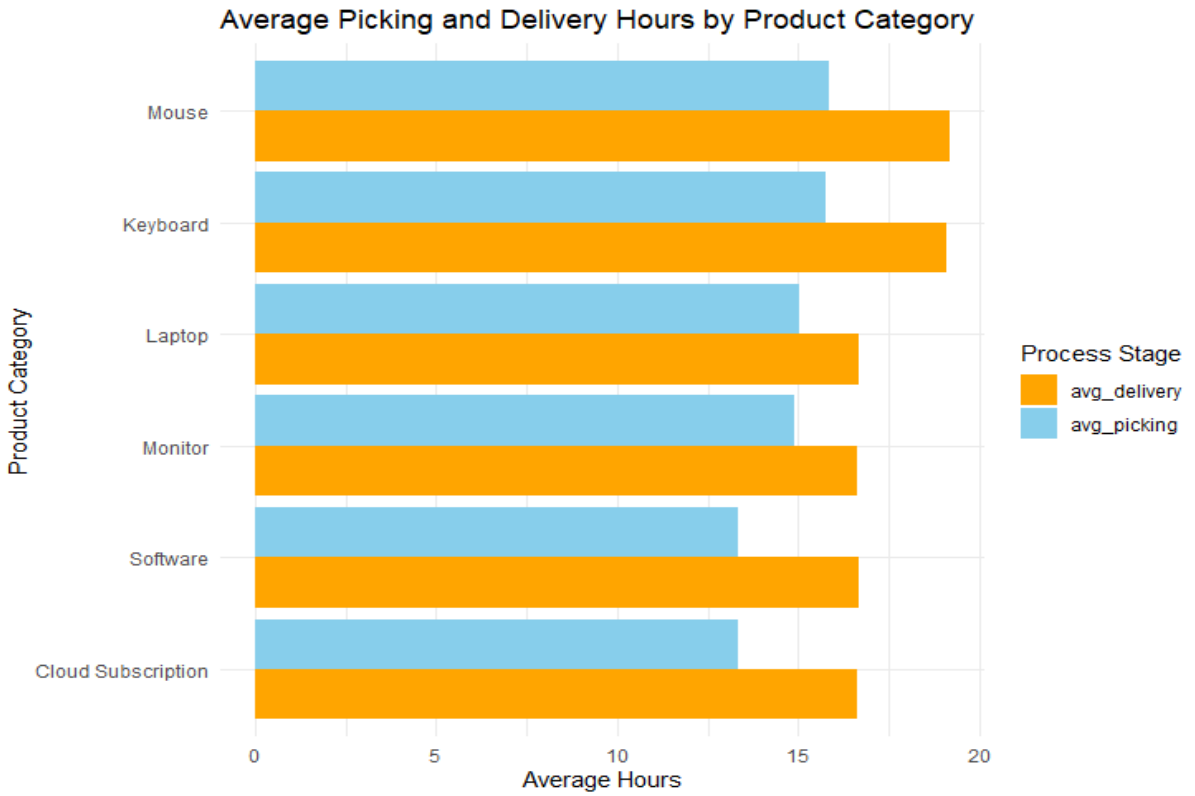

Figure 4 Average Picking and Delivery Hours By Cat.

Figure 4 above clearly shows that product categories Mouse and Keyboard have the highest order lead time (combination of picking and delivery hours), with Laptop and Monitor tying in third place and Software and Cloud Subscription having the fastest lead times. Both Mouse and Keyboard order lead times are high and could be the next focal point for company operational improvement. Delivery hours are also higher than picking hours across the board.

Figure 5 below plots the relationship between Picking and Delivery Hours. The blue points are the actual data points, with the red line showing us the trend between these two features. Notice the slight upward slant of the red trend line, indicating a mild to strong positive correlation between delivery and picking hours. In other words, as picking hours increase delivery hours also increase. From this relationship can be derived that in order to improve the delivery hours for products, we have to lower the picking hours for the respective products. This gives company management useful direction as to where they can focus to improve overall customer satisfaction. They can look at how organized their warehouses are, what type of inventory management they make use off, and how intuitive and sensible the general layout of the warehouse is. If possible, these areas should be developed or improved on.

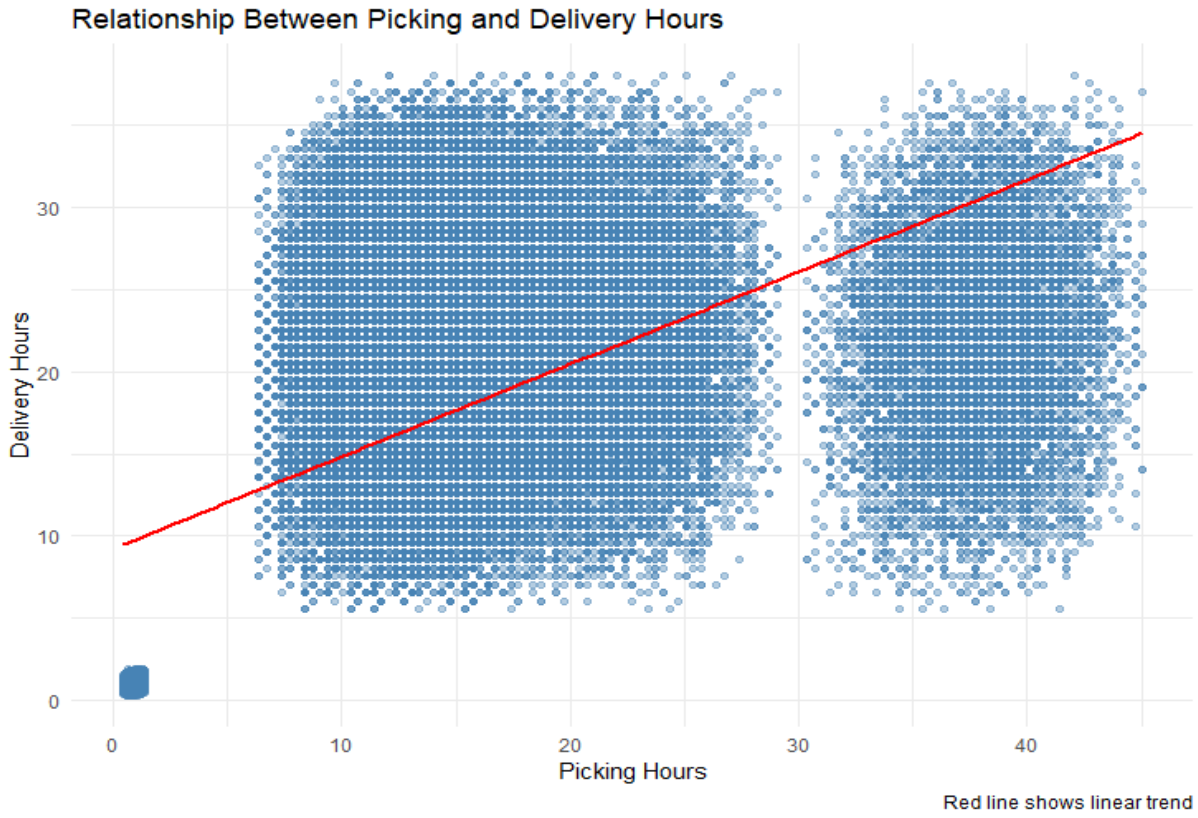

Figure 5 Relationship between Picking and Delivery Hours

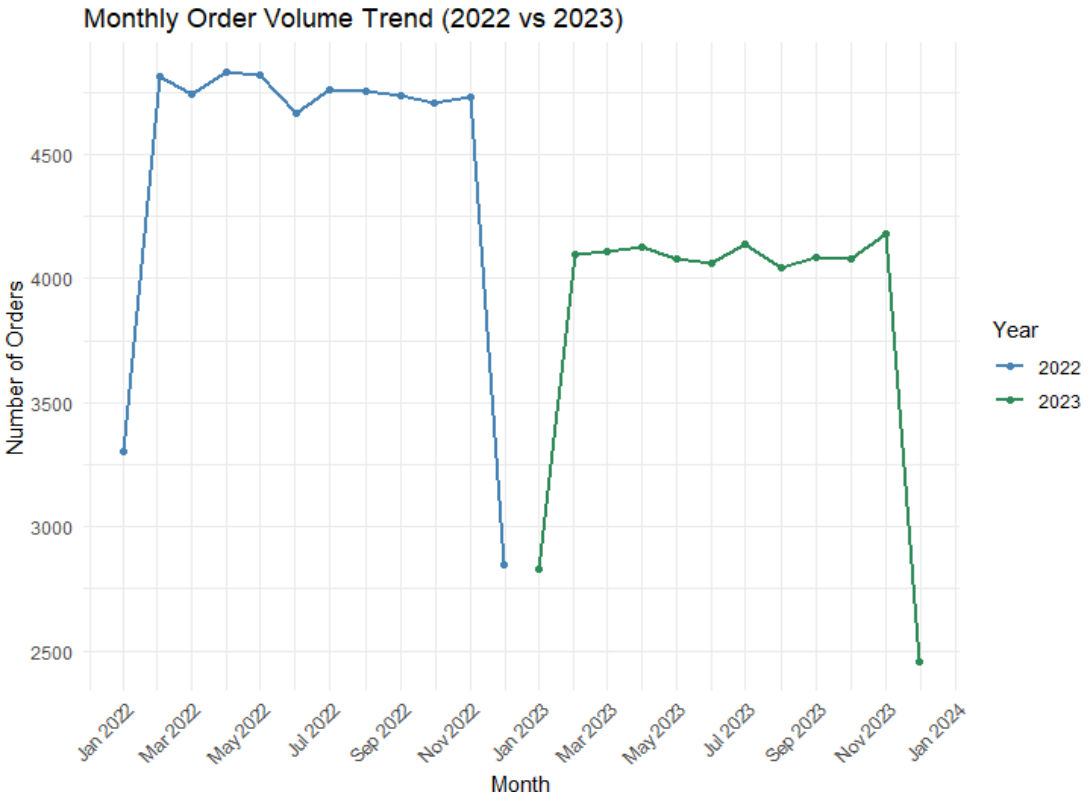

Figure 6 Monthly Order Volume Trend

In figure 6 above, the monthly order volume is plotted over time. In doing so we can identify trends and seasonality in the orders the company receives. For both 2022 and 2023, January seems to be a slow month, with February immediately receiving almost 1500 more orders. The orders then seem to be pretty stable around 4750 per month for 2022 and 4150 per month for 2023. December month is once again very slow in both years. The order volume seems to have a cyclic trend over the years. It is definitely also worth noting that 2023 clearly had less orders than 2022, possible indicating a decrease in product demand, a decrease in customer buying power, or losing customers to out-bidding competitors. Management certainly needs to have a close look at this to identify and address reasons for this decrease in order volumes, as the company preferably wants to see increase over the years.

Customer Segmentation

In this section we explore the customer segmentation of the company to see if any useful trends or patterns can be identified.

