ANALYSIS AND MANIPULATION OF SALES DATA USING R PROGRAMMING METHODS

Prepared For: The Engineering Council of South Africa

Prepared By: A Jahura (Student No: 23678941)

Date: 21 October 2022

Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Plagiarism Declaration

I have read and understand the Stellenbosch University Policy on Plagiarism and the definitions of plagiarism and self-plagiarism contained in the Policy (Plagiarism: The use of the ideas or material of others without acknowledgement, or the re-use of one's own previously evaluated or published material without acknowledgement or indication thereof (self-plagiarism or text-recycling)).

I also understand that direct translations are plagiarism, unless accompanied by an appropriate acknowledgement of the source. I also know that verbatim copy that has not been explicitly indicated as such, is plagiarism.

I know that plagiarism is a punishable offence and may be referred to the University's Central Disciplinary Committee (CDC) who has the authority to expel me for such an offence.

I know that plagiarism is harmful for the academic environment and that it has a negative impact on any profession.

Accordingly all quotations and contributions from any source whatsoever (including the internet) have been cited fully (acknowledged); further, all verbatim copies have been expressly indicated as such (e.g. through quotation marks) and the sources are cited fully.

I declare that, except where a source has been cited, the work contained in this assignment is my own work and that I have not previously (in its entirety or in part) submitted it for grading in this module/assignment or another module/assignment.

I declare that have not allowed, and will not allow, anyone to use my work (in paper, graphics, electronic, verbal or any other format) with the intention of passing it off as his/her own work.

I know that a mark of zero may be awarded to assignments with plagiarism and also that no opportunity will be given to submit an improved assignment.

Signature: …………………………… Student no: …..23678941…..

Name: ………Anesu Jahura .... Date: …...21 October 2022……

Abstract

This report entails the findings from a detailed sales data analysis and manipulation project performed by an undergraduate industrial engineering student in partial fulfilment of the graduate attributes set by the Engineering Council of South Africa (ECSA). The data was manipulated in R Studio using R programming methods, and a detailed discussion of the results thereof is given throughout this report. Multiple approaches are taken to analyse the data, ranging from descriptive statistics, statistical process control, delivery process optimisation, and MANOVA analysis. The reliability of service of products is also conducted. The report concludes with a summary of all the important problems and recommendations presented to the business.

Table of contents

List of Figures

List of Tables

1. Introduction

The BEng in Industrial Engineering degree at Stellenbosch University requires students to comply with ECSA graduate attributes, which determine the skills and knowledge that graduate engineers are required to possess. This project is done in partial fulfilment of ECSA graduate attribute four, which states that graduate engineers are required to "demonstrate competence to conduct investigations, analyse data and design experiments. Knowledge of data and visual analysis to conduct explorative data analysis, both descriptive and visually". This project aims to fulfil this requirement while providing useful insights to an online business in order to improve their processes.

2. Data Wrangling

The data supplied for this project contained 180 000 instances of sales data for various items purchased. The descriptive features were the age of the customer; the class of item bought (from seven different classes); the price of the item; the year, month, and day on which the item was purchased; the time it took for the product to be delivered; and the reason the customer bought the item. A sample of the first twenty instances of the original sales data can be seen in Figure 1:

	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	1	19966	54	Sweets	246.21	2021	7	3	1.5	Recommended
2	2	34006	36	Household	1708.21	2026	4	1	58.5	Website
3	3	62566	41	Gifts	4050.53	2027	8	10	15.5	Recommended
4	4	70731	48	Technology	41843.21	2029	10	22	27.0	Recommended
5	5	92178	76	Household	19215.01	2027	11	26	61.5	Recommended
6	6	50586	78	Gifts	4929.82	2027	4	24	14.5	Random
7	7	73419	35	Luxury	108953.53	2029	11	13	4.0	Recommended
8	8	32624	58	Sweets	389.62	2025	7	2	2.0	Recommended
9	9	51401	82	Gifts	3312.11	2025	12	18	12.0	Recommended
10	10	96430	24	Sweets	176.52	2027	11	4	3.0	Recommended
11	11	87530	33	Technology	8515.63	2026	7	15	21.0	Browsing
12	12	14607	64	Gifts	3538.66	2026	5	13	13.5	Recommended
13	13	24299	52	Technology	27641.97	2024	5	29	17.0	Browsing
14	14	77795	92	Food	556.83	2025	6	3	3.0	Random
15	15	62567	73	Clothing	347.99	2024	3	29	8.5	Website
16	16	14839	47	Technology	54650.41	2027	12	30	18.5	Recommended
17	17	96208	44	Technology	14739.09	2028	3	17	13.0	Recommended
18	18	39674	69	Technology	22315.17	2026	8	20	20.5	Recommended
19	19	98694	74	Sweets	546.48	2025	5	9	2.0	Recommended
20	20	99187	54	Luxury	81620.21	2027	9	14	3.0	Recommended

Showing 1 to 20 of 180,000 entries, 10 total columns

Figure 1: Sample of first twenty instances from sales data table (Source: Anesu Jahura)

Before the data could be analysed, it had to be cleaned to ensure that only valid instances, i.e., instances that are realistic, are included in the analysis. To achieve this, the original sales data was indexed to include only the instances for which the sales price was bigger than or equal to zero, and for which the sales price was an actual number. Each instance had to satisfy both of these conditions. The clean data was then allocated to a variable, and the unclean data was allocated to another variable. A list of indices were attached to both the clean data and the unclean data to ensure that the data can be ordered in continuous indices, even when the data "jumped" an instance to exclude an invalid instance. Both the final clean and unclean data tables were written to new Microsoft Excel Files. Twenty samples of the final tables in R are displayed in Figure 2 and Figure 3 below:

	indices_clean_data	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
12334	12334	12334	94724	45	Clothing	942.31	2021	12	9	8.0	Recommended
12335	12335	12335	37617	62	Sweets	58.88	2028	5	19	2.5	Recommended
12336	12336	12336	69090	25	Technology	50894.17	2022	5	10	15.5	Browsing
12337	12337	12337	21028	33	Sweets	197.22	2026	5	7	3.0	Recommended
12338	12338	12338	38677	64	Food	224.82	2025	5	2	2.5	Recommended
12339	12339	12339	69019	44	Gifts	1795.49	2023	6	1	10.5	Recommended
12340	12340	12340	16593	30	Clothing	633.99	2021	5	10	9.0	Website
12341	12341	12341	78109	71	Gifts	3424.01	2025	7	29	12.5	Random
12342	12342	12342	88576	54	Gifts	752.79	2028	7	21	13.5	Recommended
12343	12343	12343	27986	37	Clothing	712.19	2021	10	10	9.0	Recommended
12344	12344	12344	90260	34	Luxury	42891.66	2025	8	4	4.0	Recommended
12346	12345	12346	92286	32	Technology	38167.24	2028	7	6	19.5	Website
12347	12346	12347	89263	44	Clothing	891.71	2021	7	2	8.5	Recommended
12348	12347	12348	71191	49	Household	14936.31	2025	10	11	43.5	Recommended
12349	12348	12349	24801	28	Food	425.96	2022	1	29	2.5	Recommended
12350	12349	12350	85475	57	Luxury	78817.55	2026	3	21	5.0	Browsing
12351	12350	12351	61842	24	Clothing	1008.78	2025	7	16	8.0	Recommended
12352	12351	12352	49373	34	Technology	17277.26	2024	10	11	14.5	Browsing
12353	12352	12353	40283	45	Technology	16930.76	2025	3	9	27.5	EMail
12354	12353	12354	19084	56	Sweets	171.81	2026	10	8	1.5	Random

Showing 12,333 to 12,353 of 179,978 entries, 11 total columns

Figure 2: Sample of twenty instances from the clean data table (Source: Anesu Jahura)

The clean data table containts 179 978 entries, indicating that 22 invalid entries have been removed. In the sample data, we can see from the "X" column, which displays the same numbers as the most leftmost column automatically generated by R Studio, that the numbers jump from "12344" to "12346". This indicates that an instance was skipped due to it being invalid. The "indices_clean_data" column that was added ensures that there is a continuous number of indices for each instance of the data.

	indices_unclean_data	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
12345	1	12345	18973	93	Gifts	NA	2026	6	11	15.5	Website
16320	2	16320	44142	82	Household	-588.8	2023	10	2	48.0	EMail
16321	3	16321	81959	43	Technology	NA	2029	9	6	22.0	Recommended
19540	4	19540	65689	96	Sweets	-588.8	2028	4	7	3.0	Random
19541	5	19541	71169	42	Technology	NA	2025	1	19	20.5	Recommended
19998	6	19998	68743	45	Household	-588.8	2024	7	16	45.5	Recommended
19999	7	19999	67228	89	Gifts	NA	2026	2	4	15.0	Recommended
23456	8	23456	88622	71	Food	NA	2027	4	18	2.5	Random
34567	9	34567	18748	48	Clothing	NA	2021	4	9	8.0	Recommended
45678	10	45678	89095	65	Sweets	NA	2029	11	6	2.0	Recommended
54321	11	54321	62209	34	Clothing	NA	2021	3	24	9.5	Recommended
56789	12	56789	63849	51	Gifts	NA	2024	5	3	10.5	Website
65432	13	65432	51904	31	Gifts	NA	2027	7	24	14.5	Recommended
76543	14	76543	79732	71	Food	NA	2028	9	24	2.5	Recommended
87654	15	87654	40983	33	Food	NA	2024	8	27	2.0	Recommended
98765	16	98765	64288	25	Clothing	NA	2021	1	24	8.5	Browsing
144443	17	144443	37737	81	Food	-588.8	2022	12	10	2.5	Recommended
144444	18	144444	70761	70	Food	NA	2027	9	28	2.5	Recommended
155554	19	155554	36599	29	Luxury	-588.8	2026	4	14	3.5	Recommended
155555	20	155555	33583	56	Gifts	NA	2022	12	9	10.0	Recommended

Showing 1 to 20 of 22 entries, 11 total columns

Figure 3: Sample of twenty instances from the unclean data table (Source: Anesu Jahura)

The sample of twenty instances displayed in Figure 3 shows us all of the invalid instances that were not included in the clean data table. All of the instances either have negative values or "NA" (missing) values for the "Price" feature – this is not possible because all the items have to be sold at some price, which can only be a positive amount. The data has now been successfully wrangled.

3. Descriptive Statistics

3.1 Spread of the Data

The clean data contains many different characteristics that can be analysed and investigated. A statistical summary for three features is given in Table 1 below:

Table 1: Statistical Summary for Three Features of the Data (Source: Anesu Jahura)

	Min	Mean	Max	Range	Standard Deviation
Price	35.65	12294.10	116618.97	116583.3	20889.15
Delivery Time	0.5	14.5	75.0	74.5	13.95578
Age	18	54.57	108.00	90	20.38881

The range of prices of goods sold, as well as the high standard deviation in price, are expected due to the wide variety of different products that are sold in all the classes. However, the mean indicates that the sales data is dominated by products with high prices. This could be a good sign because it could mean that the business is making a lot of high-profit sales. Management should consult the finance department to find out whether their gross profit margins are indeed high and whether they can improve profits even more.

The maximum delivery time is very high, which could indicate problems with the delivery process. However, the data needs to be looked at in more detail to able to make a more informed analysis. The minimum value and the standard deviation for delivery is reasonable and as expected.

It is clear from the minimum age that only adults purchase goods from this business. This is very unexpected, as teenagers and pre-teens are expected customers for some of the classes of goods sold. One would expect that people under the age of 18 would purchase the most sweets, and young people are also expected to sometimes buy their own clothes (especially fashion-obsessed teenagers). The reason for this lack of data needs to be investigated by management. The standard deviation in age is expected.

3.2 Graphical Analysis

Graphical representation of the data is highly useful to identify trends that are difficult to see easily with numbers. There were a high number of different graphs all showing different insights generated in R, and the most useful of these are analysed in the following figures:

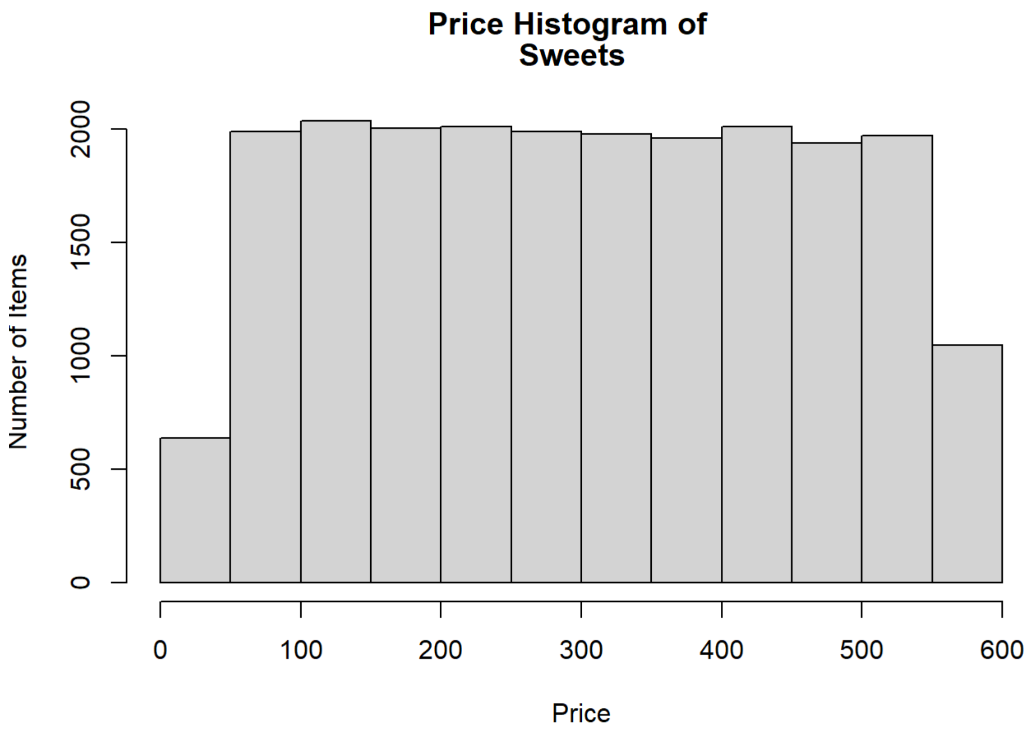

Figure 3: Histogram for Count of Sweets vs Price (Source: Anesu Jahura)

The prices of items seem to be evenly distributed across the “sweets” class. At the extreme ends of the prices for sweets items sold, the number of items is the least. This shows that most customers who purchase sweets do so in the R50 – R550 range, and less customers purchase outside this range. There weren’t any sweets items purchased for more than R600. Management should devise ways to get more customers to purchase sweets with higher prices, such as attracting wholesalers to buy more sweets by offering discounts for bulk purchases.

Figure 4: Histogram for Count of Sweets vs Delivery Time (Source: Anesu Jahura)

The delivery time for the “sweets” class shows a normal distribution which peaks at about 2.7 hours, which indicates the range around which the most delivery hours are. There are few deliveries in the one-hour range, which is expected, but the business will benefit from increased customer satisfaction if this number can be improved. The business should also try to have less deliveries in the 3–4-hour range to increase customer satisfaction.

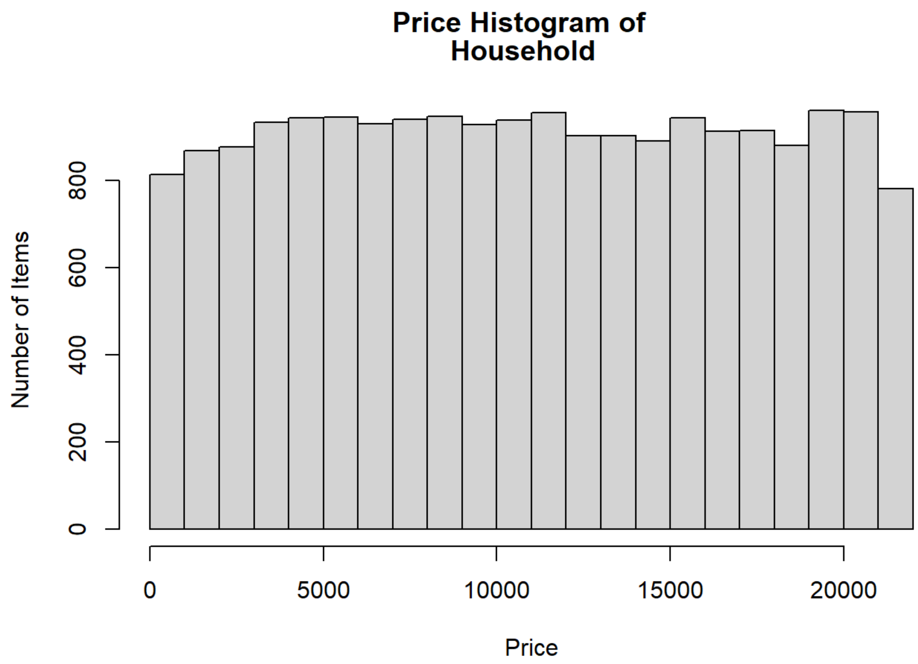

Figure 5: Histogram for Count of Household Items vs Price (Source: Anesu Jahura)

Figure 5 shows that prices for household items are evenly spread across multiple ranges between R0 and just over R20 000. This shows that the customers of the business are from multiple income bands, and they purchase items of various expenses. This diverse customer base shows that the business is able to cater for a wide range of customers and can attract customers from all walks of life.

Figure 6: Histogram for Count of Luxury Items vs Delivery Time (Source: Anesu Jahura)

Figure 6 shows that the delivery times for luxury times are slightly skewed to the right, favouring lower delivery times. This is to be expected, as people who purchase luxury items often pay for shorter delivery times. The business should aim to make delivery times even more skewed to the right, to ensure that customers who buy luxury items are kept happy – these customers can offer the business a lot in terms of profits due to the expensive items that they purchase.

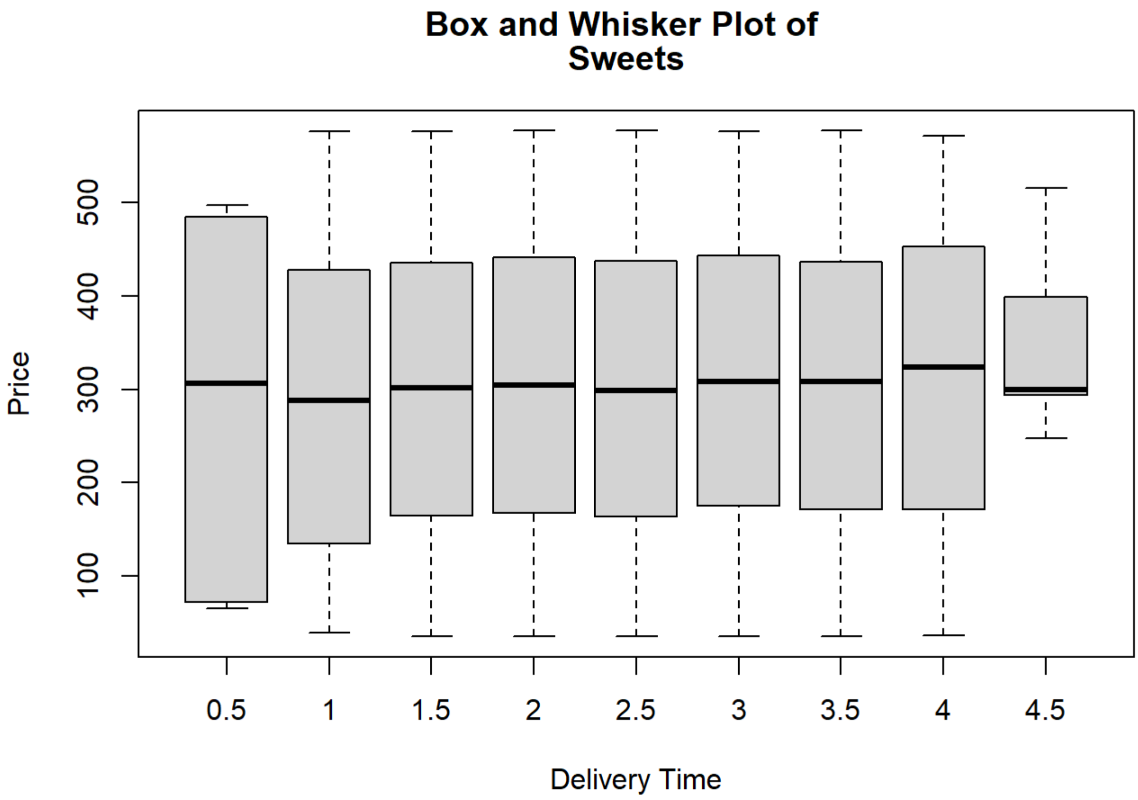

Figure 7: Box & Whisker Plot of Sweets Class (Source: Anesu Jahura)

The side-by-side box and whisker plot comparing prices vs delivery times for the “sweets class” shows unexpected results. It is expected for higher-priced items to have shorter delivery times, but this is not the case with sweets. The box plot for the 4.5-hour delivery time shows that more than half of the items are priced higher than R300, while the other boxplots show similar results across delivery times up till 0.5 hours. Therefore, the price doesn’t seem to influence the delivery times for the sweets class. Management can use this discovery to make more profits by offering customers free or expedited delivery if they buy products for a certain price or higher. This will encourage customers to buy higher-priced items from which the business can make more profits.

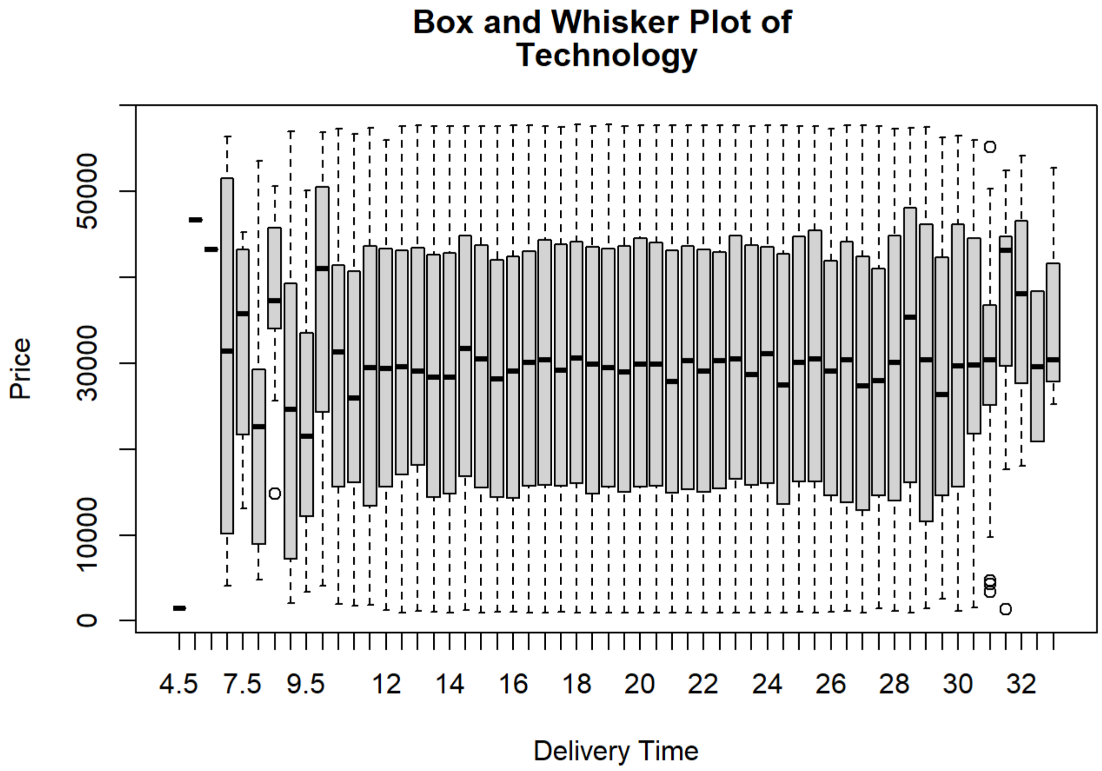

Figure 8: Box & Whisker Plot of Technology Class (Source: Anesu Jahura)

The side-by-size box & whisker plot shows a somewhat expected pattern. The medians for the lowest delivery times (between 4.5 and about 9 hours) are the highest in terms of price compared to the other classes, and more than half of the samples are well above R3000. This means that for technology items, customers who buy more expensive items tend to receive their goods in a much shorter time. However, the medians slightly increase between about 30 hours and 32 hours, indicating relatively high-priced items with very long delivery times. Management needs to look into this discrepancy and figure out why this has happened. For example, it could be that these items are very expensive pieces of technology that have to be sourced from further distances than most other products.