Figure 7 Income Distribution by City

Figure 7 above shows us the Customer Income versus the city they live in. At first glance we notice that the spread of the income for the respective cities are very similar, with little to no variation between minimum, 1st quartile, median, 3rd quartile and maximum values. In the previous section we found that Los Angeles and San Francisco were the cities with the highest quantity of products sold, yet now we see that there is almost no financial indication that this would be the case. Figure 8 below further proves our point:

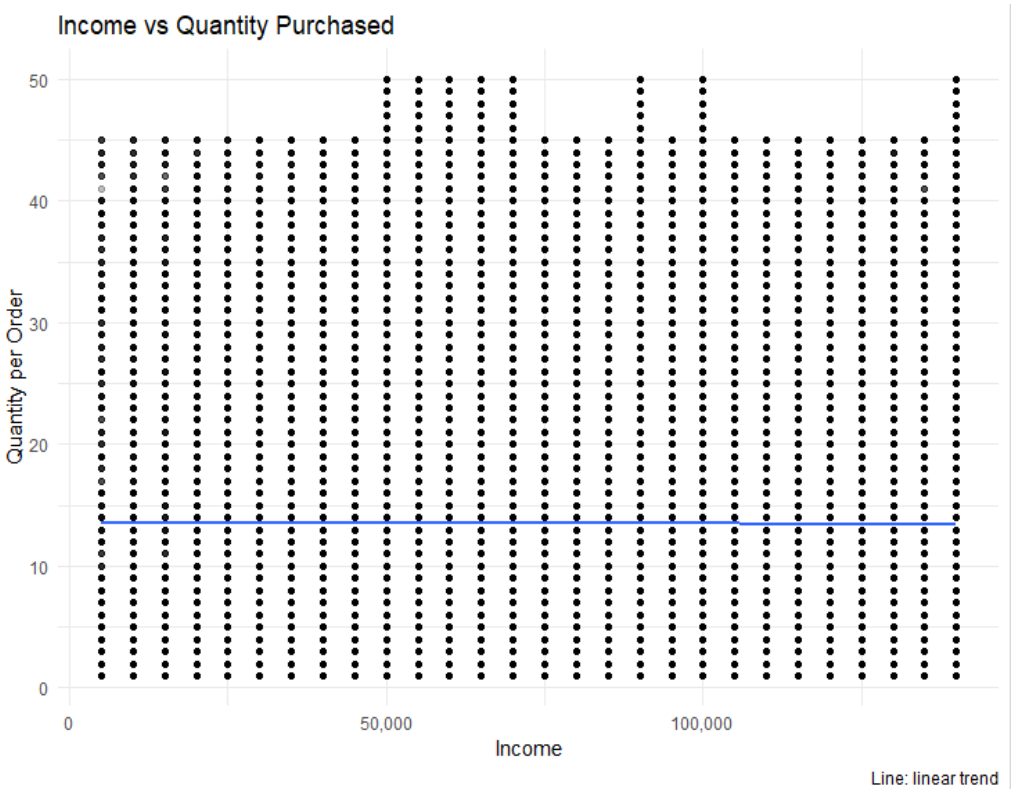

Figure 8 Income Versus Quantity Purchased

In the plot above, the quantity of products per respective order is plotted versus the income of the customer making the order, with the black dots showing the actual data entries and the blue line showing the trend of the data. As the customer income increases, the trend line shows that the quantity per order remains the same. This means that high income customers aren't, per say, more 'valuable' customers.

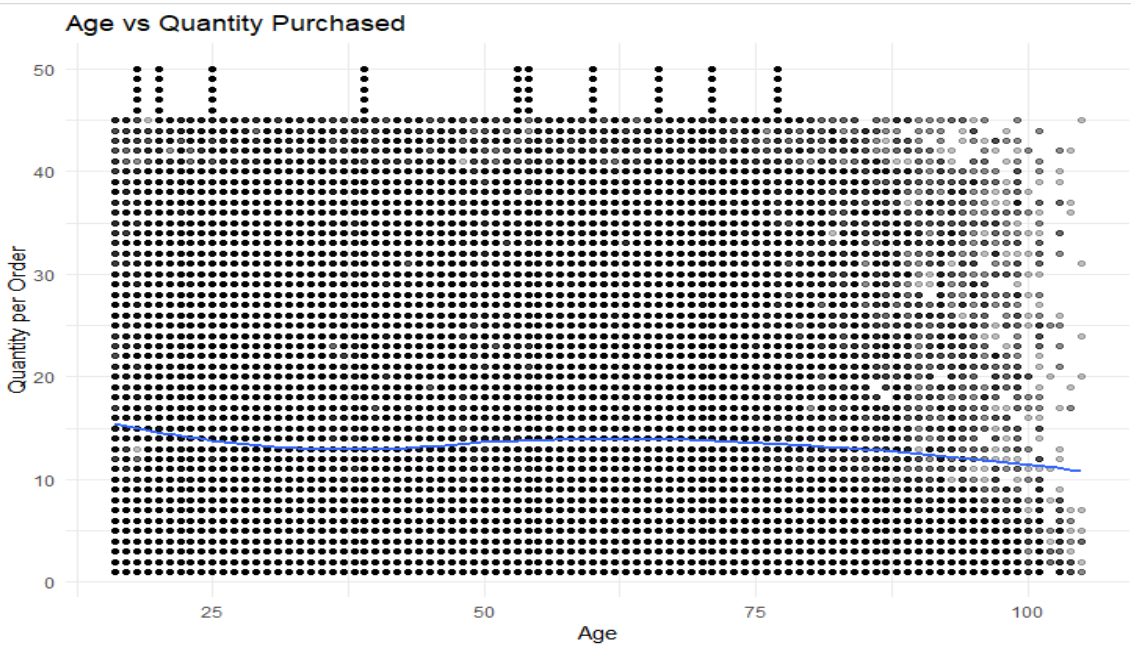

Figure 9 Age vs Quantity Purchased

In figure 9 above, we further look at customer demographics. Once again, the product quantity per order is plotted, but this time versus the age of the customer making the order. The black dots show the actual data entries and the blue line shows the trend across all entries. There is also not too much to derive from the plot above, as with the previous one, although we do see a slight increase in order quantity over the ages 45 to 75, after which it drops down again. Middle aged customers can thus be said to be the more valuable customers to the company.

Revenue and Product Insight

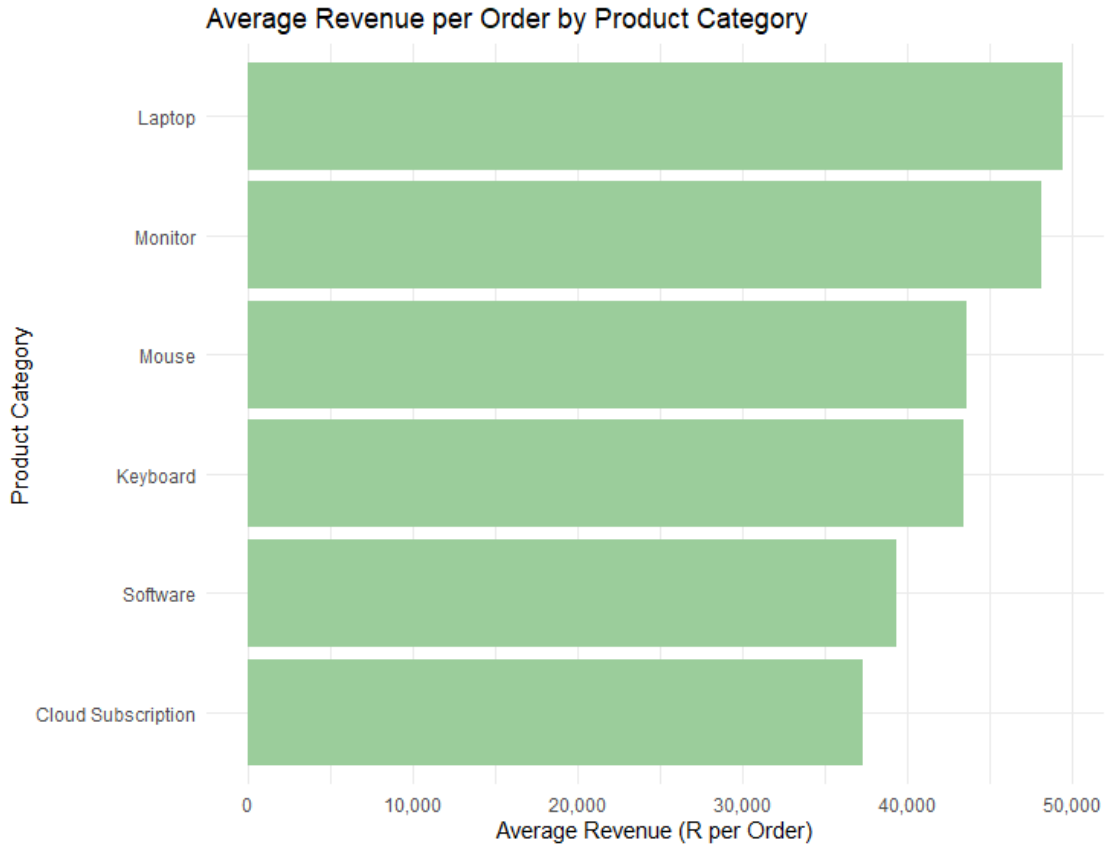

Figure 10 Average Revenue per Order by Cat.

Figure 10 above shows the average revenue per order by respective product categories. The average revenue can be calculated as the total revenue for a specific category divided by the number of orders within said category. Doing so gives us valuable insight into the revenue-generating potential for each type of product. The graph shows an expected result (based off the total revenue per category plot from figure 1), although Mouse and Keyboard swopped places, indicating that the Mouse category brings in higher average revenue than Keyboard.

In figure 11 below we plot the product markup versus the selling price of the product. This can help us identify any trends between markup percentage and selling price potential of our products. The blue dots indicate the actual data entries and the red line shows the linear trend between said entries. The trend line is practically horizontal, showing very little to no correlation between markup and selling price. We also notice that the product selling price can be expected to fall in any of three possible ranges, namely 0 – 1250, 5000 – 7500 and 15000 – 20000.

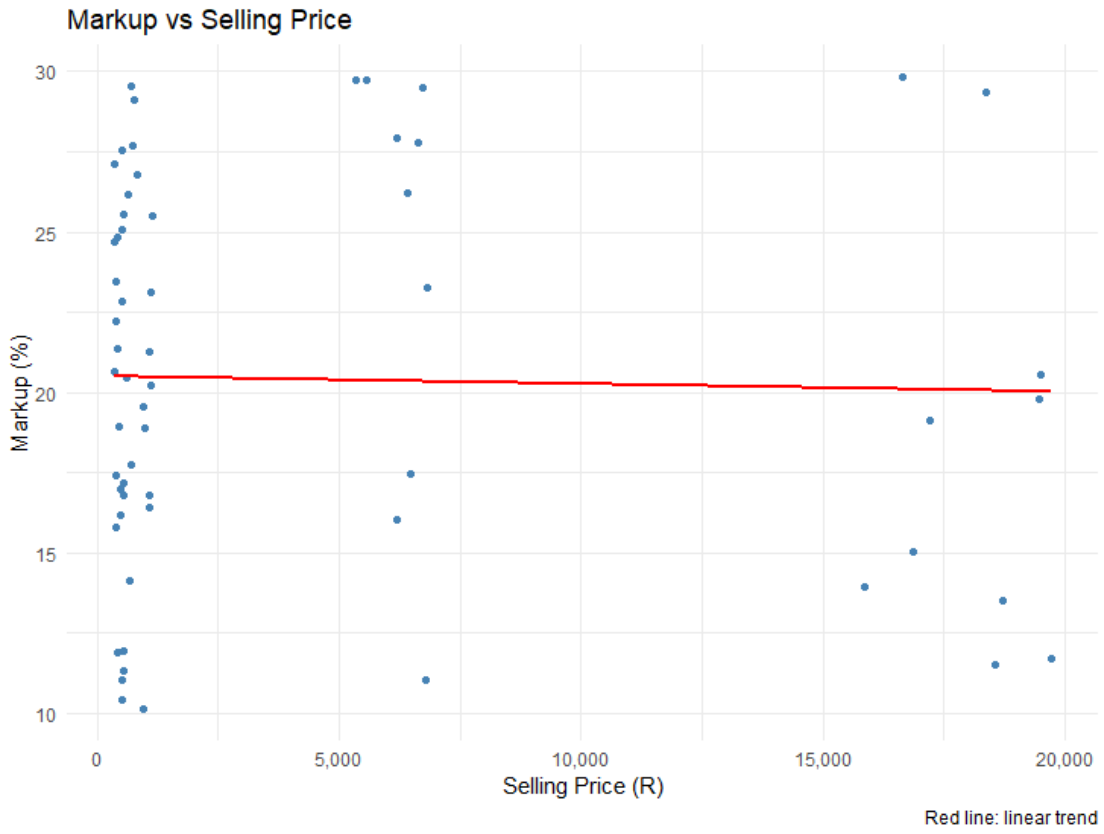

Figure 11 Markup vs Selling Price

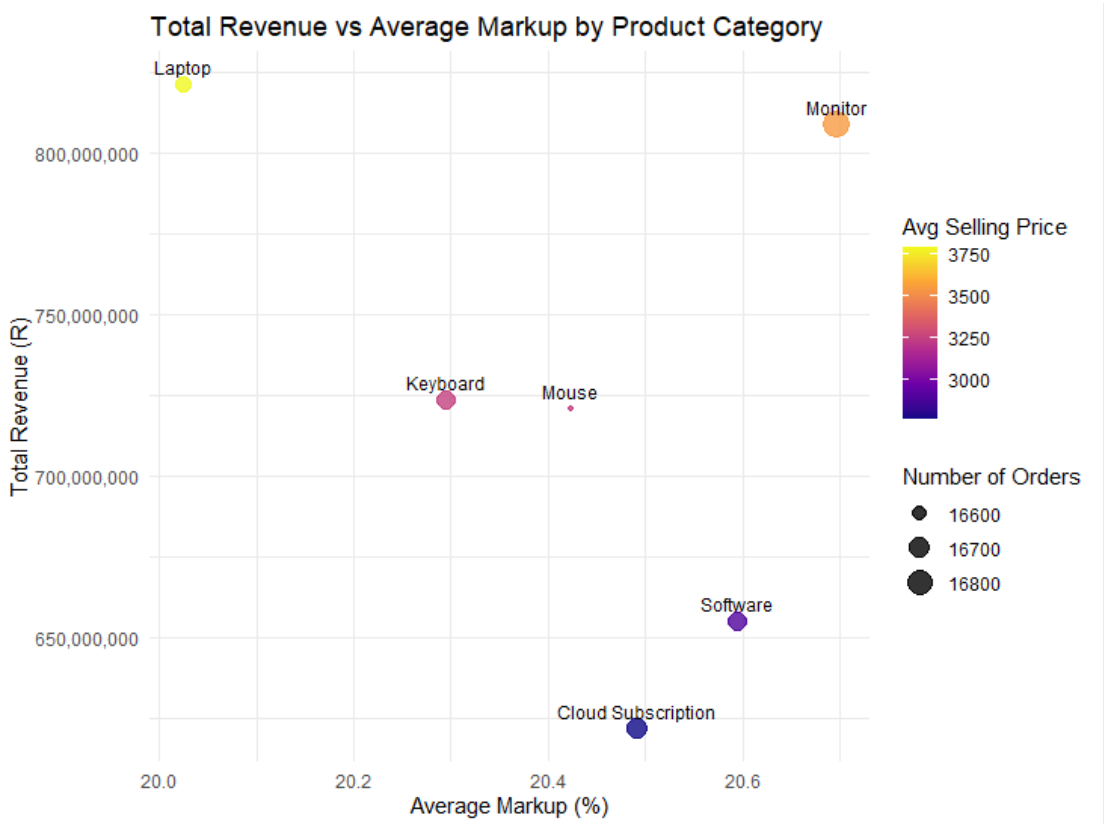

Figure 12 Total Revenue vs Average Markup by Cat.

In figure 12 above we plotted the total revenue versus average markup by product category. Laptop has the highest contribution to company revenue with the highest average selling price, despite its markup being the lowest of all categories. Mouse and Keyboard have almost the same contribution to company revenue, although Mouse has significantly less orders.

Part 3 – Statistical Process Control (SPC)

The goal of this part of the report is to perform statistical process control. SPC is a statistical method for controlling and monitoring a process or production method to ensure it remains efficient, stable and within acceptable quality limits. We have been provided with a future sales dataset (for the years 2026 and 2027), which we will use to demonstrate SPC.

It is possible to construct control charts (X-s charts) for the delivery times of every product type (Laptops – LAP, Monitors – MON, Mouse – MOU, Keyboard – KEY, Cloud subscription – CLO and Software - SOF).

3.1 – 3.2:

To set up control charts we first need to identify feasible control limits. By ordering our dataset chronologically, then splitting the data into samples of 24 entries each, we can mimic orders coming in in batches from least recent to most recent. The initial 30 samples will be comprised of the first 30x24 = 720 entries in the dataset, ordered by year, month, day and order time (per each respective product category). These 30 samples will be used to identify center-lines, outer control limits, 2-sigma control limits and 1-sigma control limits for the charts.

We use the constants A_3, B_3 and B_4 to calculate the control limits. These constants were calculated for n=24 to be:

- $A_3 = 0.35$
- $B_3 = 0.38$
- $B_4 = 1.62$

X-bar charts:

- Center line: the mean of incoming samples
- Upper Control Limit (UCL): X-bar center line + $A_3 \times$ (s-bar center line)
- Lower Control Limit (LCL): X-bar center line - $A_3 \times$ (s-bar center line)
- Upper 1-sigma limits: X-bar center line + (1/3) × (UCL – X-bar center line)
- Lower 1-sigma limits: X-bar center line - (1/3) × (UCL – X-bar center line)
- Upper 2-sigma limits: X-bar center line + (2/3) × (UCL – X-bar center line)
- Lower 2-sigma limits: X-bar center line - (2/3) × (UCL – X-bar center line)

s-charts:

- Center line: the mean standard deviation of samples
- Upper Control Limit (UCL): $B_4 \times$ (s-bar center line)
- Lower Control Limit (LCL): $B_3 \times$ (s-bar center line)
- Upper 1-sigma limits: s-bar center line + (1/3) × (UCL – s-bar center line)
- Lower 1-sigma limits: s-bar center line - (1/3) × (UCL – s-bar center line)
- Upper 2-sigma limits: s-bar center line + (2/3) × (UCL – s-bar center line)

- Lower 2-sigma limits: s-bar center line - (2/3) × (UCL – s-bar center line)

Analyzing X-s charts

Cloud Subscription (CLO)

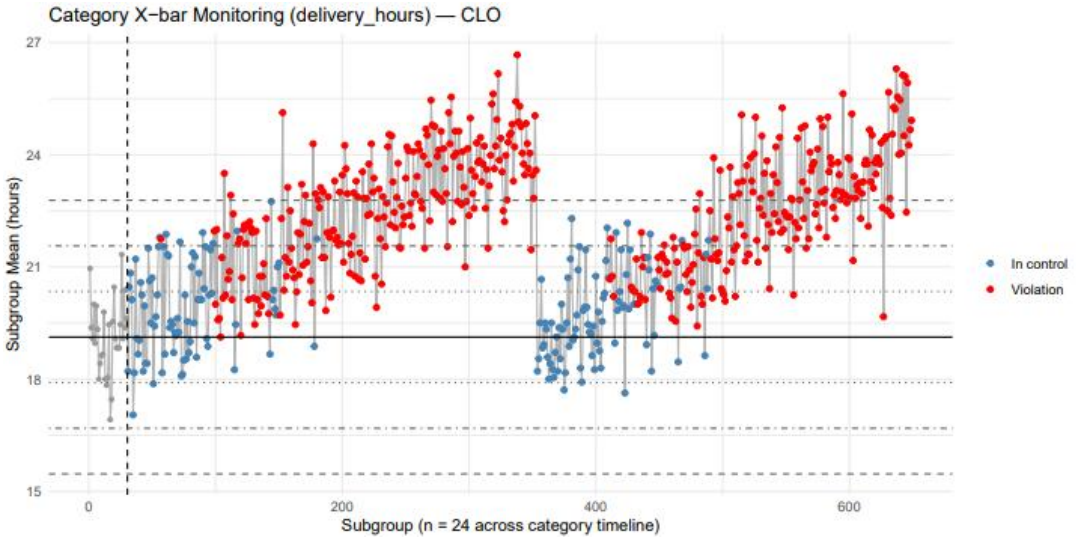

Figure 13 CLO - X-bar Chart

Figure 13 above shows the X-bar chart for Cloud Subscription. Each point represents the mean delivery time for given samples of 24 entries. The gray points to the left of the chart are the first 30 samples that were used for finding the control limits of delivery times, where the blue and red points show new incoming samples that are in and out of control respectively. Clearly, the delivery process for CLO starts out within acceptable limits, however as the year progresses the process becomes increasingly inefficient. Around sample 350 we notice a sudden reset of the mean delivery times that slowly escalate into an unstable process as the samples keep coming in. This 'reset' happens as the years change over from 2026 to 2027, which might be indicative of a couple of things. Firstly, we remember from figure (…) the order volume over time, that is high throughout the year, but low at the beginning and the end of the year. It seems like the company is capable of servicing the low order quantities, but develops a stacking backlog as the order volumes jump higher. These order delivery times that are in violation of our control limits could also provide a strong case as to why the order volume decreased from 2022 – 2023 (past dataset), as lagging deliveries lead to unsatisfied, non-returning customers.

Figure 14 below shows the s-chart for CLO. We can read from this graph the spread of Cloud Subscription delivery times. The gray points to the left of the graph are the 30 samples used to find the control limits, where the green points are the are the incoming samples as the year progresses. The standard deviation of delivery times for CLO seems in control, with a total of only 10 samples exceeding both the upper and lower 2-sigma thresholds.

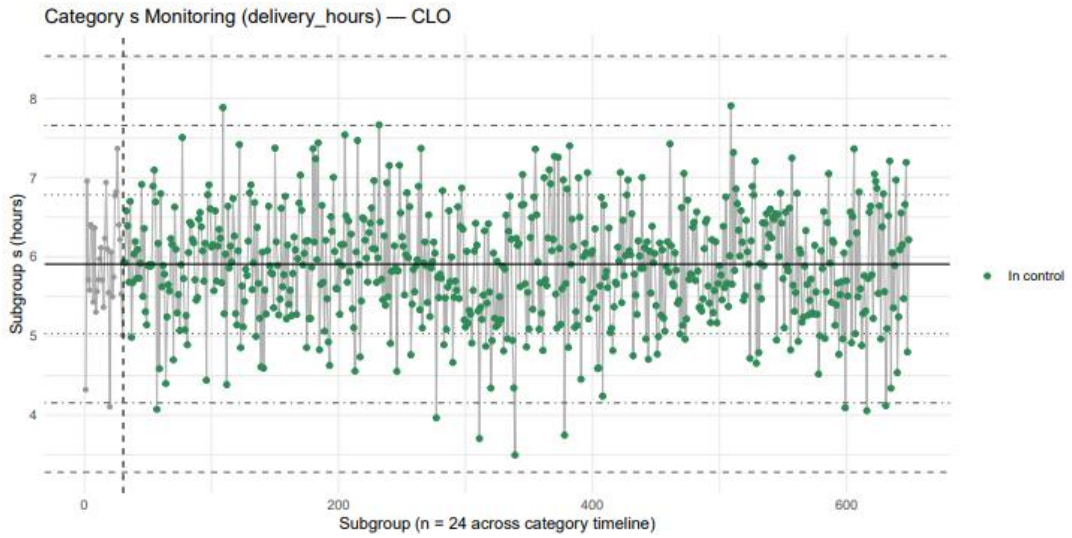

Figure 14 CLO - s-chart

Keyboards (KEY)

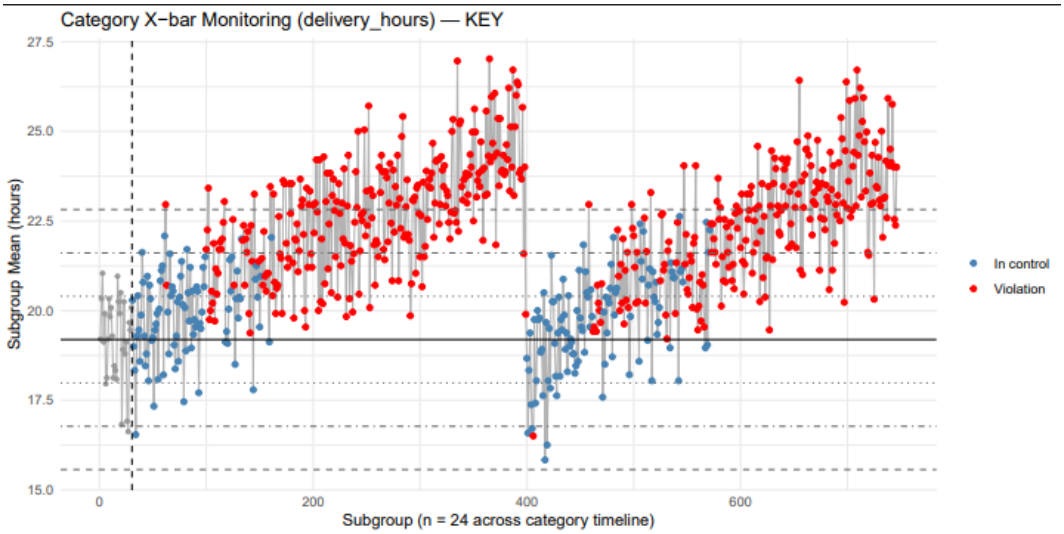

Figure 15 KEY - X-bar chart

Figure 15 above shows the X-bar chart for Keyboards. Each point represents the mean delivery time for given samples of 24 entries. The gray points to the left of the chart are the first 30 samples that were used for finding the control limits of delivery times, where the blue and red points show new incoming samples that are in and out of control respectively. Once again, the delivery process for KEY starts out within acceptable limits, however as the year progresses the process becomes increasingly inefficient. Around sample 400 we notice a sudden reset of the mean delivery times that slowly escalate into an unstable process as the samples keep coming in. This ‘reset’ happens around the changeover from 2026 to 2027, which might indicate that the company is capable of servicing the low order quantities, but develops a stacking backlog as the order volumes increase. These order delivery times that

are in violation of our control limits could also provide a strong case as to why the order volume decreased from 2022 – 2023 (past dataset), as lagging deliveries lead to unsatisfied, non-returning customers.