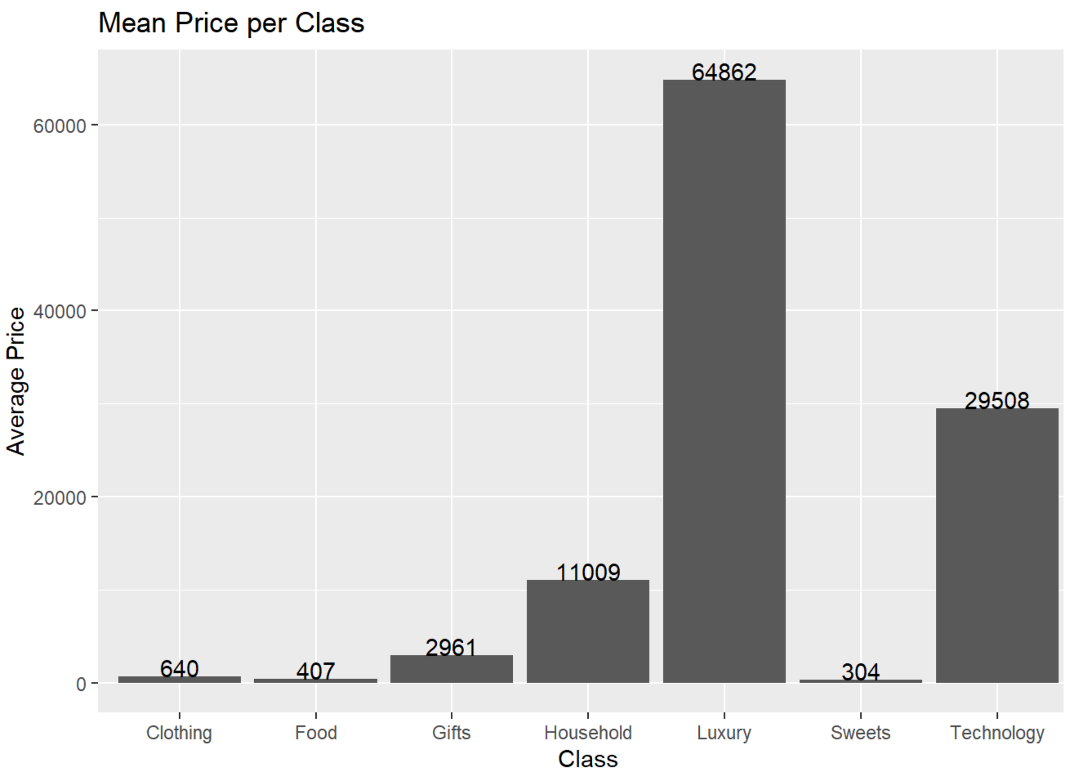

Figure 9: Histogram of Mean Price per Class (Source: Anesu Jahura)

Figure 9 shows that luxury items have by far the highest average selling price, which is expected. Technology items have the second highest average price, which is also expected. Household follows thereafter. The “gifts” class only has an average selling price of R2961, which is strange considering that gifts that people buy online often consist of technology and household items, both of which have much higher average selling prices. Management needs to investigate the composition of gifts bought to determine why this selling price is so low.

If the data still makes sense after the “gifts” class investigation, then the business should find ways to get customers to buy more expensive gifts, such as running Cyberweek promotions to get people to buy more technology items. The clothing average is as expected. The food average of R407 is quite high, and if this is for a single meal, then customers are spending generously in this category – management should ensure that remains the case. The sweets average of R304 is also very high, and it indicates that people are already buying in bulk. Again, encourage customers to buy even more sweets in bulk through discounts will increase the average price of sweets and, in turn, increase profits.

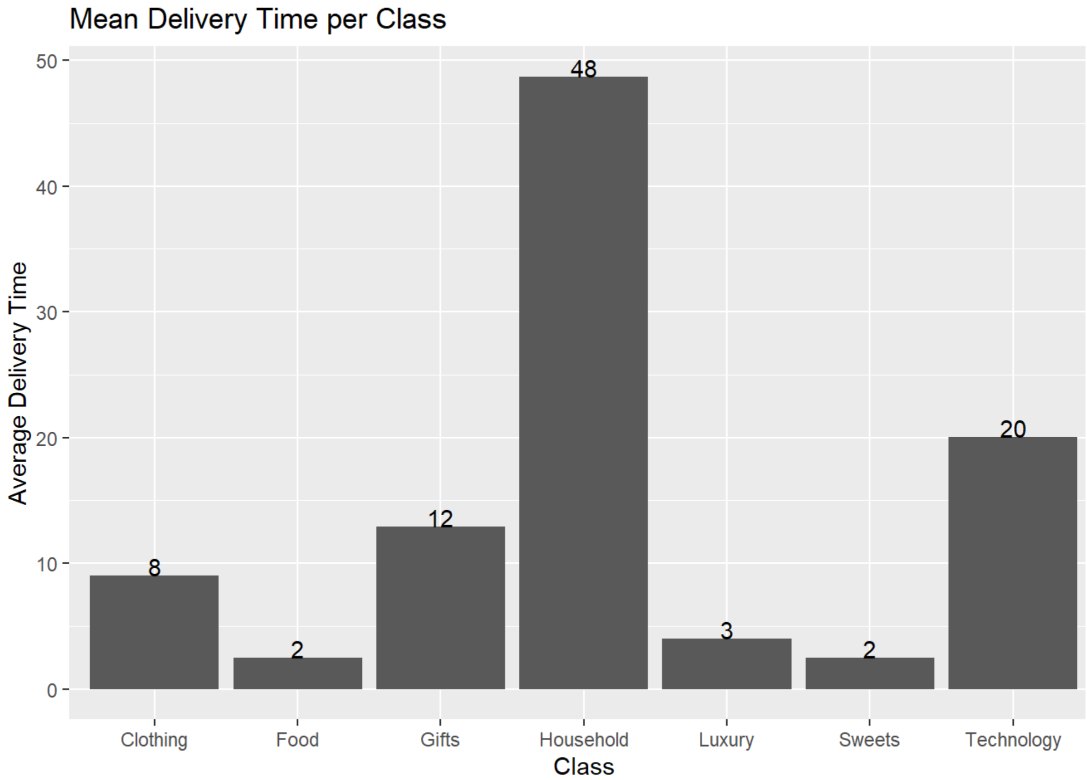

Figure 10: Histogram of Mean Delivery Time per Class (Source: Anesu Jahura)

Figure 10 reveals that household items are by far the class with the longest delivery times. Management needs to investigate why household items have the longest delivery times. It could be because many household items are custom-made or not easily available and have to be sourced from elsewhere, or it could be that the department in charge of household items is underperforming. The technology class average is satisfactory, but management needs to determine how this could possibly decrease to increase customer satisfaction. The clothing delivery time is acceptable. The gifts class is good considering that some gifts have special requirements such as having to be wrapped before they are shipped off. The sweets delivery time is good. The food delivery time is satisfactory, but 2 hours is a long time for hungry customers to wait for their food, and management needs to determine how they can deliver food to customers faster.

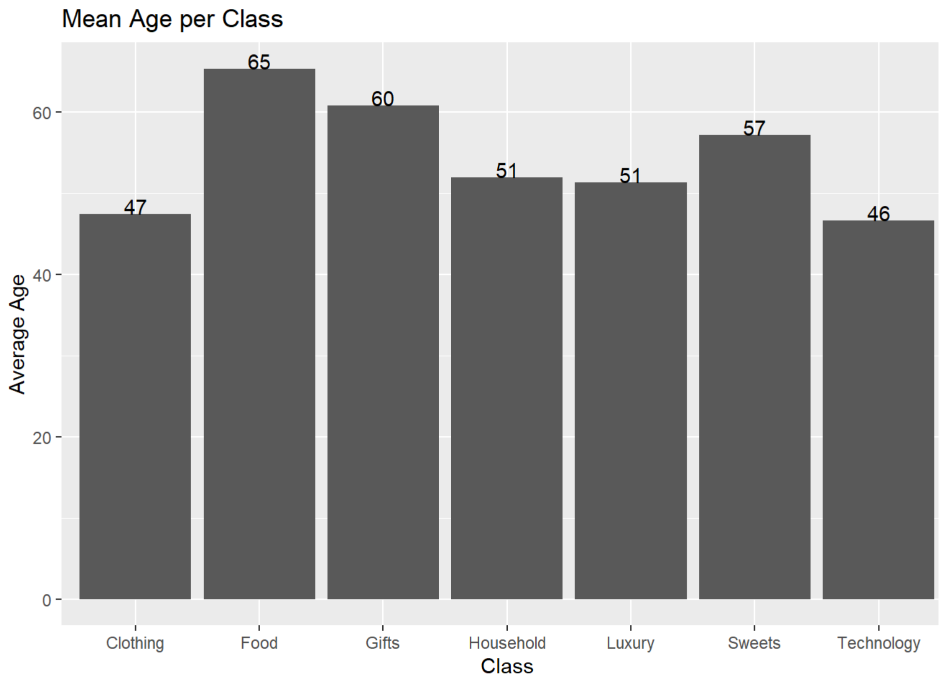

Figure 11: Histogram of Mean Age per Class (Source: Anesu Jahura)

The ages of customers in each class seems to be randomly distributed. The average of the "clothing" class is higher than expected, as higher people are expected to be more interested in spending on buying clothes. The "technology" class average is also higher than expected due to the same reasoning. The "sweets" average is very expected because younger people are expected to be more interested in spending on sweets. However, the mean price per class of R304 (as seen in Figure 9) shows that the sweets customers likely buy in bulk, meaning that they own businesses that sell sweets. This makes the age of sweets customers more reasonable. The household, gifts, and luxury averages are as expected. Management needs to investigate the reason for the "food" average, as it seems slightly higher than expected (expected to be in the 40 range).

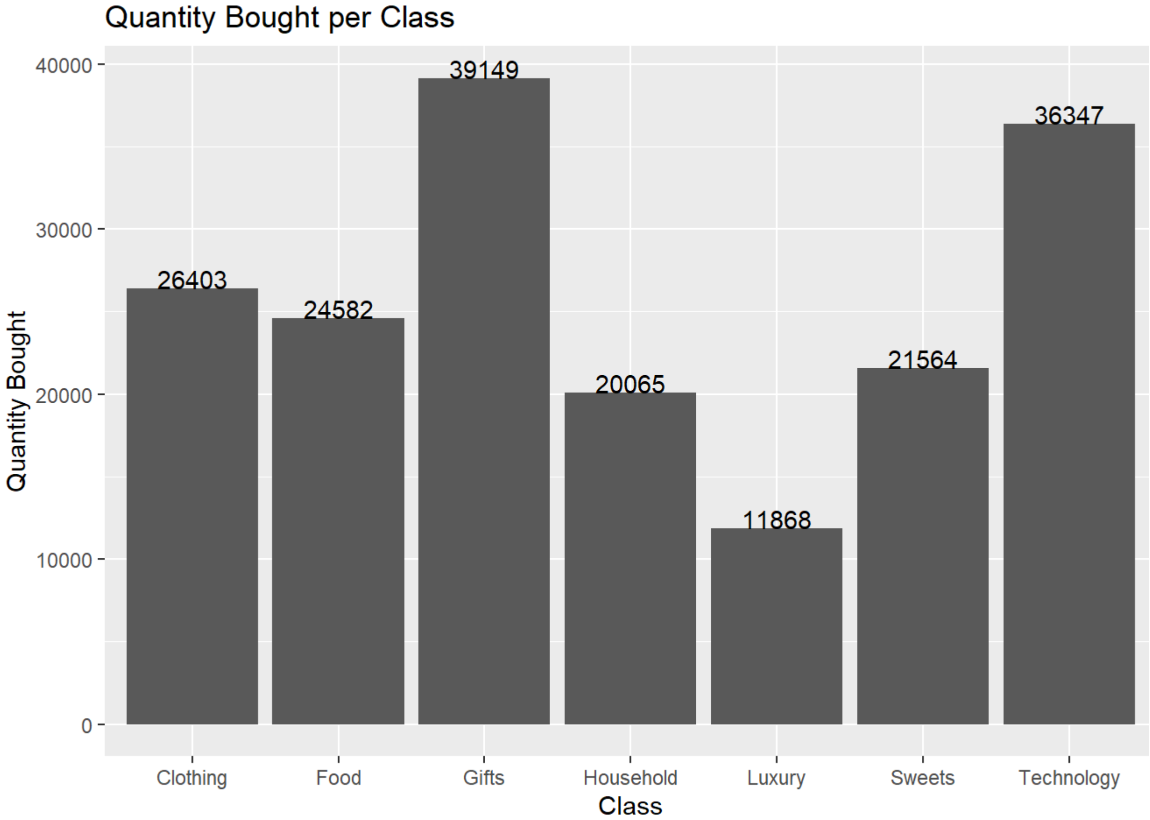

Figure 12: Histogram of Quantity Bought per Class (Source: Anesu Jahura)

Gifts are the most popular class of items. Figure 9 indicated that the average selling price of gifts is R2961, so if the business can attract customers that buy more expensive gifts, then this will be very good for the business's profits. Technology items are the second most popular class, which is expected. The luxury class is the least popular class, which makes sense when considering that these are also the most expensive items. Again, if the business can sell more luxury items, they will boost the profits of the business. The rest of the classes are as expected, and there are no major recommendations to management in this regard.

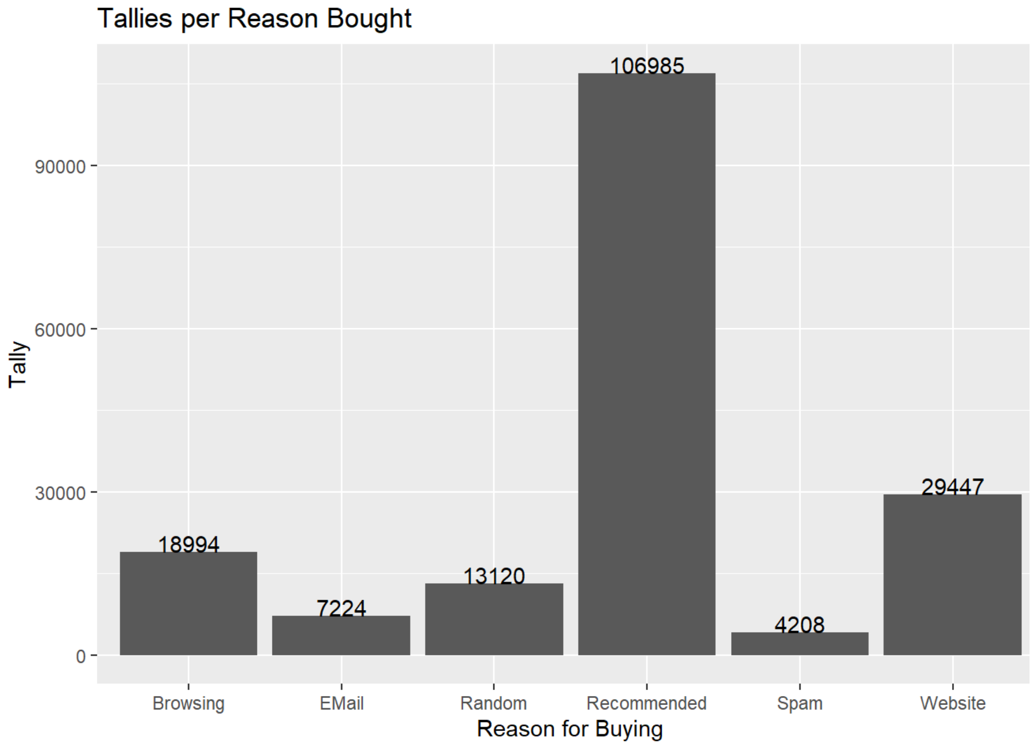

Figure 13: Histogram of Tallies per Reason Bought (Source: Anesu Jahura)

Figure 13 indicates that by a very large margin, customers manly buy products from the business due to recommendations. This speaks volumes of the quality of products that the business sells, because it means that customers who have been impressed with products have recommended it to their friends and families. Therefore, the business should continue to produce products that will satisfy customers and encourage them to recommend the business's products. The "browsing" and "website" reasons indicate that the business needs to focus on increasing its online presence, for example, by using search engine optimisation (SEO) methods to appear more in users' online searches. Furthermore, the business should optimise its website to attract buyers to spend more, for example, by using more enticing graphics or hosting "flash sales" where prices are reduced for a limited time. The business can also find out how to better target customers through email and spam to increase sales. The random reasons for purchase will always be present, but it could be that the business is incorrectly classifying purchase reasons as "random" so the manager should ensure that these purchase reasons are indeed random and not due to inefficient data collection and classification.

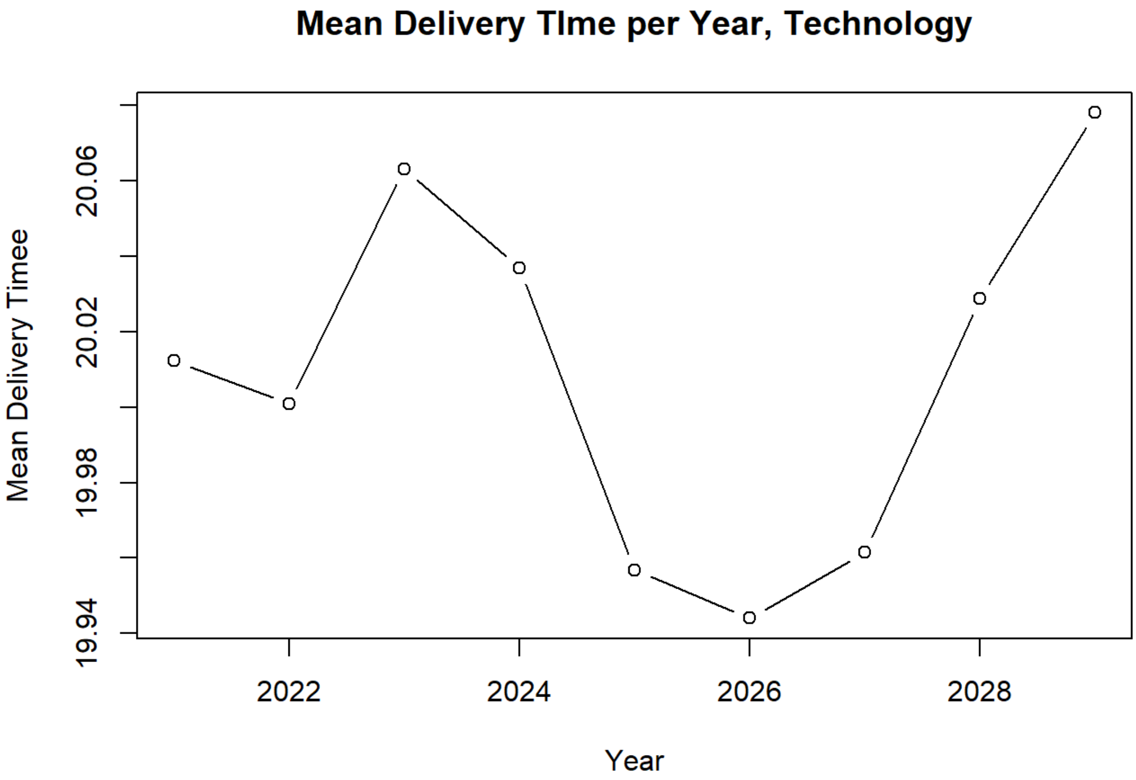

Figure 14: Temporal Line Graph for Technology Delivery Time (Source: Anesu Jahura)

Figure 14 shows that the mean delivery time decreased in some years, which is positive, then increased in other years, which is negative. However, the trend from 2021 to 2026 showed a general decrease in mean delivery time. This showed the business had good control over their delivery process and satisfied customers. The mean delivery time increased from 2026 and continues to increase up to 2029. This shows that the business is having poor control over their delivery process for technology items. However, it could also be a result of suppliers becoming more unreliable. Therefore, the business needs to investigate this and fix the problem to ensure it doesn't adversely affect customer satisfaction.

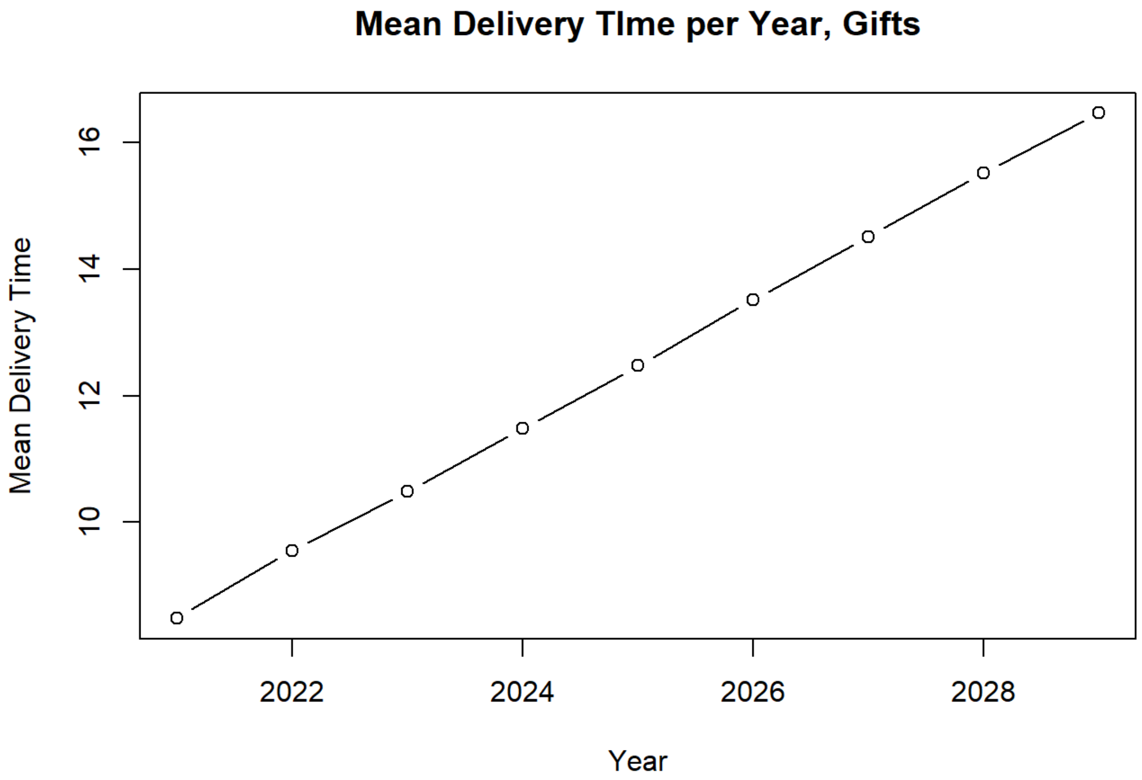

Figure 15: Temporal Line Graph for Gifts Delivery Time (Source: Anesu Jahura)

Figure 15 shows a linear increase in mean delivery times from 2021 to 2029. This means that year after year, the business's control over its delivery process has gotten worse. This is bad for customer satisfaction and needs to be investigated. If it is found that the business itself is to blame, then the manager needs to figure out to how fix the company's control over its delivery process. If it is due to unrealiable suppliers, the business could look for alternative sources of supply. Other outside factors, such as increased traffic in popular routes, might need more creative solutions to be applied.

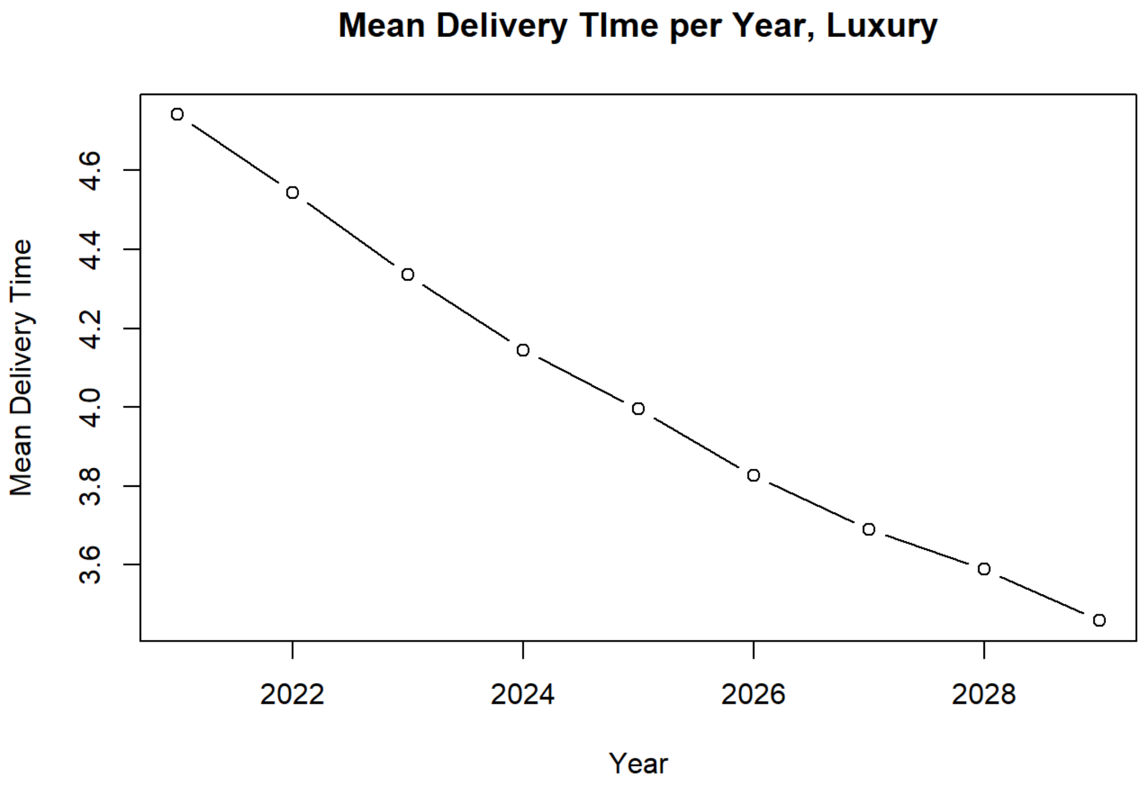

Figure 16: Temporal Line Graph for Luxury Delivery Time (Source: Anesu Jahura)

The mean delivery time for luxury items decreases year-on-year from 2021 to 2029. This could be an indication that the business has been having better control over its delivery process for luxury items. It could also be due to suppliers becoming more reliable or on other factors such as improved transport routes (shorter; less traffic) for the luxury items. The business is doing well in this regard, and it needs to maintain this performance to ensure that luxury customers remain satisfied.

3.3 Process Capability

Process capability is defined as the ability of a process to meet specifications (van Schalkwyk, 2022). In assessing a process's ability to meet specifications, there are multiple metrics we should calculate and analyse to make an informed opinion.

The process capability indices are calculated using the following formulas:

$$C_p = \frac{(USL-LSL)}{6\sigma} \quad (1)$$

$$C_{pu} = \frac{(USL - \mu)}{3\sigma} \qquad (2)$$

$$C_{pl} = \frac{(\mu - USL)}{3\sigma} \qquad (3)$$

$$C_{pk} = \min(C_{pl}, C_{pu}) \qquad (4)$$

When analysing the delivery process times for "technology" class items, the C_p was found to be 1.14, the C_{pu} was found to be 0.380, the C_{pl} was found to be 1.90, and the C_{pk} was thus found to be 0.380.