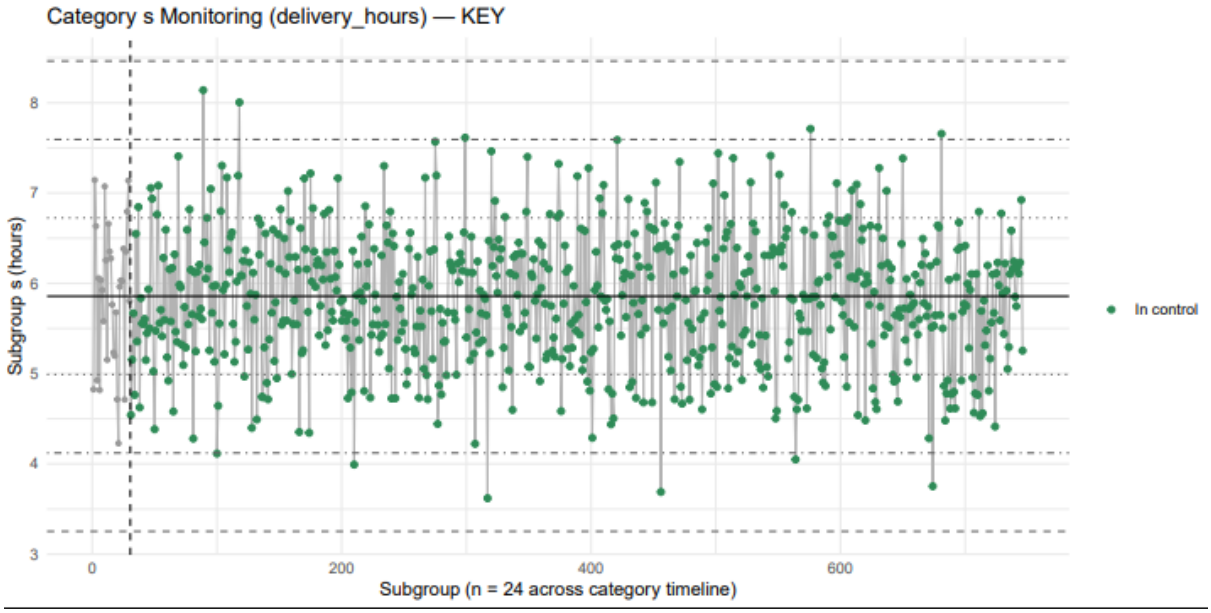

Figure 16 KEY - s-chart

Figure 16 above shows the s-chart for KEY. We can read from this graph the spread of Keyboard delivery times. The gray points to the left of the graph are the 30 samples used to find the control limits, where the green points are the are the incoming samples as the year progresses. The standard deviation of delivery times for KEY seems in control, with a total of only 10 samples exceeding both the upper and lower 2-sigma thresholds.

Laptop (LAP)

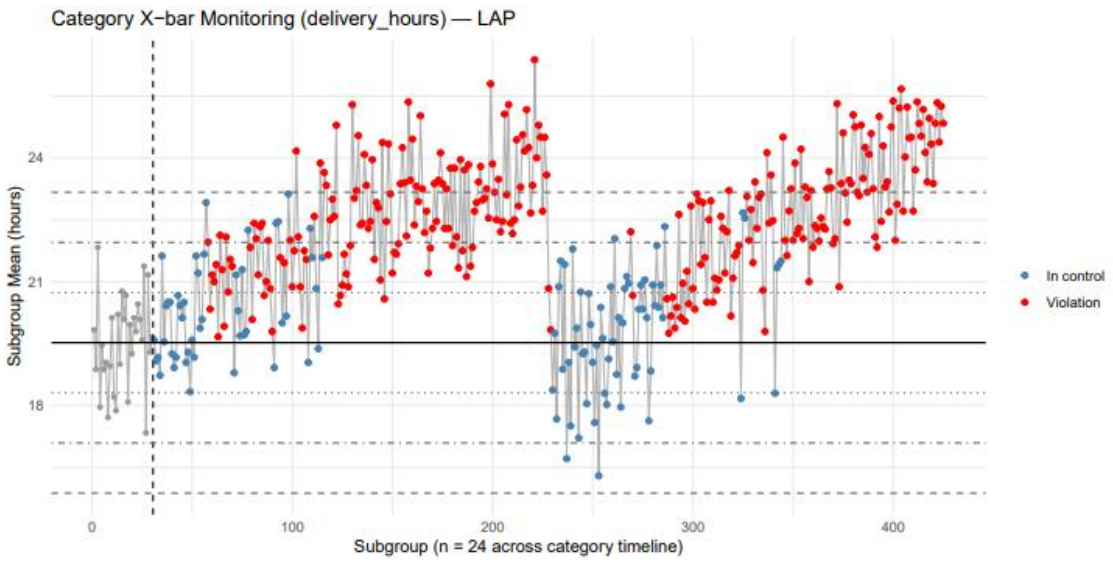

Figure 17 LAP - X-bar chart

Figure 17 above shows the X-bar chart for Laptops. Each point represents the mean delivery time for given samples of 24 entries. The gray points to the left of the chart are the first 30 samples that were used for finding the control limits of delivery times, where the blue and red points show new incoming samples that are in and out of control respectively. Once again, the delivery process for LAP starts out within acceptable limits, however as the year progresses the process becomes increasingly unstable. Around sample 250 we notice a sudden reset of the mean delivery times that slowly escalate into an unstable process as the samples keep coming in. This 'reset' happens around the changeover from 2026 to 2027, which might indicate that the company is capable of servicing the low order quantities, but develops a stacking backlog as the order volumes increase. These order delivery times that are in violation of our control limits could also provide a strong case as to why the order volume decreased from 2022 – 2023 (past dataset), as lagging deliveries lead to unsatisfied, non-returning customers.

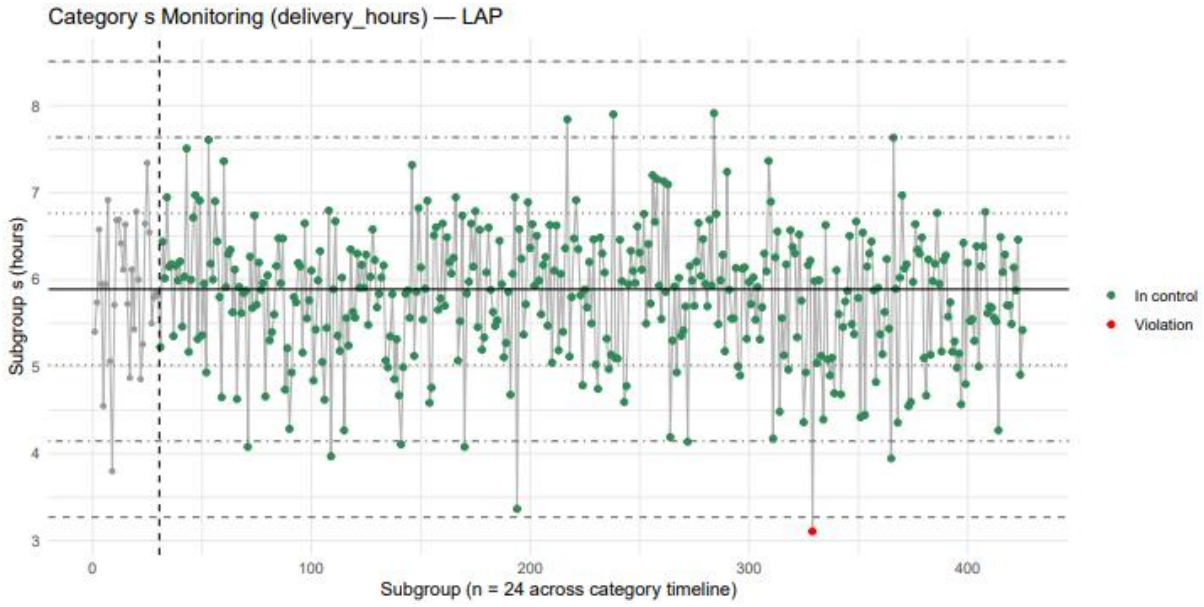

Figure 18 LAP - s-chart

Figure 18 above shows the s-chart for LAP. We can read from this graph the spread of Laptop delivery times. The gray points to the left of the graph are the 30 samples used to find the control limits, where the green points are the are the incoming samples as the year progresses. The standard deviation of delivery times for LAP seems in control in general, with a total of only 9 samples exceeding both the upper and lower 2-sigma thresholds. We do find on this s-chart our first sample that is in violation of a control limit (LCL). This indicates our first encounter of what is know as special cause variation, meaning that the process is not in a state of statistical control. Further investigation is advised to identify and prevent such violations in the future.

Monitor (MON)

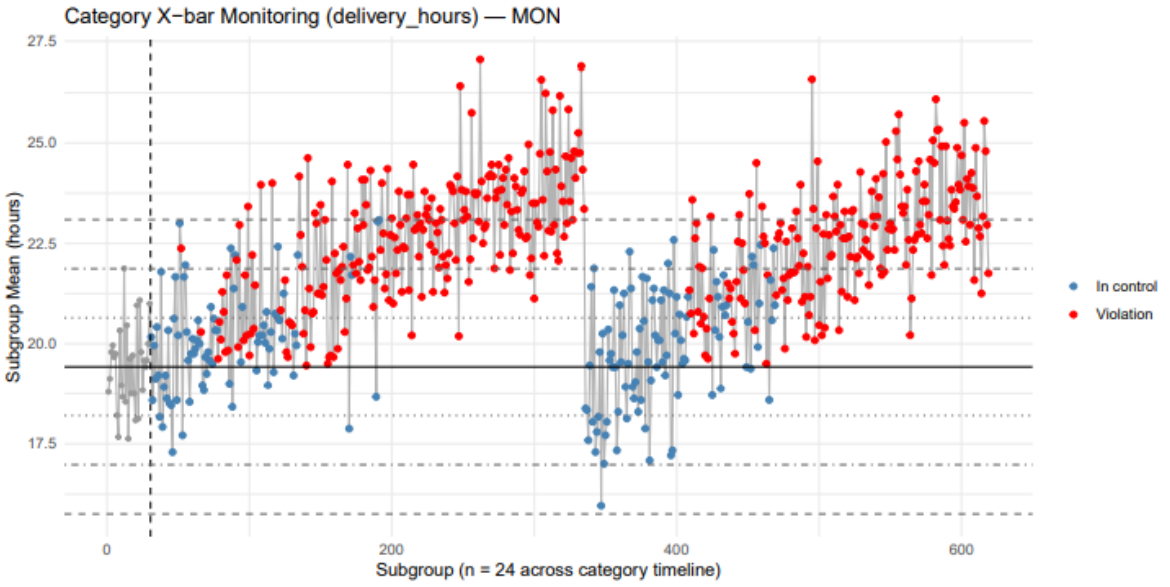

Figure 19 MON - X-bar chart

Figure 19 above shows the X-bar chart for Monitor. Each point represents the mean delivery time for given samples of 24 entries. The gray points to the left of the chart are the first 30 samples that were used for finding the control limits of delivery times, where the blue and red points show new incoming samples that are in and out of control respectively. Once again, the delivery process for MON starts out within acceptable limits, however as the year progresses the process becomes increasingly inefficient. Around sample 350 we notice a sudden reset of the mean delivery times that slowly escalate into an unstable process as the samples keep coming in. This 'reset' happens around the changeover from 2026 to 2027, which might indicate that the company is capable of servicing the low order quantities, but develops a stacking backlog as the order volumes increase. These order delivery times that are in violation of our control limits could also provide a strong case as to why the order volume decreased from 2022 – 2023 (past dataset), as lagging deliveries lead to unsatisfied, non-returning customers.

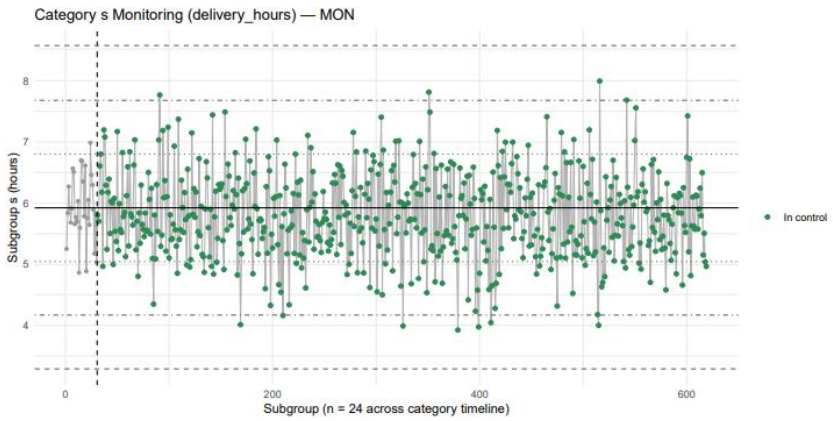

Figure 20 MON - s-chart

Figure 20 above shows the s-chart for MON. We can read from this graph the spread of Monitor delivery times. The gray points to the left of the graph are the 30 samples used to find the control limits, where the green points are the are the incoming samples as the year progresses. The standard deviation of delivery times for MON seems in control in general, with a total of only 9 samples exceeding both the upper and lower 2-sigma thresholds.

Mouse (MOU)

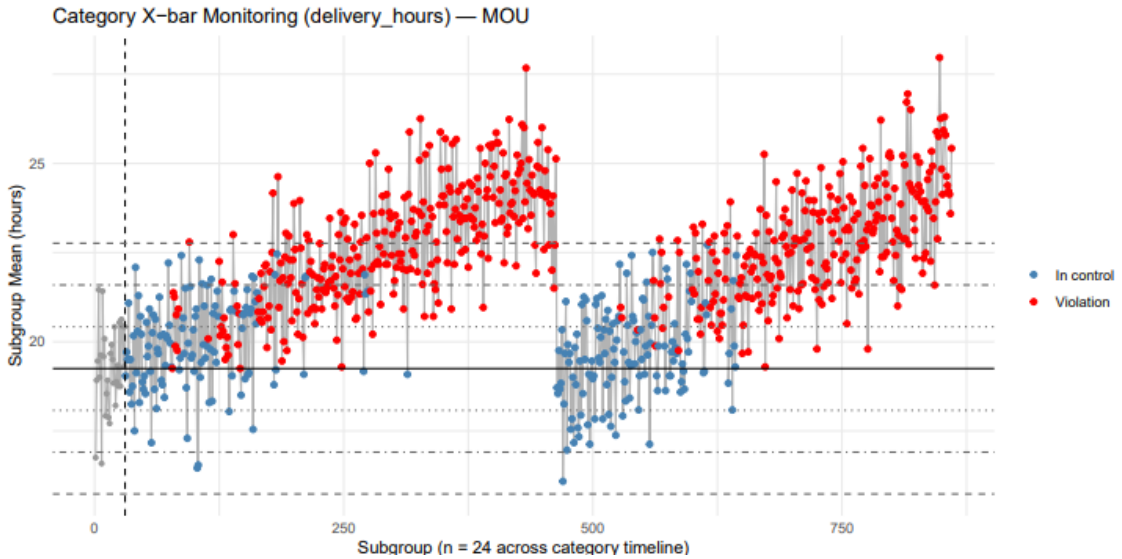

Figure 21 MOU - X-bar chart

Figure 21 above shows the X-bar chart for Mouse. Each point represents the mean delivery time for given samples of 24 entries. The gray points to the left of the chart are the first 30 samples that were used for finding the control limits of delivery times, where the blue and red points show new incoming samples that are in and out of control respectively. Once again, the delivery process for MOU starts out within acceptable limits, however as the year progresses the process becomes increasingly inefficient. Around sample 475 we notice a sudden reset of the mean delivery times that slowly escalate into an unstable process as the samples keep coming in. This ‘reset’ happens around the changeover from 2026 to 2027, which might indicate that the company is capable of servicing the low order quantities, but develops a stacking backlog as the order volumes increase. These order delivery times that are in violation of our control limits could also provide a strong case as to why the order volume decreased from 2022 – 2023 (past dataset), as lagging deliveries lead to unsatisfied, non-returning customers.

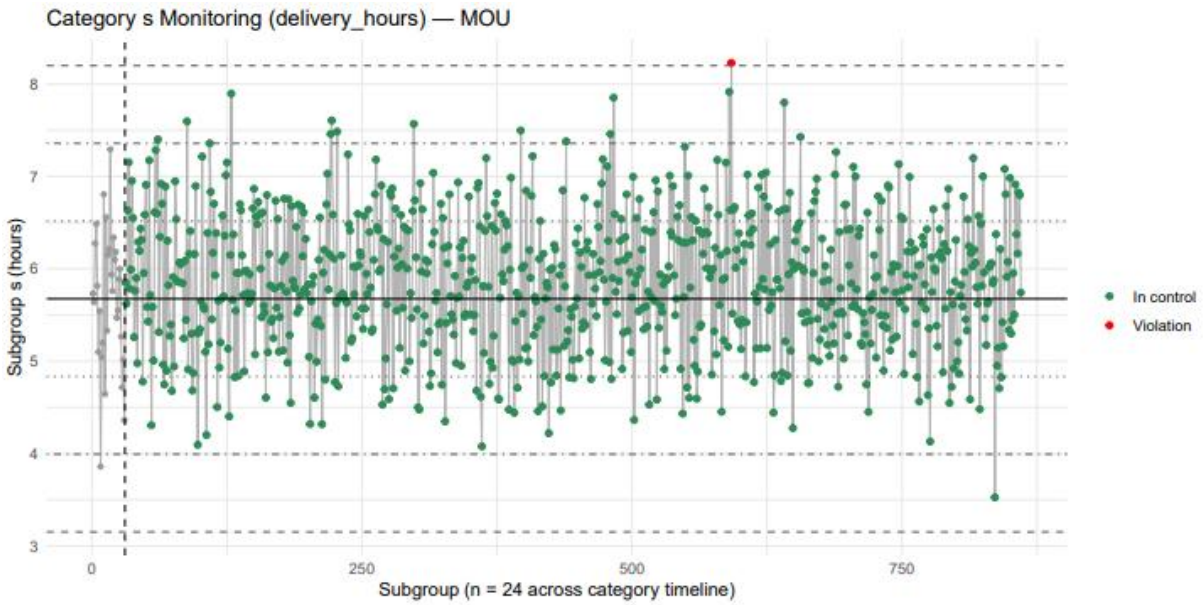

Figure 22 MOU - s-chart

Figure 22 above shows the s-chart for MOU. We can read from this graph the spread of Mouse delivery times. The gray points to the left of the graph are the 30 samples used to find the control limits, where the green points are the are the incoming samples as the year progresses. The standard deviation of delivery times for MOU seems in control in general, with a total of only 14 samples exceeding both the upper and lower 2-sigma thresholds. We do find on this s-chart our second sample that is in violation of a control limit (LCL). This indicates another encounter of what is known as special cause variation, meaning that the process is not in a state of statistical control. Further investigation is advised to identify and prevent such violations in the future.

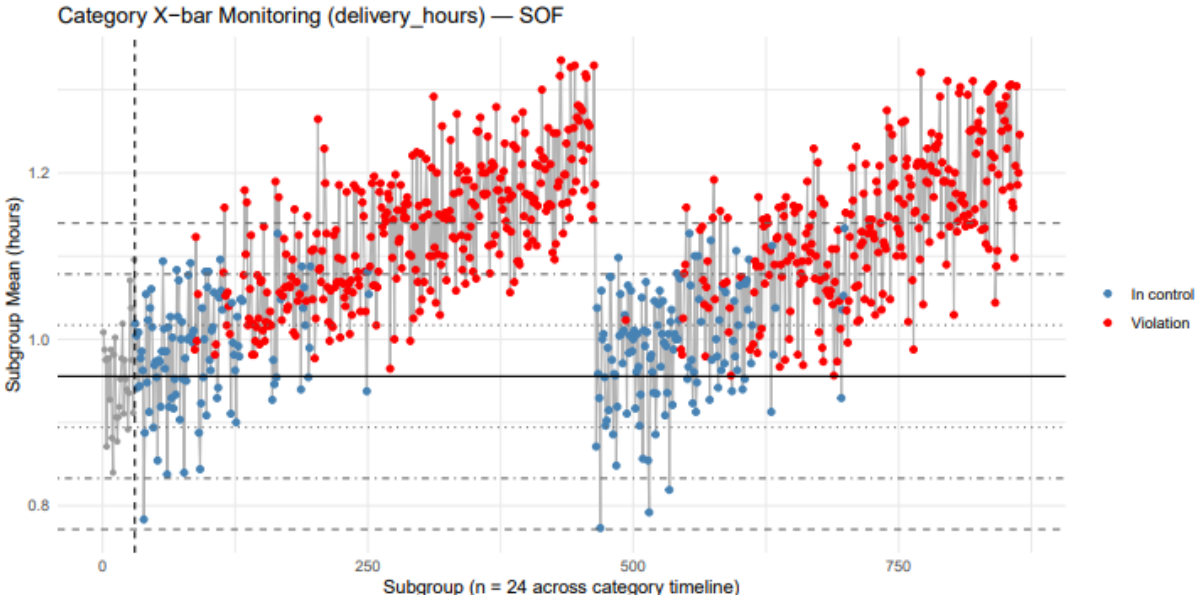

Figure 23 SOF - X-bar chart

Figure 23 above shows the X-bar chart for Software. Each point represents the mean delivery time for given samples of 24 entries. The gray points to the left of the chart are the first 30 samples that were used for finding the control limits of delivery times, where the blue and red points show new incoming samples that are in and out of control respectively. Once again, the delivery process for SOF starts out within acceptable limits, however as the year progresses the process becomes increasingly inefficient. Around sample 475 we notice a sudden reset of the mean delivery times that slowly escalate into an unstable process as the samples keep coming in. This 'reset' happens around the changeover from 2026 to 2027, which might indicate that the company is capable of servicing the low order quantities, but develops a stacking backlog as the order volumes increase. These order delivery times that are in violation of our control limits could also provide a strong case as to why the order volume decreased from 2022 – 2023 (past dataset), as lagging deliveries lead to unsatisfied, non-returning customers.