The C_{pk} value of 0.380 is far below 1.0, which is the acceptable value for a capable process (Process Capability Analysis, n.d.). This shows that the delivery process for technology items is currently incapable. If we look again at Figure 14, we can see that the average delivery times for technology items begins to increase from 2026 and continues to increase up to 2029 (and will probably increase again in the next year), which confirms the current poor capability. It is thus vital that management looks at this process and figure out how to improve it. The business needs to bring the C_{pk} to 1, and then ideally further increase it to above 2 to ensure a highly capable process.

An LSL of 0 is logical because delivery time is measured in hours, and hours cannot be negative. Therefore, the only delivery times possible in practise would be those that are bigger than or equal to 0. Although a delivery time of 0 is not probable, it is still the smallest theoretical value for delivery time that can be achieved.

4. Statistical Process Control

Statistical process control (SPC) is a methodology for monitoring a process to identify special causes of variation and signal the need to take corrective action (Evans & Lindsay, 2020:401). Businesses can use SPC to measure the quality of their goods and services and maintain a state of continuous quality control by analysing samples of data representing goods or services. For example, manufacturers can use SPC to ensure that tolerances on parts are maintained, and that manufacturing processes do not result in many parts that have to be discarded or reworked due to poor conformance to tolerances (i.e., poor quality).

4.1 Control Chart Initialisation

4.1.1 X-Charts

The mean of the x-bar data was calculated and this was selected as the centreline for each class of items. Using the statistical process control formulas from the formula sheet supplied for the QA344 module, the other control limits were calculated as follows:

$$UCL = cl + 0.789\,\bar{s} \tag{5}$$

$$LCL = cl - 0.789\,\bar{s} \tag{6}$$

These upper and lower control limits were used to determine whether the x-bar samples were in control, where samples that were in control lied between the limits, and those that were out of control lied outside the limits. R code was added in the loop to partition the control limits into three equal sections each, allowing a more detailed analysis of samples. Graphs showing the x-bar samples vs the control limits for each class of item were generated, and this enables a detailed analysis of each group of samples for each class. An example of such an analysis, using the four general rules for analysing patterns in control charts given in Evans & Lindsay, is done with the "Luxury" class below:

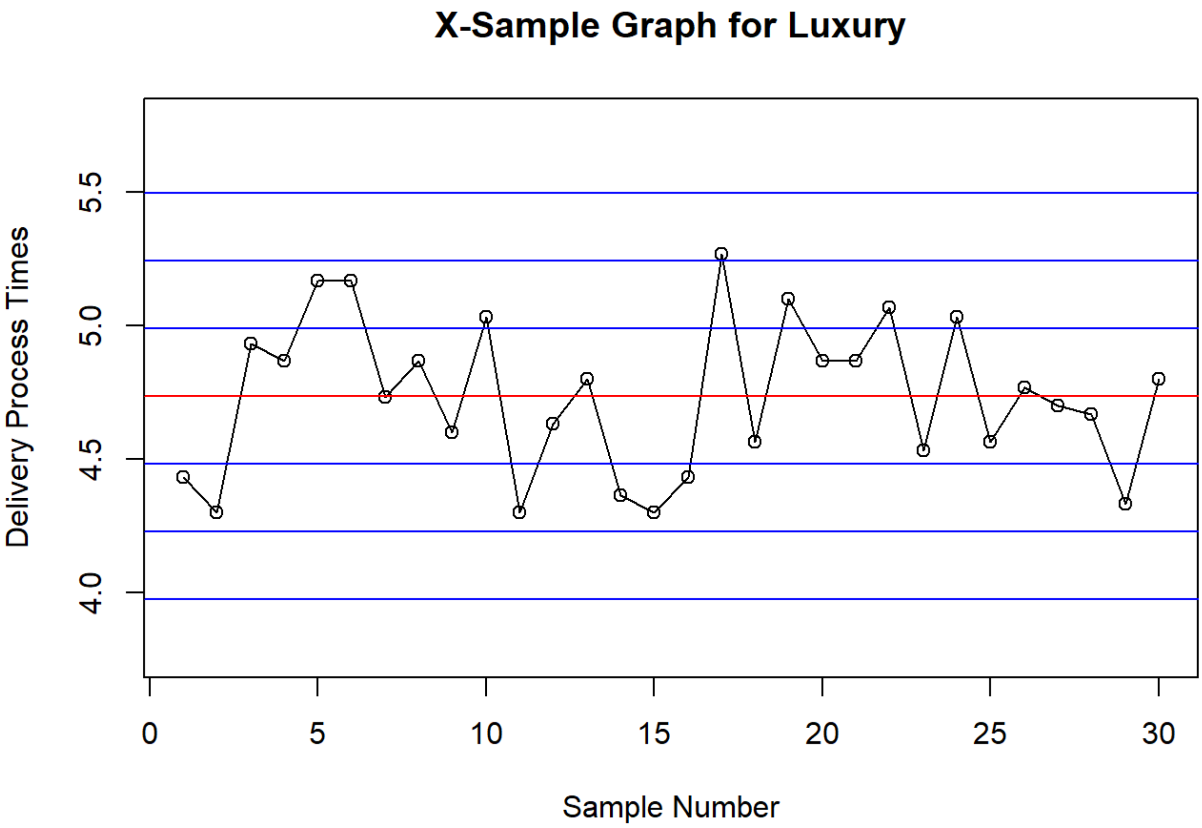

Figure 17: X-Sample Graph for “Luxury” Class (Source: Anesu Jahura)

Figure 17 displays the first 30 samples of x-bar data for the “Luxury” class’s delivery times. This graph shows that there aren’t any samples that are higher than the UCL or lower than the LCL, so it conforms with the first rule. There are 14 samples below the centre line, 15 samples above the centre line, and one sample on the centre line, so it also confirms with the second rule. The points appear to randomly fall above and below the centre line, so the third rule is satisfied. Most points are somewhat near the centre line, so the fourth rule is also satisfied. Therefore, we can ascertain that this process is currently in control.

The values for the x-bar control limits for all the classes were tabulated in R, and the values rounded to 4 decimal places are given in Figure 18:

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22.9746	22.1079	21.2412	20.3744	19.5077	18.641	17.7743
Clothing	9.4049	9.26	9.115	8.97	8.825	8.68	8.5351
Household	50.2483	49.0196	47.7909	46.5622	45.3335	44.1048	42.8761
Luxury	5.494	5.2412	4.9884	4.7356	4.4828	4.2299	3.9771
Food	2.7095	2.6363	2.5632	2.49	2.4168	2.3437	2.2705
Gifts	9.4886	9.1127	8.7369	8.3611	7.9853	7.6095	7.2337
Sweets	2.897	2.7573	2.6175	2.4778	2.338	2.1983	2.0585

Figure 18: SPC x Chart (Source: Anesu Jahura)

This gives us some insight about what we expect the process for controlling delivery process times to look like. We expect “food” and “sweets” items have delivery times close to 2.49 and 2.4778 hours (the values for the centre lines) respectively, which makes sense because food orders should be delivered timeously to hungry customers. “Luxury” items also have a relatively low centre line value of 4.7356 items, and this can be attributed to the fact that luxury items are high-cost and ordered by customers who can afford to pay for expedited delivery. “Household” items have the highest centre line value of 46.5622 hours, and this is reasonable because household items are usually not needed immediately, thus, customers wouldn’t mind waiting. For all of the classes of items, the size of the UCL and LCL correspond closely to the centre line.

4.1.2 S Charts

To get the s-bar data, the standard deviation function was applied to the control data. The centre lines for all the classes taken as the average of the standard deviations of each sample of 15 instances. As per the statistical process control formulas from the formula sheet supplied for the QA344 module, the other control limits were calculated as follows:

$$UCL = 1.572\ cl \tag{7}$$

$$LCL = 0.428\ cl \tag{8}$$

Again, R code was added in the loop to partition the control limits into three equal sections each, allowing a more detailed analysis of samples. This time, graphs showing the s-bar samples vs the control limits for each class of item were generated. An example of such an analysis, again using the four general rules for analysing patterns in control charts given in Evans & Lindsay, is done with the "Sweets" class below:

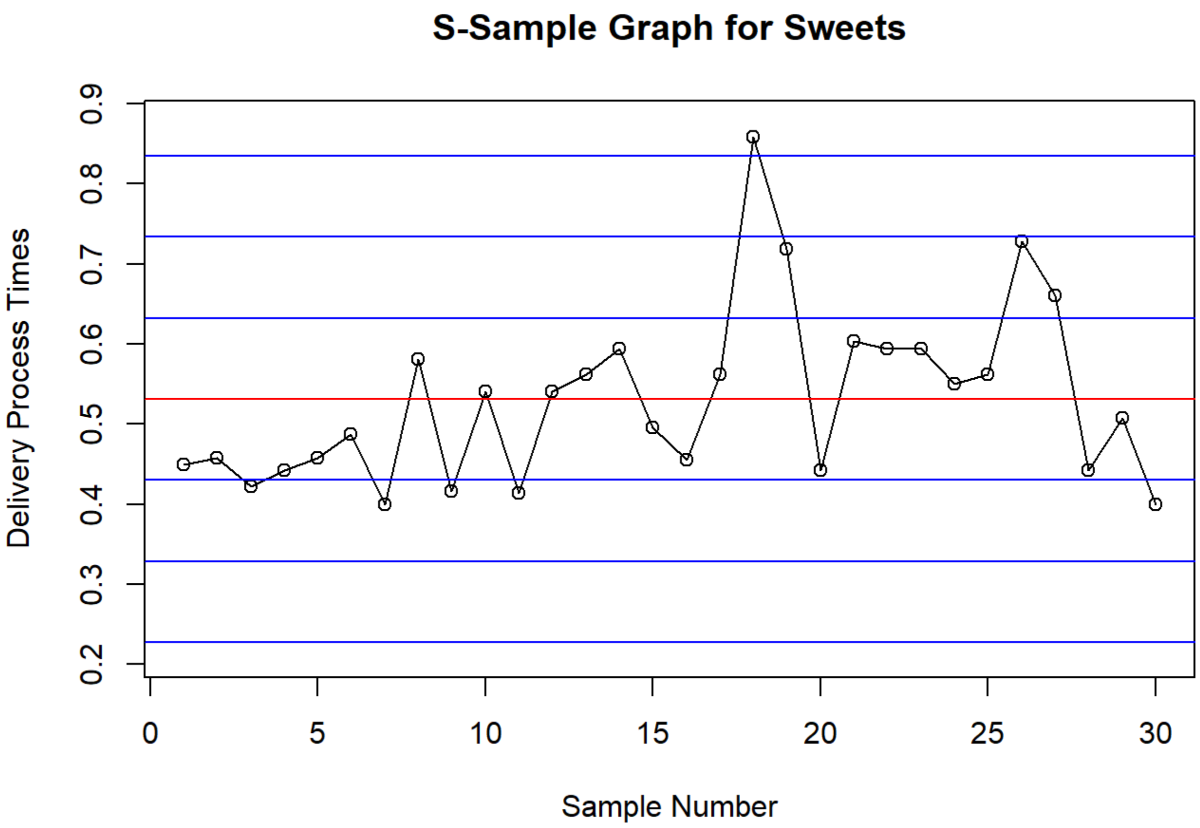

Figure 19: S-Sample Graph for "Sweets" Class (Source: Anesu Jahura)

The S-Sample graph shows that sample 18 is above the UCL, i.e., it is out of control. Practically, this could be a delivery that resulted due to a "special case" such as a broken-down delivery vehicle (McNeese, n.d.). Considering that this is the only point out of the control limits, we recommend for the management to investigate the reason for this sample being out of control to determine if the issue was a once-off occurrence or if the root cause needs to be dealt with. There are an equal number of samples (15) above and below the centre line, so the samples confirm to the second rule. The points fall randomly above and below the centre line, so the third rule is also satisfied. The points are also somewhat close to the centre line, so the fourth rule is satisfied as well. Besides the out-of-control sample that is possibly due to a special case of variation, the process is in control.

The values for the s-bar control limits for all the classes were tabulated in R, and the values rounded to 4 decimal places are given in Figure 20:

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.1806	4.5522	3.9239	3.2955	2.6672	2.0388	1.4105
Clothing	0.8666	0.7615	0.6564	0.5512	0.4461	0.341	0.2359
Household	7.3442	6.4534	5.5626	4.6719	3.7811	2.8903	1.9996
Luxury	1.5111	1.3278	1.1445	0.9612	0.778	0.5947	0.4114
Food	0.4372	0.3842	0.3312	0.2781	0.2251	0.1721	0.119
Gifts	2.2463	1.9739	1.7014	1.429	1.1565	0.8841	0.6116
Sweets	0.8353	0.734	0.6327	0.5314	0.4301	0.3288	0.2274

Figure 20: SPC s Chart (Source: Anesu Jahura)

Analysing the centre line values for each class gives an idea of the variation of the samples in each class. Household items had the greatest variation of samples, meaning that delivery times were highly variable. The UCL for household items is also very high. This could be an issue that management needs to look at, because both a reliable and a consistent delivery process is essential for a business that delivers goods. The technology class also has a relatively high average delivery variability, and this can also be looked at by management to investigate whether this can be brought down to similar values for classes with lower average delivery variability. All other classes have acceptable average delivery variability. Looking at the UCL and LCL for all the classes, they are not very far from the centre line values and thus there is nothing management needs to look at specifically in this regard.