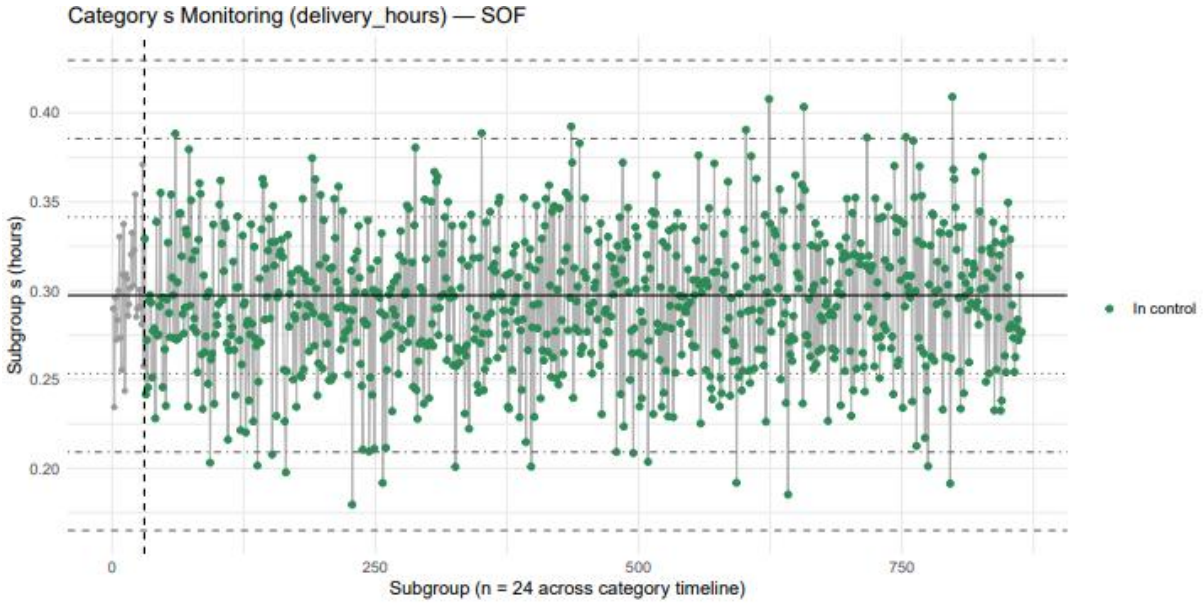

Figure 24 SOF - s-chart

Figure 24 above shows the s-chart for SOF. We can read from this graph the spread of Software delivery times. The gray points to the left of the graph are the 30 samples used to find the control limits, where the green points are the are the incoming samples as the year progresses. The standard deviation of delivery times for SOF seems in control in general, with a total of only 19 samples exceeding both the upper and lower 2-sigma thresholds. This is the highest 2-sigma control line violation we've gotten on any of our charts and it is advisable that, in case the company wants to prevent future violations, the reliability of this process can be improved.

Process Capability (3.3)

Process capability can be seen as the ability of a process to meet specifications. Capability potential (Cp) shows us how well the process spread fits into the specification range. Capability performance (Cpk) does the same as Cp, but additionally measures how close the process mean is to the target value of said specification.

The following formulas are used to calculate the process capability:

$C_p = (USL - LSL)/6\times sigma$

$C_{pu} = (USL - \mu)/3\times sigma$

$C_{pl} = (\mu - LSL)/3\times sigma$

$C_{pk} = \min(C_{pl}, C_{pu})$

Using the first 1000 deliveries (per product type) we calculate the process capability indices, export to a .csv file and find the following:

product_id	n	mean_delivery	sd_delivery	Cp	Cpu	Cpl	Cpk	capable
SOF008	1000	1.075675	0.292453121	18.23654098	35.24704501	1.226036951	1.226036951	Capable
SOF003	1000	1.069425	0.295476278	18.04995439	34.89346675	1.20644203	1.20644203	Capable
SOF010	1000	1.069425	0.29646765	17.98959628	34.77678482	1.20240775	1.20240775	Capable
SOF007	1000	1.08575	0.304488688	17.51570272	33.8428008	1.188604639	1.188604639	Capable
SOF009	1000	1.085725	0.305005701	17.48601193	33.78546135	1.186562519	1.186562519	Capable
SOF004	1000	1.069825	0.304294396	17.52688647	33.88185411	1.171918833	1.171918833	Capable
SOF006	1000	1.0594	0.301903222	17.66570524	34.16171997	1.169690508	1.169690508	Capable
SOF005	1000	1.078225	0.308372207	17.29511681	33.42473191	1.165501708	1.165501708	Capable
SOF002	1000	1.064625	0.30822881	17.30316297	33.4549897	1.151336242	1.151336242	Capable
SOF001	1000	1.069425	0.310050489	17.20149951	33.25326692	1.149732101	1.149732101	Capable
CLO020	1000	20.895046	5.957779299	0.895188134	0.621313941	1.169062327	0.621313941	Not capable
MON032	1000	21.143636	5.987536703	0.890739147	0.604386775	1.177091518	0.604386775	Not capable
MON039	1000	21.078656	6.074802419	0.877943506	0.59927019	1.156616822	0.59927019	Not capable
MON037	1000	21.465962	5.920142964	0.900879145	0.593118447	1.208639844	0.593118447	Not capable
LAP022	1000	21.708244	5.81366668	0.917378589	0.590089787	1.24466739	0.590089787	Not capable
MOU055	1000	21.40194	6.008164308	0.887681005	0.587981035	1.187380976	0.587981035	Not capable
MOU058	1000	21.384764	6.03097589	0.884323438	0.586706375	1.181940501	0.586706375	Not capable
MOU057	1000	21.404176	6.032999094	0.884026875	0.585437073	1.182616676	0.585437073	Not capable
CLO014	1000	21.335194	6.077119833	0.877608716	0.584970419	1.170247013	0.584970419	Not capable

Figure 25 Snippet of process capability .csv

The table above is a snippet of the full range of process capabilities (for all products). The plot in figure 26 below shows a graphed summary of our findings:

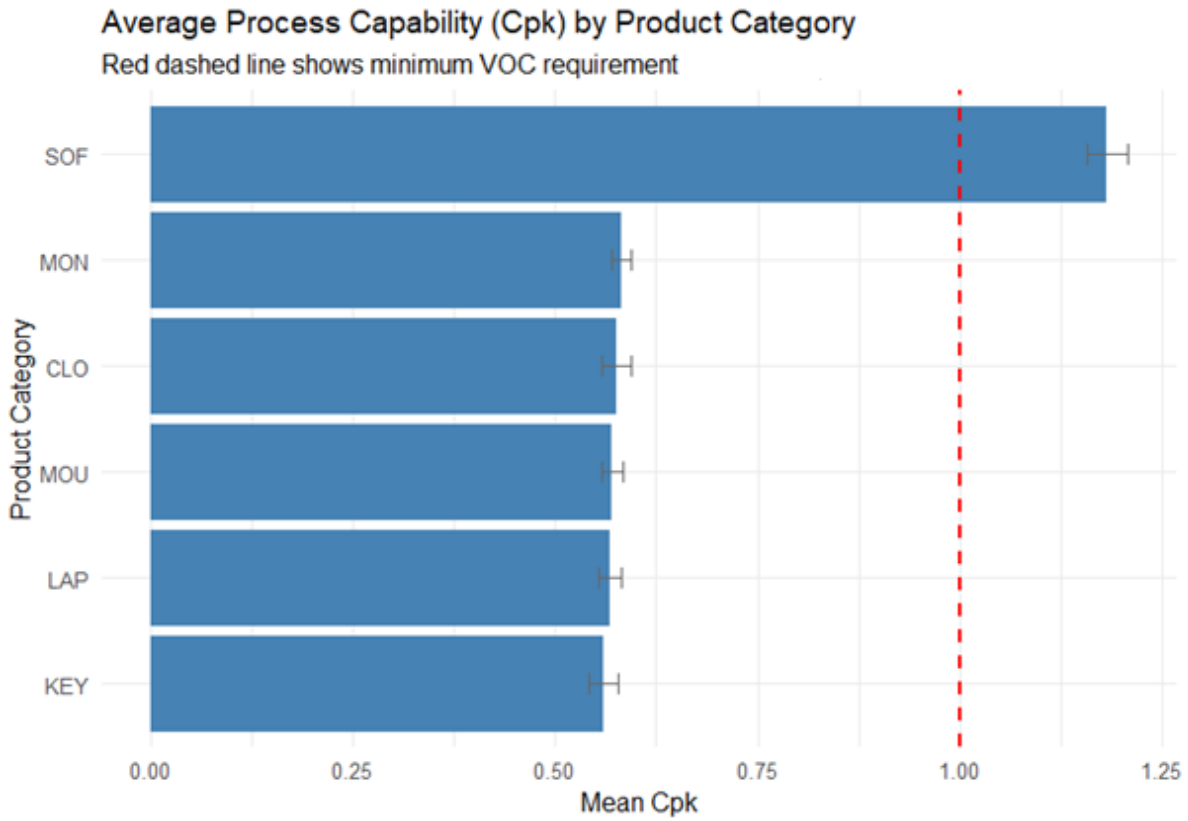

Figure 26 Average Process capability by Category

Typically, a process is considered barely capable with a C_{pk} >= 1.0 and fully capable with a C_{pk} >= 1.3. None of our categories are fully capable per the above definition, with only software being marginally capable. The rest of the product categories are not capable, meaning that a substantial portion of its output can be expected to fall outside of specification limits.

Process Control Issues (3.4)

In this subsection, the objective is to identify samples that show process control issues in accordance with a given ruleset. The ruleset is as follows:

A) 1 s sample outside of the upper +3 sigma-control limits for all product types (if many, list only the first 3 and last 3 and total number identified).
B) Find the most consecutive samples of s between the -1 and +1 sigma-control limits for all product types. This signifies good control.
C) 4 consecutive X-bar samples outside of the upper, second control limits for all product types (if many, list only the first 3 and last 3 and total number identified).

Rule A

This rule is simply a check for process control. If the rule is violated, we know that at that instance, for that category, the process was out of control. We only found one instance from the MOU category to break this rule (sample 592). This should get a manager's attention as

out-of-control processes can have severe downstream impact if unattended. The rest of the categories showed no sign of control loss per this rule.

Rule B

Identifying lasting periods of proper control is done by this rule. It indicates process consistency and low/negligible variability.

CLO had the best control according to this rule, with a staggering 35 consecutive samples (474-508) within 1-sigma control lines. A close second place was taken by MON with 34 consecutive samples, followed by SOF, LAP, MOU and lastly KEY.

Rule C

If samples start qualifying for rule C, it might be the case that there is a shift in process mean.

KEY and SOF tied first place with a total of 25 runs where 4 (or more) X-bar samples exceeded the 2-sigma control limits. Because they show multiple shifts in process mean, we might have to look for systematic bias within our process control. MON, MOU, CLO and LAP also had several runs where the rule was qualified for. In general, the process does not seem to be in control and requires attention. Investigators can be on the lookout for calibration issues, staff/workforce problems or simply capacity requirements as a result of product demand (as per an earlier suggestion).

Part 4 – Risk and Data Correction

4.1 Type I error – Manufacturer's error

Making a type I error would mean that we conclude that a process is out of control even though it is not. Our hypothesis is as follows:

H_0: The process is in control and centered around the center line.

H_1: The process is not in control.

As per the guidance of Hint 2 in the project brief, we use an imagined scenario to answer the question. If we find seven (or more) samples above the center line of the SPC graph, we investigate to identify possible causes. The probability that one sample is above the center line is:

$$P(\text{sample} > \text{center line}) = 0.5$$

The probability, then, of finding 7 consecutive samples above the center line is:

$$0.5^7 = 0.0078$$

This result means that the likelihood of us flagging an error in the process when it is in actual fact functioning as expected is 0.78%. This is very low odds of us making a type I error.

4.2 Type II error – Consumer's error

A type II error is made when you flag the process as in control, when in actual fact it is not. We were provided a scenario in this question with the following values given:

- Process Mean: 25.05l

- UCL: 25.058l
- LCL: 25.011l

‘Unknown’ given values:

- Process mean moved to: 25.028l
- Standard deviation is 0.017 instead of 0.013l

In this situation, we can calculate the chances of, after the mean shifted, a sample mean still falling within the old control limits. First, we convert the control limits to z-scores under the shifted distribution:

$$Z_{lcl} = (25.011 – 25.028)/0.017 = -1.0$$

$$Z_{ucl} = (25.089 – 25.028)/0.017 = 3.58$$

Now we calculate the probability that a random sample mean will lie between these limits:

$$P(Z<3.59) – P(Z<-1.0) = 0.8411$$

This means that the probability of us making a type II error is 84.11%, meaning that there is a high risk of us failing to detect that the process mean has moved.

4.3

There were differences between the standard product dataset and the head office product dataset. In this section we fix the dataset in accordance with an email received from the head office and redo the data analysis from part 1.

After fixing the head office product dataset, remerging the corrected set with the other sets (as in part 1) and plotting the same graphs as we had plotted before, we found that none of the graphs were affected by the correction. This result is, in a sense, expected since all plots were created from aggregated category-level metrics, meaning that small corrections within respective categories would not alter the summary statistics.

Part 5 - Optimizing for Maximum Profit

In this section, we were given two datasets called timeToServe.csv and timeToServe2.csv. These files contain data from two different coffee shops and lists the number of baristas working and their respective service times. Our task is to optimize the profit for the two given datasets. We were given further information on the topic:

- There are problems if there are less than 2 people on duty
- R30 profit is made per served customer
- It costs R1000 per day, per person, to appoint more personnel (max 6 people).

We can then calculate the total profit as:

$$\text{Total profit} = (\text{customers served} \times 30) - (\text{baristas} \times 1000)$$

If we then plot the amount of baristas versus the estimated profit, we find the following graphs:

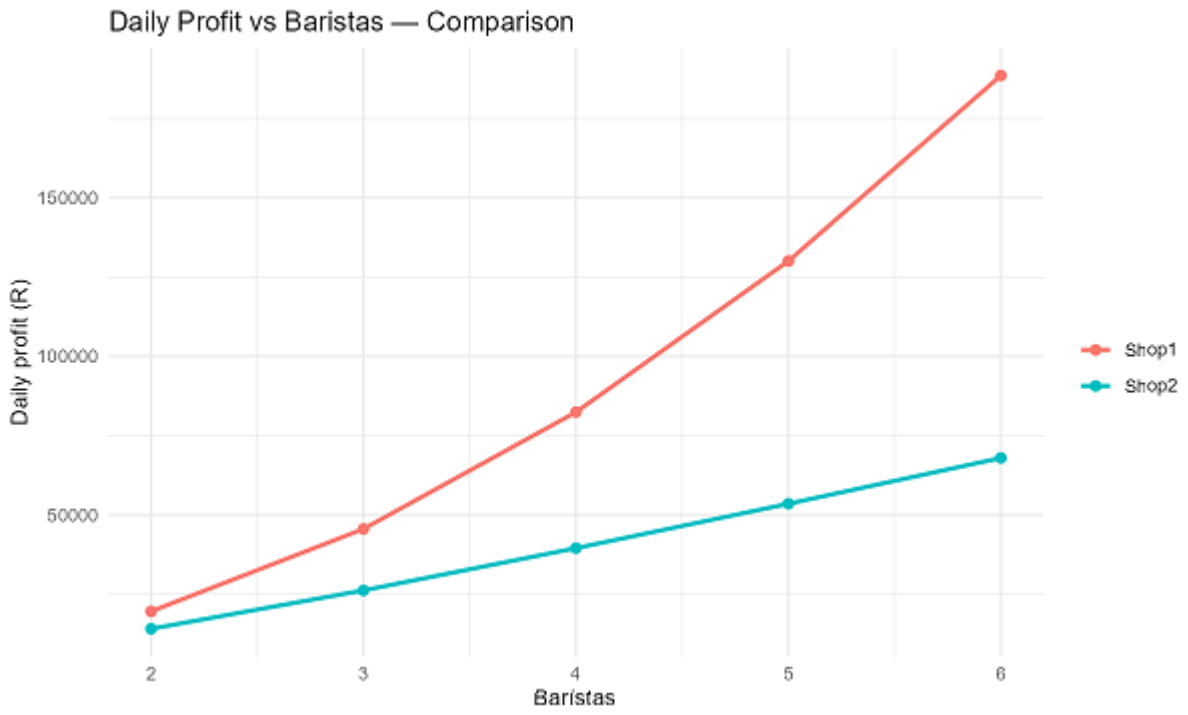

Figure 27 Profit vs Baristas (shop 1 and 2)

Figure 27 above shows that shop 1 seems to be more profitable for the same number of baristas employed. We also note that for both shops, as we increase the amount of employed baristas, our profit increases.

We can measure service reliability as whether customers were served on time. In our case, we chose 2 minutes (120s) as the service threshold for reliability.

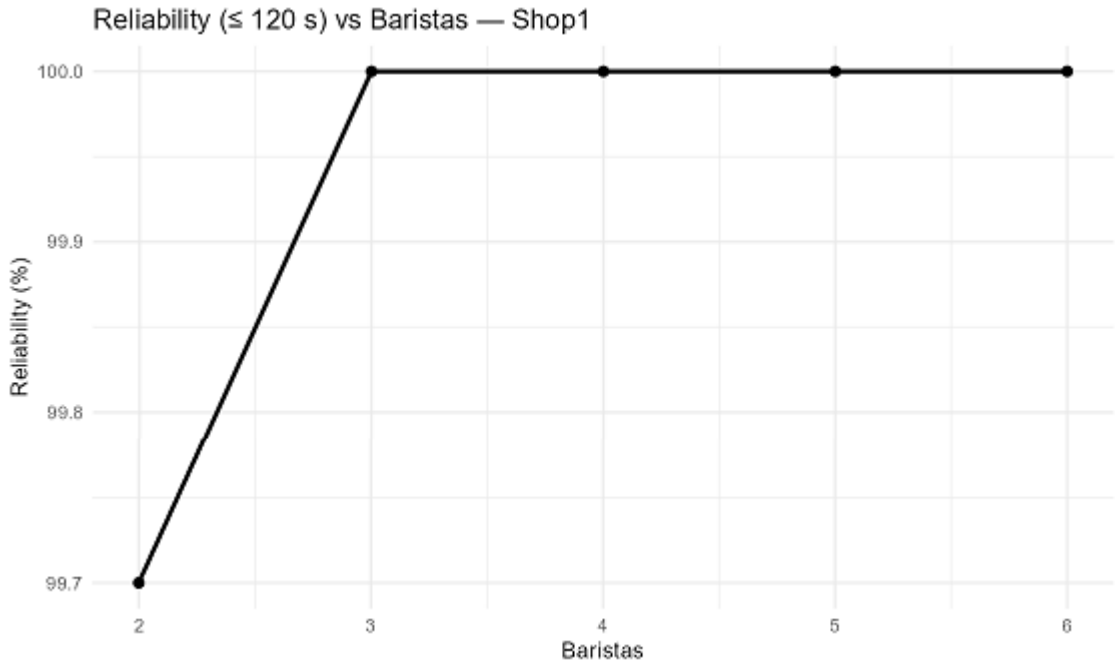

Figure 28 Reliability vs Baristas (shop 1)

Figure 28 above shows that our service reliability peaks at 100% once there are 3 baristas (or more) working in shop 1. At 2 baristas, the reliability is still high (99.7%), meaning that we can still employ 2 people and have reliable service. If costs are an issue, this might be good to keep in mind.

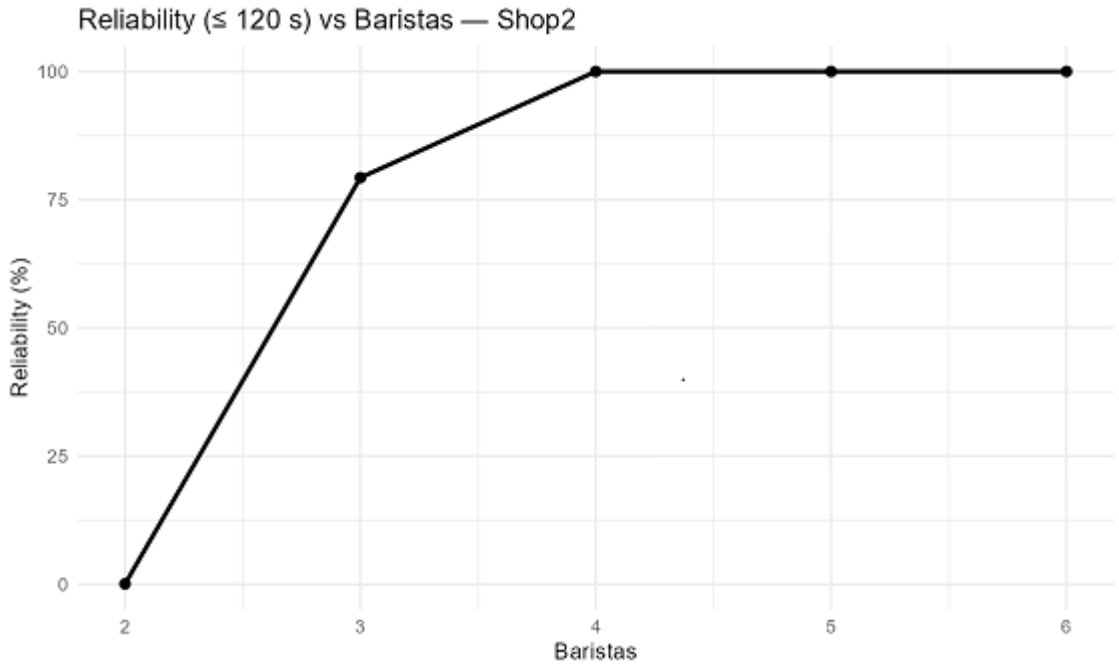

Figure 29 Reliability vs Baristas (shop 2)

Figure 29 above shows the same reliability versus number of baristas plot, this time for shop 2. Notice that the reliability is practically unacceptable for any number of baristas below 4.

Summary

We can thus conclude that employing 4-5 baristas at shop 1 and 4 baristas at shop 2 maximizes profit. Shop 1 also is more flexible in terms of number of staff working, while shop 2 requires at least 4 baristas.

Part 6 – Analysis of Variance (ANOVA)

This section of the report requires us to revisit Part 3. Based off those results, we are to decide what data we would like to use to prove our own hypothesis. Following are three hypothesis that we have tested:

Hypothesis 1:

Is there a significant difference between the delivery hours over the years across all product categories?

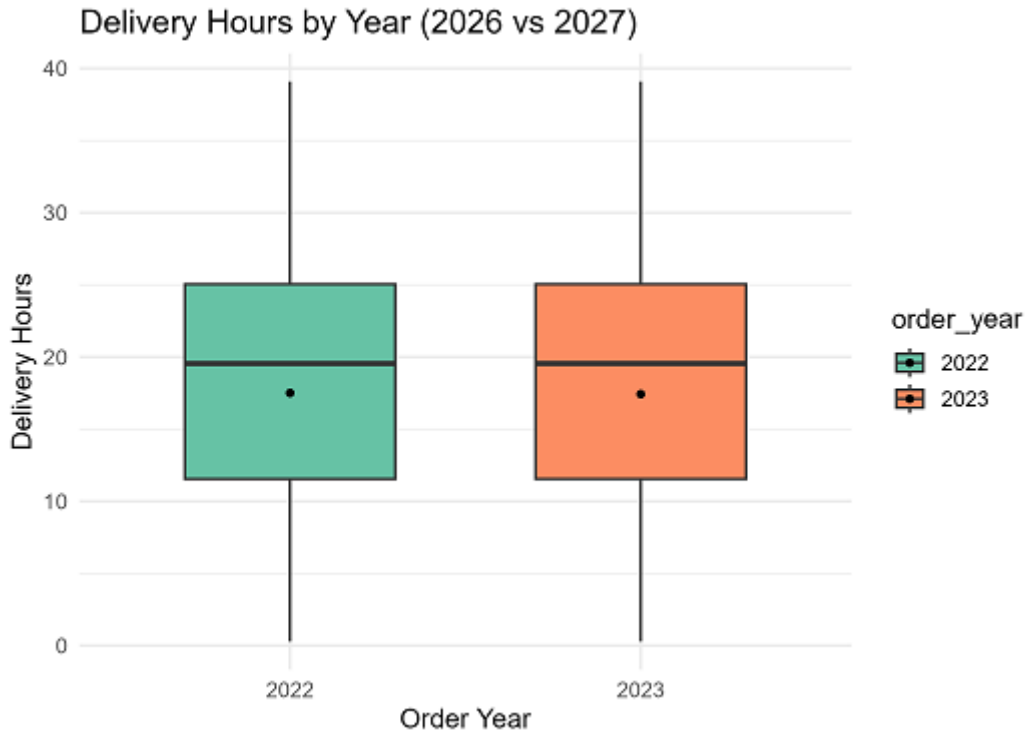

Figure 30 Delivery Hours over 2022-2023

Figure 30 above shows that there is no notable difference between the delivery hours of 2022 and 2023. In fact, the spread looks identical with minimum, 1st quartile, median, mean, 3rd quartile and maximum values being the same.

Hypothesis 2:

Is there a significant difference between the delivery hours over the years for the respective product categories?