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

5. Optimising the Delivery Process

5.1 Out-of-Control Samples

Both the x-bar and s-bar samples consisted of out-of-control samples across various classes. Using the specified limits, the out-of-control samples were calculated in R and are tabulated below, limited to the first and last 6 samples:

Table 2: Summary of Out-of-Control x-bar and s-bar Samples (Source: Anesu Jahura)

Class of Item	Nr x-bar Samples out of Control	Nr s-bar Samples out of Control	First & Last 6 x-bar Samples	First & Last 6 s-bar Samples
Sweets	5	1	942; 1104; 1243; 1294; 1403;	18
Household	400	54	252; 387; 629; 1335; 1336; 1337	65; 89; 147; 1294; 1299; 1323
Gifts	2290	8	213; 216; 218; 2607;2608;2609	193; 746; 1342; 1855; 2493; 2576
Technology	17	16	37; 398; 483; 1872; 2009; 2071	129; 230; 251; 2095; 2290; 2400
Luxury	434	4	142; 171; 184; 789; 790; 791	103; 254; 543; 600
Food	5	16	75; 663; 1203; 1467; 1515	19; 57; 96; 1422; 1429; 1553
Clothing	17	98	455; 702; 1152; 1677; 1723; 1724	289; 530; 780; 1754; 1756; 1757

It is clear from Table 2 that the delivery process times for household, gifts, and luxury items have a relatively large number of x-bar samples out of control, and further analysis of these out-of-control points is needed. It is also clear that the household class has a relatively high number of s-samples out of control, and this also needs to be looked at more closely.

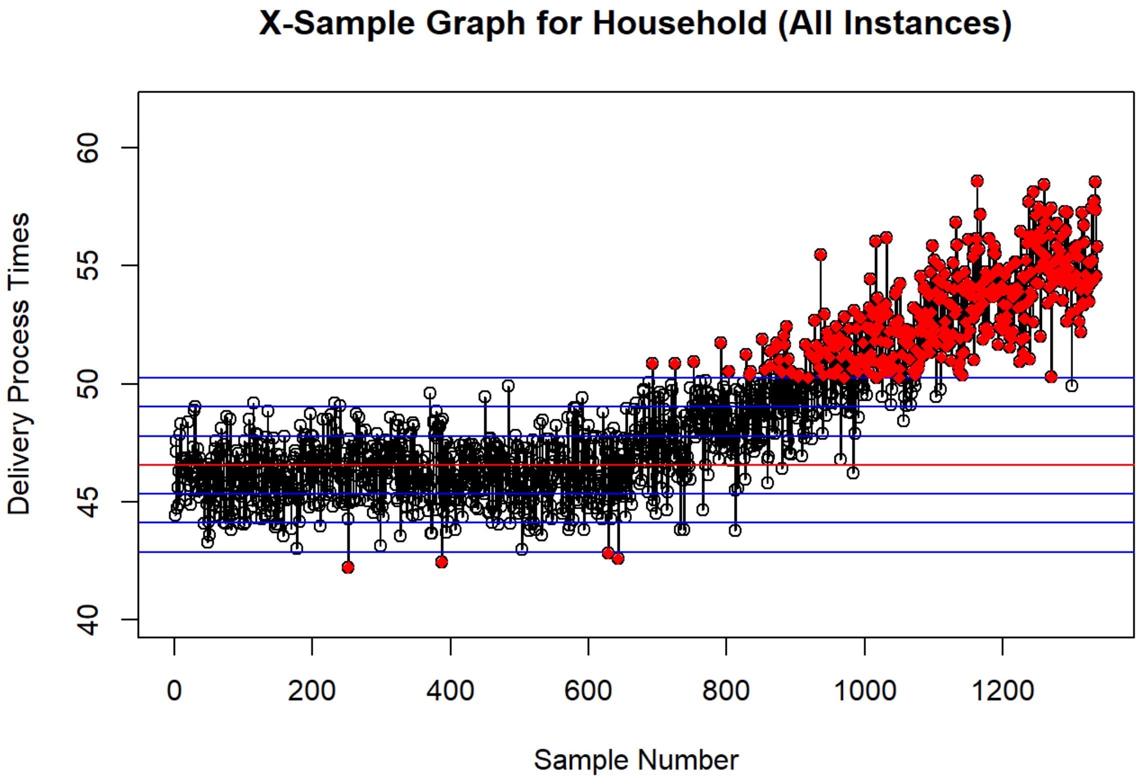

Figure 21: X-Sample Graph for all Household Items (Source: Anesu Jahura)

There are about 7 out-of-control samples in the first 800 samples, which represents an out-of-control rate of just under 1%, which is very good. However, as the samples increase from 800, the process begins to go way out of control, until almost all the samples are out of control after 1200 samples. This is a worrying indication because it shows that the business has poor control over its delivery process. Figure 10 also showed that household items have the highest average delivery times, so the delivery process for household items is clearly a big issue that the business urgently needs to address.

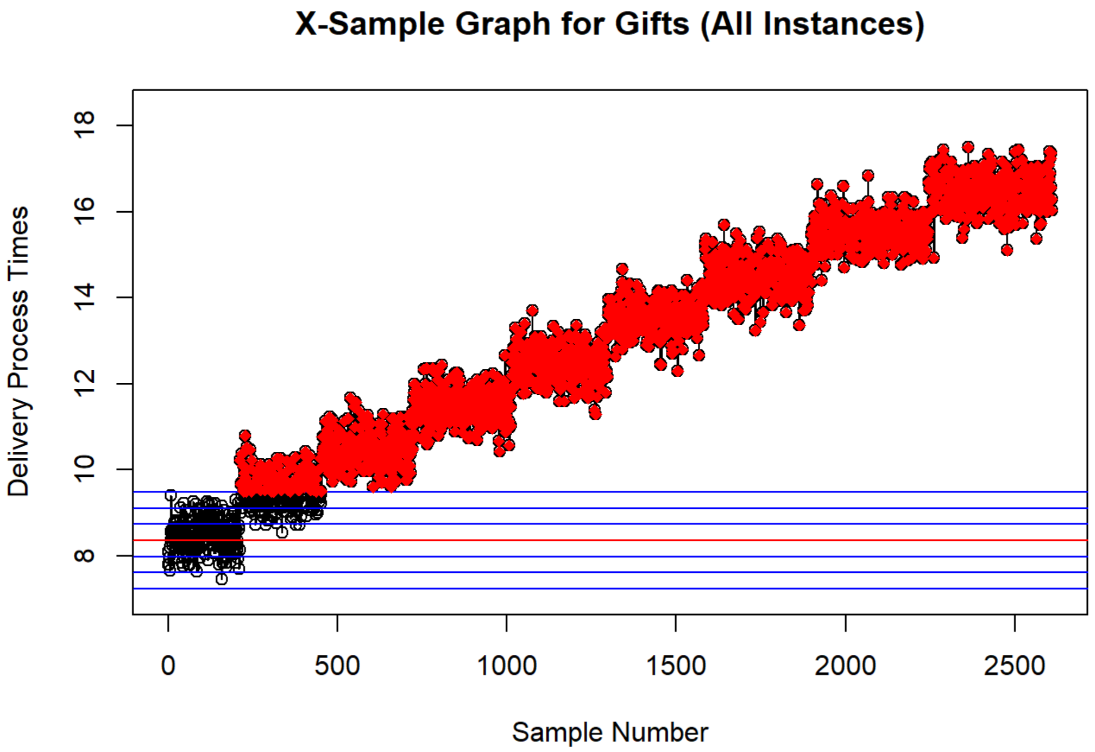

Figure 22: X-Sample Graph for all Gifts Items (Source: Anesu Jahura)

We can see from the X-Sample graph of the "gifts" class that the delivery process goes very far out of control at about 300 samples onwards. The delivery process times increase linearly, indicating a consistent worsening of the control over the delivery process over time. Management needs to investigate the reason for this and determine how they can fix the process to remain within acceptable limits.

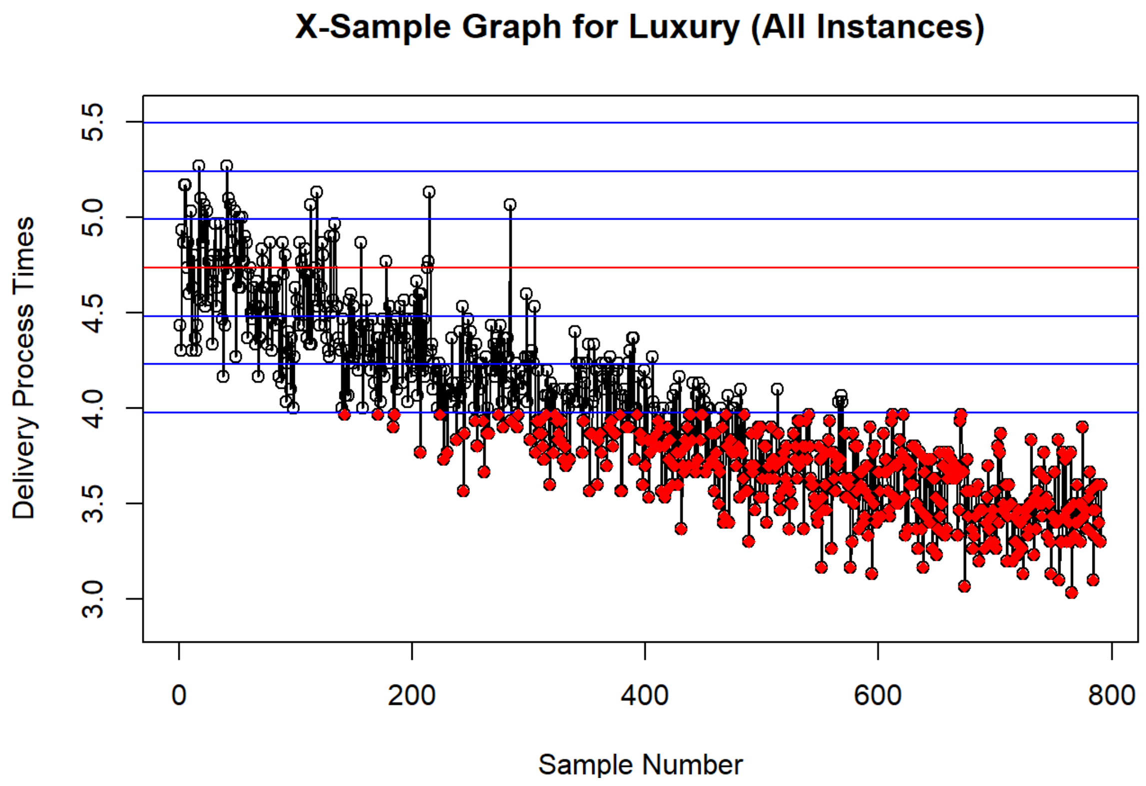

Figure 23: X-Sample Graph for all Luxury Items (Source: Anesu Jahura)

The x-sample graph for the “luxury” graph gives an interesting insight. Although the delivery process begins to gradually fall out of control after about 200 instances, all the out-of-control samples are below the LCL. This means that the delivery process times are lower than should be for the system. This is a good result for the business because customers will be more satisfied with shorter delivery times, which could result in higher profits due to retained customers and more recommendations from happy customers. However, new control limits should be calculated so that the business can look at the process as it is now and increase their expectations (“raise the bar”) so that they can improve even further.

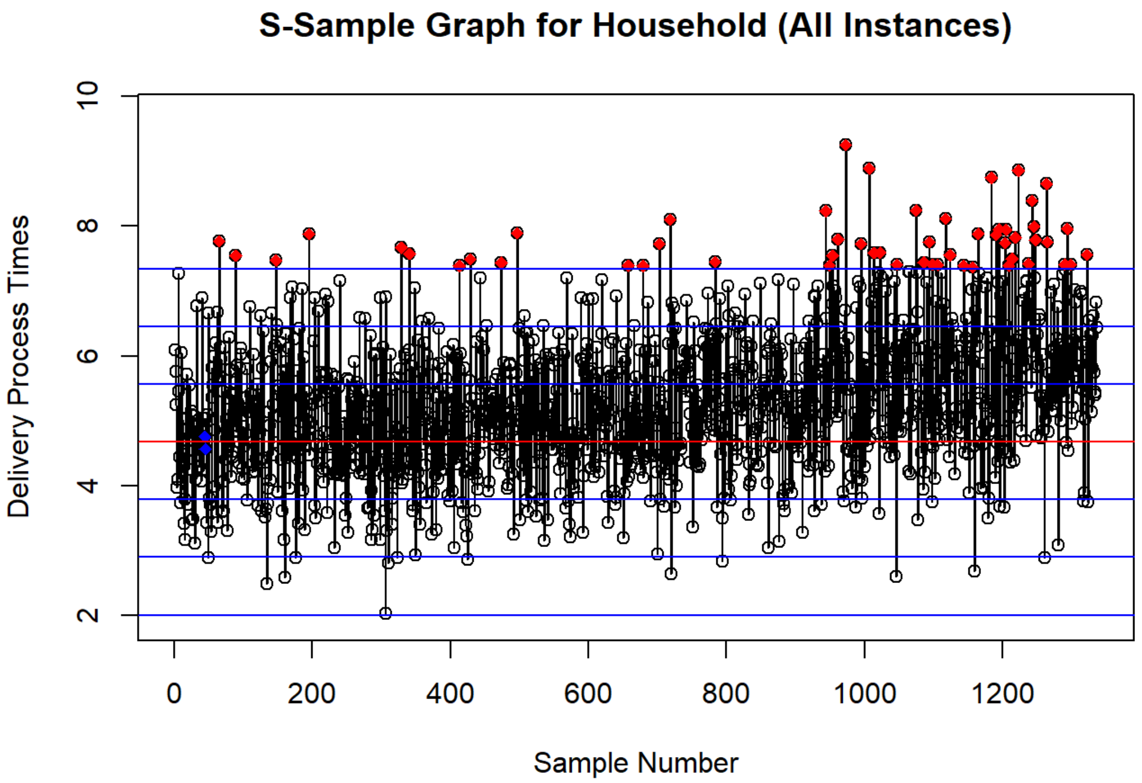

Figure 24: S-Sample Graph for all Household Items (Source: Anesu Jahura)

The s-sample graph shows us that the delivery process times for household items are mostly within the control limits, but there is a straight band of samples above the UCL throughout the process. This means that is a constant presence of variation in the delivery process times for household items, and the company should reduce this variation in order to achieve greater control over the delivery process. The out-of-control samples start to increase after about 1000 samples, indicating that the process is becoming even more variable – this issue should be addressed with urgency to stop it before its effects worsen.

5.2 Consecutive Samples

The number of consecutive s samples between -0.3 and 0.4 σ control limits was calculated in R, and the highest number of consecutive samples in each class was recorded. These results are given in Table 3:

Table 3: Consecutive s Samples Between Specified Limits Per Class (Source: Anesu Jahura)

Class of Item	Longest Nr Consecutive Samples	Ending Sample Number	Full List of Consecutive Samples
Sweets	3	94	92; 93; 94
Household	2	45	44;45
Gifts	4	254	251; 252; 253; 254
Technology	5	372	368; 369; 370; 371; 372
Luxury	3	63	61; 62; 63
Food	6	952	947; 948; 949; 950; 951; 952
Clothing	3	1013	1011; 1012; 1013

We can see from Table 3 that food items have the most consecutive number of samples between the specified limits, followed by technology items and gift items, respectively. This is an indication of a good run of consistency within the delivery processes. Clothing, luxury, and sweets items have less consecutive samples than the aforementioned items, but they still had a relatively consistent run of samples. Notably, the household items only had two consecutive items within the specified limits, which indicates that they are the least consistent class. This is yet another indication that the household item delivery process needs to be investigated urgently by management and solutions need to be created to address this problem.

5.3 Type I Error

In statistical analysis, a type I error occurs when we reject the null hypothesis when it is in fact true (Huecker & Shreffler, 2022). Using the normal distribution in R, the chance of making a type I error for the delivery processes was calculated to be 0.002699796, or 0.270 %. This means that we only have a 0.270 % chance of assuming that the delivery process is out of control when it is actually in control, so there isn't a high risk of making a type I error.

5.4 Centring Delivery Process for Best Profit

If we deliver technology items slower than 26 hours, there are various costs incurred to consider. Given the supplied cost parameters, a plot showing the cost vs hours to centre the deliver process was generated in R:

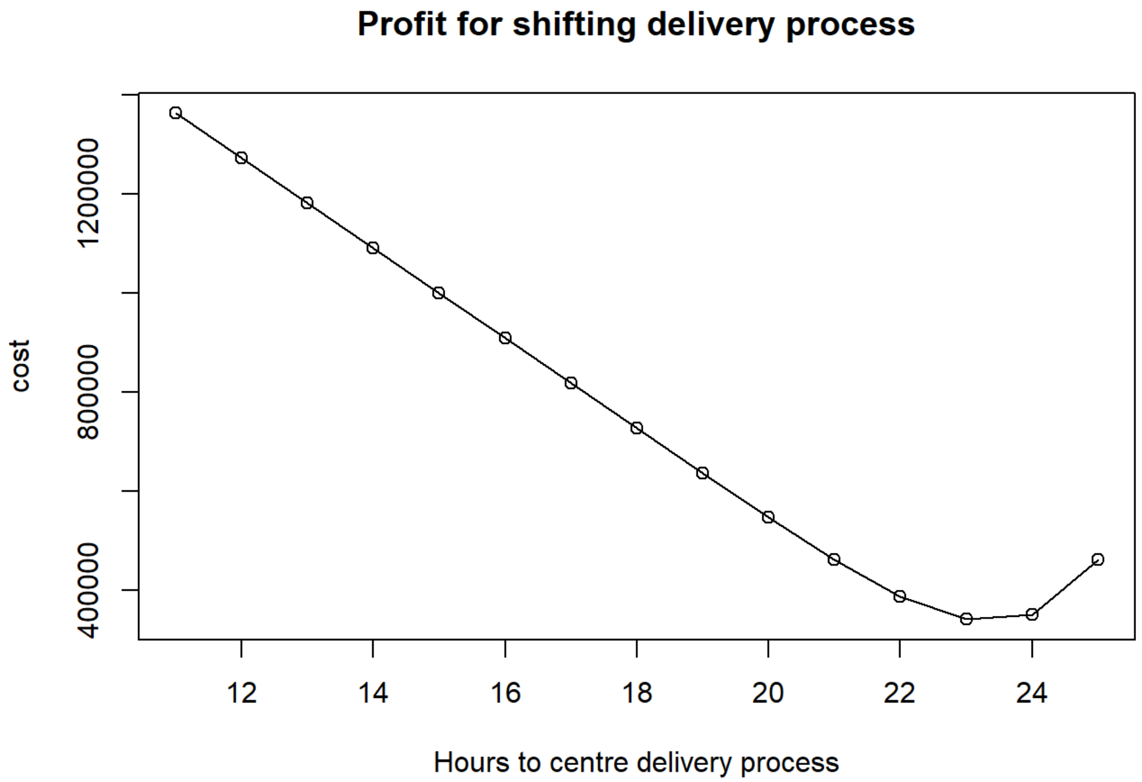

Figure 25: Profit for shifting the delivery process (Source: Anesu Jahura)

We can see from Figure 25 that the cost of shifting the delivery process decreases linearly until 23 hours are reached, where the cost begins to increase thereafter. Therefore, to achieve the highest profit the business must centre the deliver process at 23 hours. In this process, we can see that as we move away from the optimal 23 hours in either direction, the cost increases exponentially. This is similar to a Taguchi loss.

5.5 Type II Error

A type II error occurs when we fail to reject the null hypothesis when the alternative hypothesis is true. Using the normal distribution in R, along with the control limits for the "technology" class, the chance of making a type II error for the delivery process was calculated to be 0.4861932, or 48.62%. This means that there is a 48.62% chance of assuming that the technology delivery process is in control when it is in fact out of control. This is dangerous because there is currently a high chance that the business will fail to realise when its delivery process is not functioning well to satisfy customers, therefore, the manager needs to address this urgently.

6. Design of Experiments and MANOVA

6.1 Influence of Price & Age on Purchase Reasons

The MANOVA is a useful tool to test hypotheses about data. Looking at the information analysed in the previous sections of the report, the first null hypothesis that will be tested is: "the price and age of items do not have an effect on the reasons for purchasing". A MANOVA was calculated in R to test this hypothesis and the result is as follows:

Table 4: MANOVA Testing Price & Age vs Purchase Reasons (Source: Anesu Jahura)

Response	DF	∑ Squared Error	Mean Squared Error	F Value	Pr(>F)
Price	5	1.57421E+12	3.14842E+11	736.26	<2.2e-16
Age	5	106542	21308.4	51.33	<2.2e-16
Price*Age	5	3.75281E+15	7.50562E+14	599.54	<2.2e-16

Assuming a significance value of α = 0.001, we see that the p values calculated are much smaller than the significance value, meaning that we reject the null hypothesis. Therefore, we can conclude that price and age do, in fact, have an influence on the reason for purchasing. To confirm this conclusion with a graph, the interactions for each purchase reason can be plotted against the price and age:

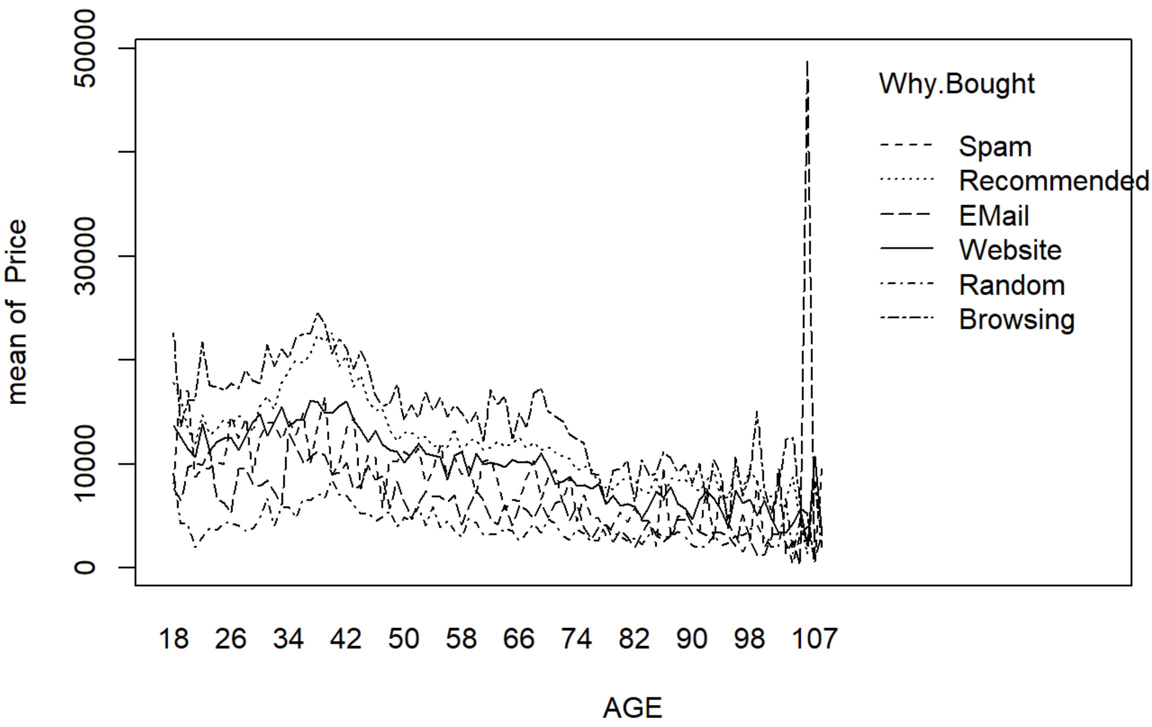

Figure 26: Purchase Reasons Interaction Plot for Age & Price (Source: Anesu Jahura)

We can only confirm that there is no interaction when the lines are parallel to each other. The lines are highly erratic with many steep slopes in alternating vertical directions, and they cross each other multiple times. Therefore, we can ascertain that there is a high amount of interaction, and we have to reject the null hypothesis. The business thus needs to consider the relationship between price and age and the reasons customers purchase items when they do advertising, and when they decide how to service different customers.

6.2 Influence of Price & Delivery Time on Class

The second null hypothesis that will be tested is: “the price and delivery time of items do not have an effect on the class of items”.