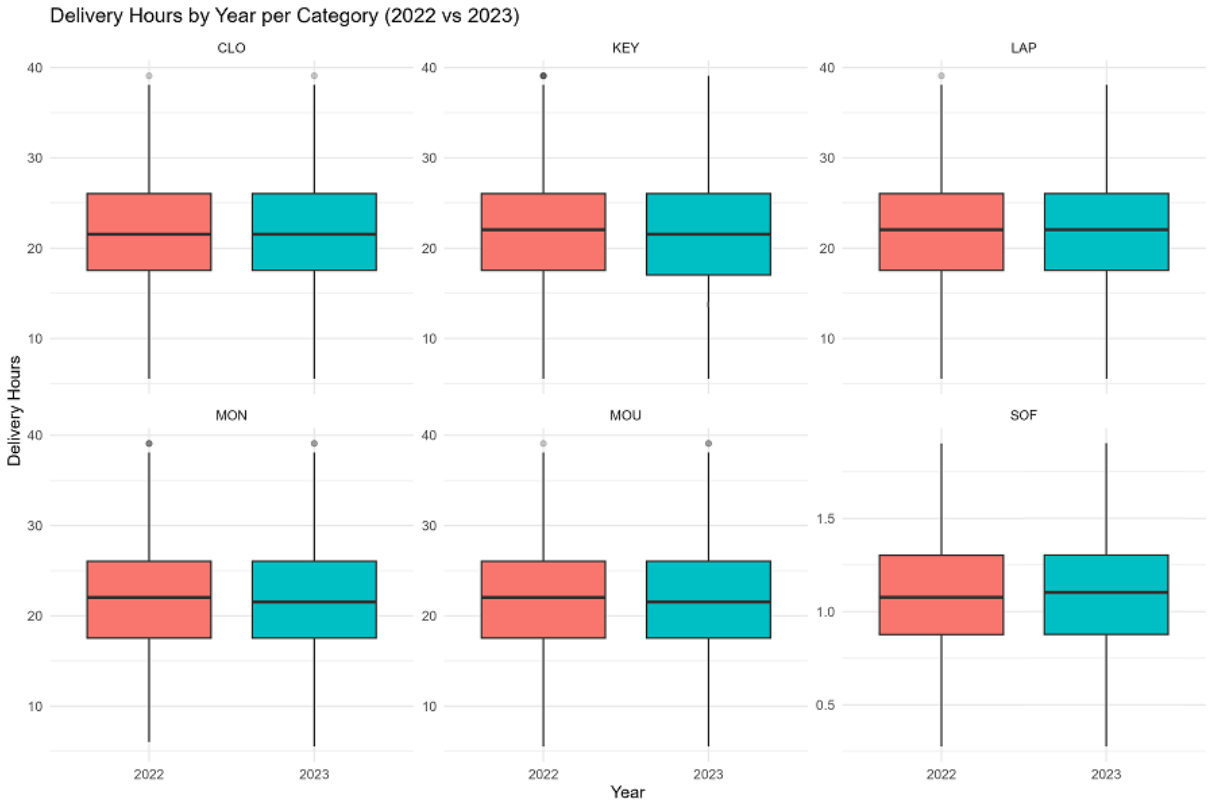

Figure 31 Delivery Hours over years per category

Figure 31 above shows what we might have suspected based of the previous plot: there is no visible difference between delivery hours from one year and the next, even for specific categories.

Hypothesis 3:

Is there a difference between delivery hours as we go through the months of a given year?

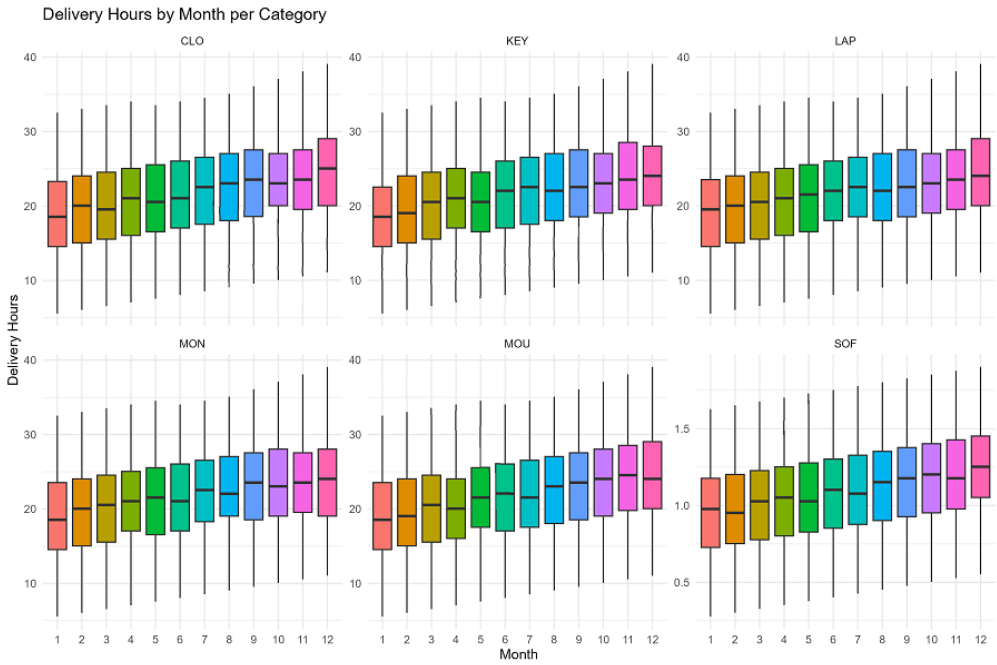

Figure 32 Delivery Hours by Month per Category

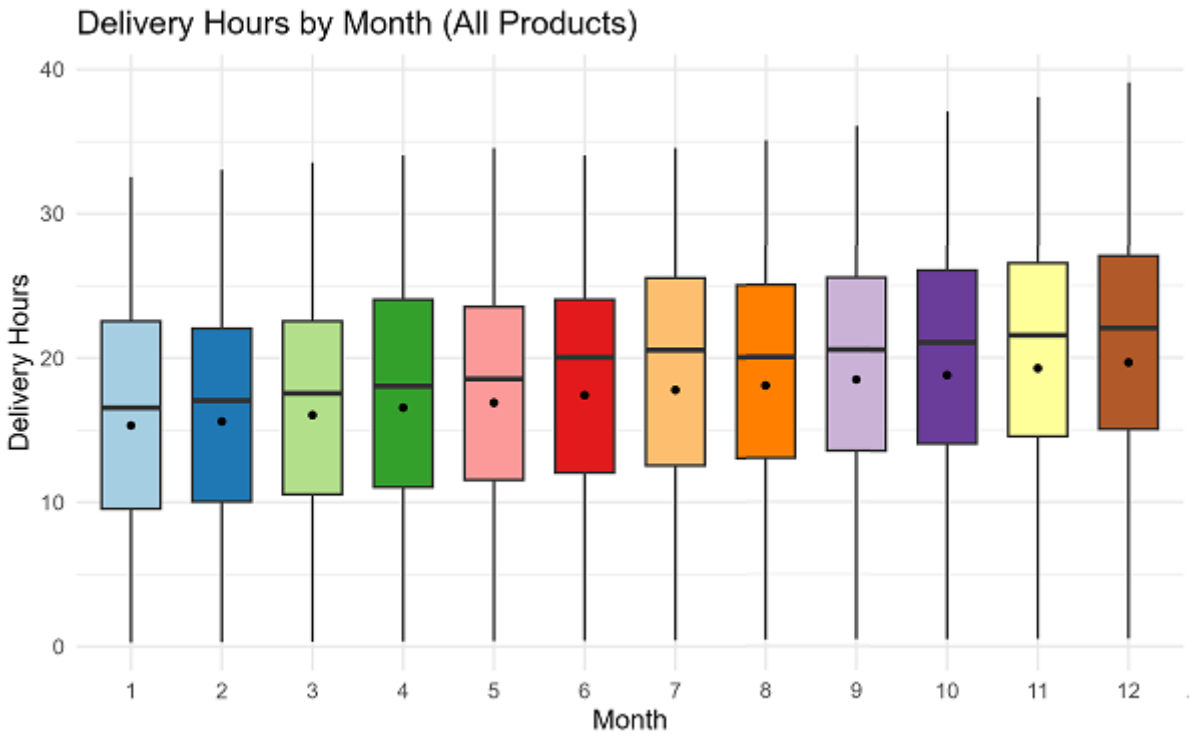

Figure 33 Delivery Hours by Month (All Products)

Figures 32 and 33 above lead us to encounter our first valuable hypothesis, as there is indeed a change in delivery hours as the months pass by. This result was expected based of previous analysis, specifically the SPC section, where we saw that delivery hours go out of control as the months go by. This, once again, is definitely a part of the business that requires attention as long delivery hours lead to unsatisfied, non-returning customers. Competitors might also see this as their chance to outperform our company by having better delivery lead times. We should consider looking at expanding order handling capacity, improving inventory management or training staff/workforce to execute their duties better and quicker.

Part 7 – Reliability of Service

In this section of the report, we have another entirely new scenario. A car rental agency provides the following table:

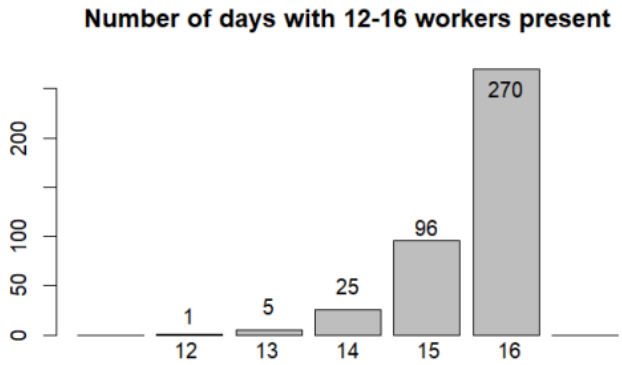

Figure 34 Number of days worked with x workers present

Figure 34 above reflects the number of people on duty at the agency over 397 days.

7.1

We are required to estimate how many days per year we are expected to provide reliable service. First, we define 15 as the minimum number of workers for reliable service.

Now, we have the following:

Total days observed: 397

Days with reliable service: (96 + 270) = 366

The expected number of reliable days can then be calculated as follows:

$$(366/397) \times 365 = 336.4987\ldots = 336 \text{ days}$$

7.2

We are required to optimize profit for the company based of the following information:

- There are problems if there are less than 15 workers on a given day
- Problematic days yield R20 000 less in sales for that day
- More people can be employed at R25 000 per month per worker

Let us model the data as a binomial problem:

Let n_s be days with staffing s an element of {12, 13, 14, 15, 16}

I (s < 15) = 1, 0 otherwise.

The expected problem days after hiring k people is:

$$D(k) = \sum ns \times \{s + k < 15\}$$

Where s + k represents the effective staffing.

The total annual cost of staffing k is then:

$$C(k) = loss\ per\ problem\ day \times D(k) \times \left(\frac{365}{397} + k\right) \times 300\ 000$$

We then compute the annual cost and D(k):

Our problem days are:

$n_{12} = 1$

$n_{13} = 5$

$n_{14} = 25$

and we have a scale factor of f = 365/397 = 0.914

For k = 0

Problem days = 1 + 5 + 25 = 31

Ann. Loss = 31 × f × 20 000 = 570 000

Staff Cost = 0

Total = R570 000

For k = 1

Problem days = 1 + 5 = 6

Ann. Loss = 6 × f × 20 000 = 110 400

Staff Cost = 300 000

Total = R 410 400

For k = 2

Problem days = 1

Ann. Loss = 1 × f × 20 000 = 18 400

Staff Cost = 2 × 300 000

Total = R 618 400

For k = 3

Problem days = 0

Ann. Loss = 0

Staff Cost = 3 × 300 000

Total = R 900 000

Conclusion:

Adding 0 staff members, although incurring no staff cost, still costs the company R570 000 of annual losses. By adding 1 extra staff member, we lower the annual losses significantly. Adding another staff member for a total of 2 drops the annual loss even more, however now the staff cost incurred becomes more significant. At 3 added staff members, we find no losses, however the staff costs are now at almost a million Rand. Adding one extra staff member is thus the recommendation as this minimizes annual loss while maximizing company profit.

Bibliography

Stellenbosch University, 2025. QA344Statistics. [pdf] Stellenbosch University. Available at: QA344Statistics.pdf [on SUNLearn]

Illowsky, B. Dean, S. De Anza College. Introductory Statistics. OpenStax.