Table 5: MANOVA Testing Price & Delivery Time vs Class (Source: Anesu Jahura)

Response	DF	∑ Squared Error	Mean Squared Error	F Value	Pr(>F)
Delivery Time	6	33458565	5576427	629429	<2.2e-16
Price	6	5.71684E+13	9.52807E+12	80258	<2.2e-16
Delivery Time*Price	6	1.17012E+16	1.95021E+15	52792	<2.2e-16

Again assuming a significance value of α = 0.001, the p values calculated are much smaller than the significance value, so we reject the null hypothesis. We can, thus, conclude that delivery time and price influence the class of items. The interactions between each class of items against the delivery time and price can then be plotted:

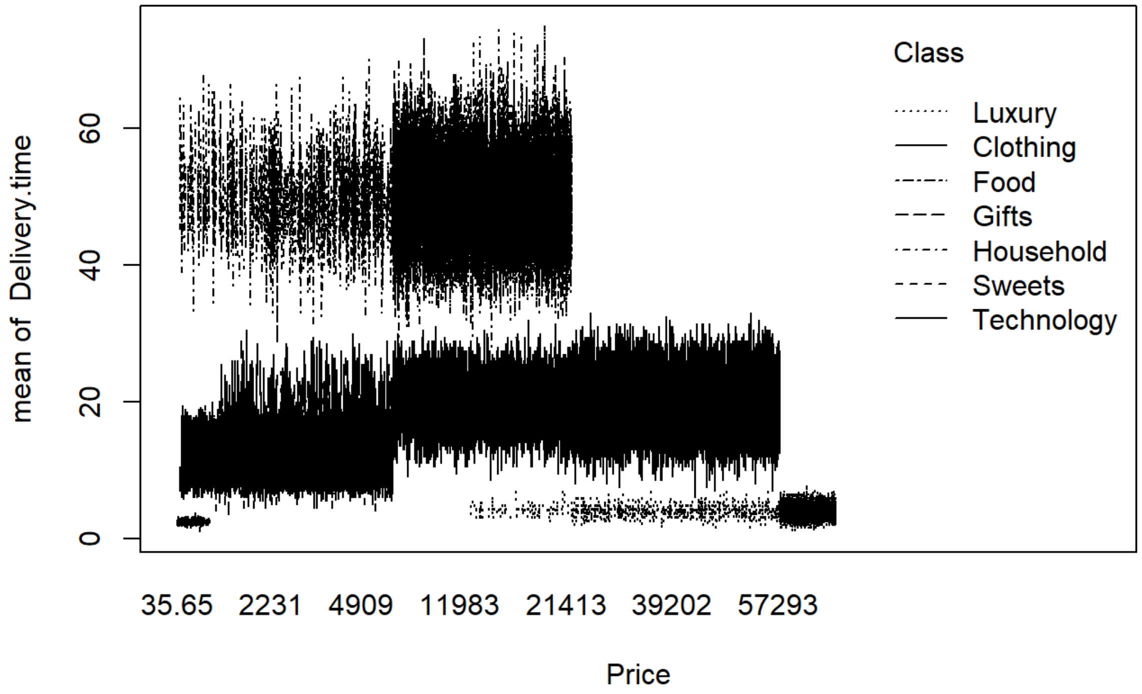

Figure 27: Class Interaction Plot for Price & Delivery Time (Source: Anesu Jahura)

The interactions plot once again affirms the results of the MANOVA. There are extremely steep slopes present in all of the lines, and the lines cross each other many times. This indicates a very high level of interaction. Therefore, delivery time and price have a very large influence on the class of items. The business thus has to ensure that its systems and processes are capable of meeting price and delivery time expectations for customers that purchase each class of items.

7. Reliability of the Service and Products

7.1 Taguchi Loss Function Analysis

The Taguchi loss function enables a visual display of how a marginal increase in the variation of specifications results in an exponential increase in the dissatisfaction of customers (Taguchi Loss Function, n.d.). The Taguchi loss function is modelled with the following exponential function:

$$L = k(y - m)^2 \qquad (9)$$

Where k represents the cost associated with some deviation from the target, y represents the observed value, m represents the target value, and thus, $(y - m)$ represents the deviation rom the target. After analysing the problem involving the thickness of a refrigerator part at Cool Food, Inc., the Taguchi loss function was calculating using R programming methods and found to be:

$$L = 28125(y - 0.06)^2 \tag{10}$$

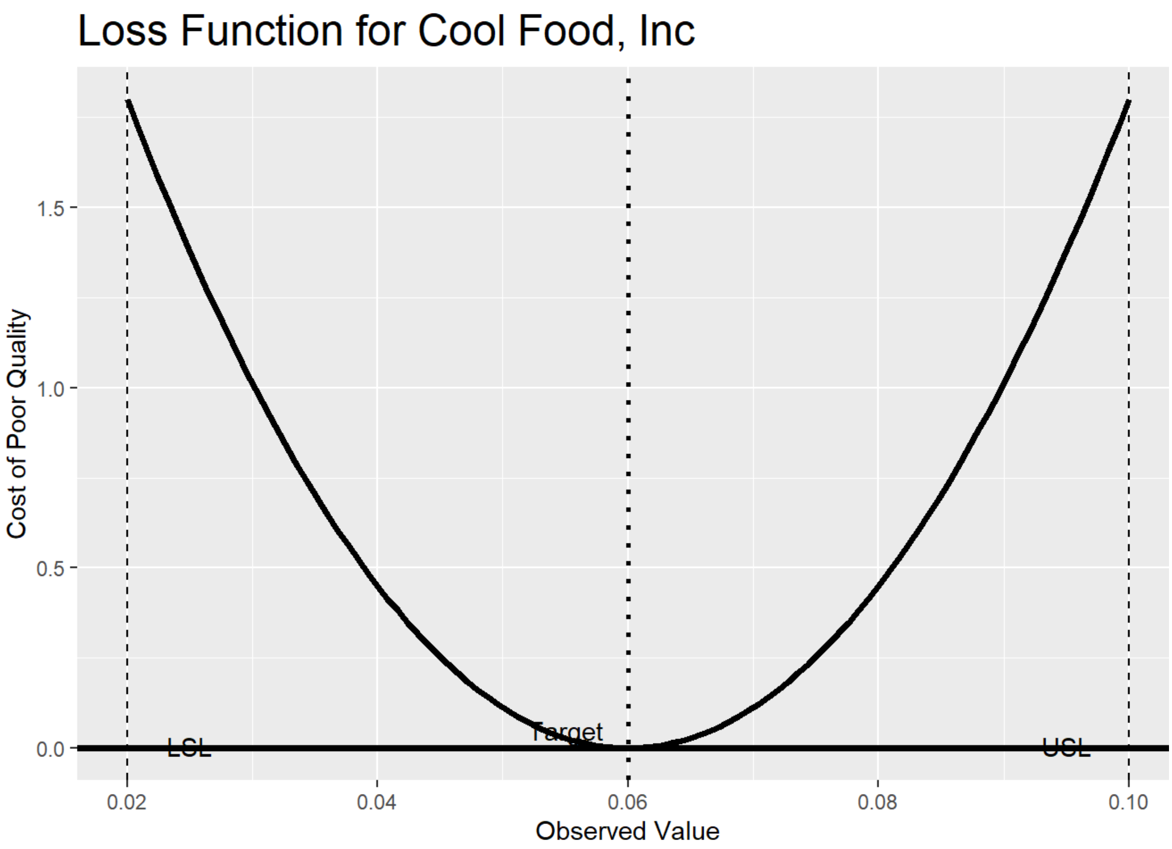

Figure 28: Taguchi Loss Function for Cool Foods (Source: Anesu Jahura)

The Taguchi Loss Function plotted in Figure 7 was plotted assuming that 10 000 parts in total were produced and that the sample measured has a size of 50 parts. From this plot, we see that as we start to deviate from the target value of 0.06 cm, the cost of poor quality – which is due to scrapped parts in this case – grows exponentially. As the observed value goes from 0.08 to 0.010 cm and from 0.04 to 0.02 cm, the growth in cost of poor quality is almost three times as much as in the range between 0.06 to 0.08 and 0.06 to 0.04 cm. Therefore, management

should aim to ensure that the thickness of the refrigerator part is kept within a healthy tolerance, preferably not deviating from the target value by more than 0.02 cm.
If the cost of scrap is reduced to $35 per part, the new Taguchi loss function is:

$$L = 21875(y - 0.06)^2 \tag{11}$$

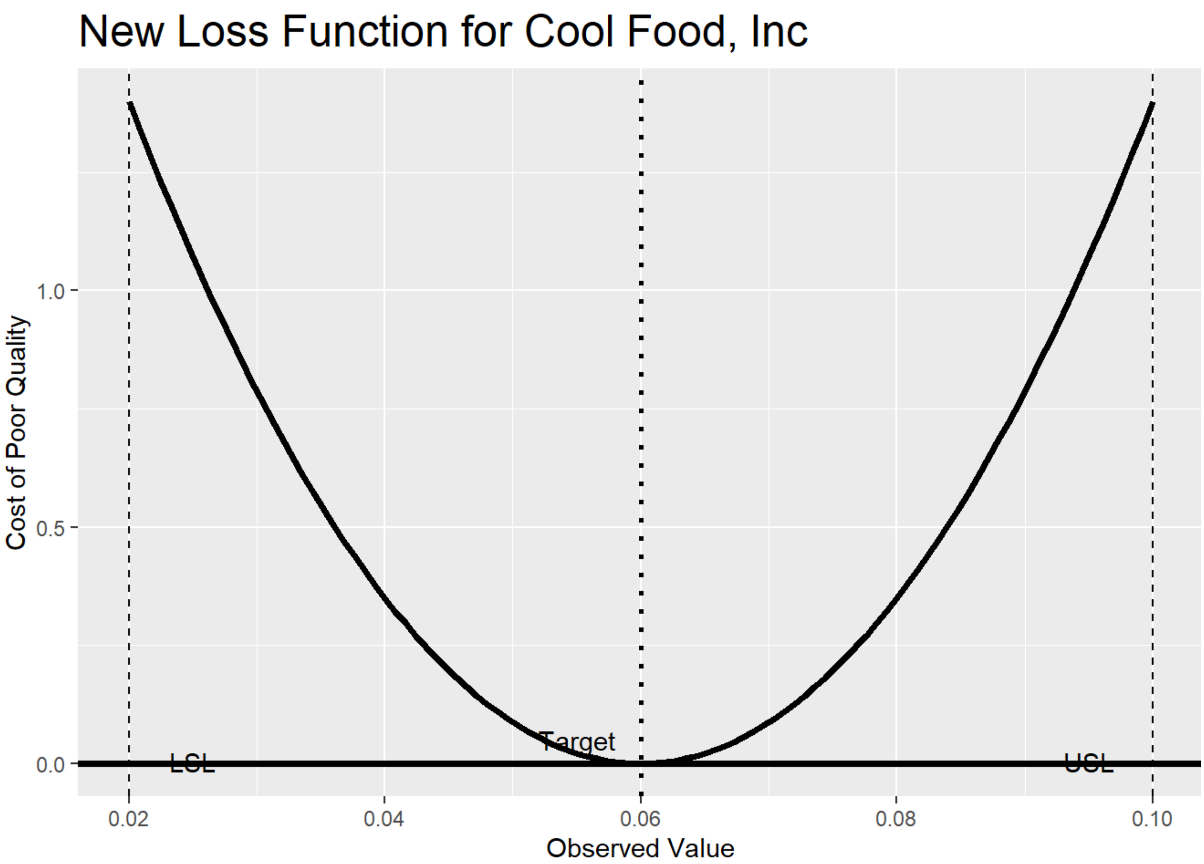

Figure 29: New Taguchi Loss Function for Cool Foods (Source: Anesu Jahura)

The new Taguchi loss function, plotted under the same assumptions of 10 000 parts produced and a sample size of 50 parts, shows us that the exponential increase in cost of poor quality will now cost the business less. Therefore, management should ensure that the cost of scrap is kept as low as possible to reduce the cost of poor quality.

When the process deviation from the target is reduced to 0.027 cm, i.e., $y - m$ = 0.027 cm, the new Taguchi loss, L, is 15.95.

7.2 Magnaplex Process Analysis

The manufacturing process at Magnaplex, Inc. consists of three processes connected in series, where if we assume only one machine at each stage, the reliability of the system is calculated using the following formula:

$$R_{system} = R_a \, R_b \, R_c \tag{12}$$

Where R_a = 0.85, R_b = 0.92, and R_c = 0.90 .

Using this formula and the given values, the system reliability is equal to 0.7038. This means that the system is only reliable 70.38 % of the time. This is not a good reliability, as it means that the business will lose 29.62 % of its production time due to breakdowns. Management needs to figure out how the reliability of the systems can be improved in order to maximise on available production time and increase profits.

If we assume three machines at each stage where two identical machines are connected in parallel at each process (a, b, and c), we can use the following formula, with the same reliability values used for equation (12), to calculate the reliability of each pair of machines:

$$R_{a,parallel} = R_a + R_a - R_a \, R_a \tag{13}$$

$$R_{b,parallel} = R_b + R_b - R_b \, R_b \tag{14}$$

$$R_{c,parallel} = R_c + R_c - R_c \, R_c \tag{15}$$

We then modify equation (12) to get the new reliability of the system:

$$R_{system} = R_{a,parallel} \, R_{b,parallel} \, R_{c,parallel} \tag{16}$$

This gives us a new system reliability of 0.9613156. This means that the system is now reliable 96.13% of the time, which represents an improvement of 25.77%. Therefore, management is strongly recommended to use two machines connected in parallel at each stage in order to ensure a high reliability and, in turn, higher profits.

7.3 Delivery Process Reliability

The functions created in R to calculate binomial probabilities allows the delivery process to be analysed in terms of reliability.

Considering that at least 19/21 vehicles need to be available, as well as the historical reliability information provided, the probability that there is zero vehicles unavailable (i.e., all the vehicles are available) is 0.861511, or 86.15%. This is a good indication because it means that the business will be able to achieve a high service level for its customers by delivering products on time. The number of days per year that there is zero vehicles unavailable is 314.46. It is reasonable to expect that there will be a few days where at least one vehicle will be unavailable (13.85 % in this case). However, this number can still be improved. Management needs to investigate exactly what causes vehicles to be unavailable and implement solutions to solve these issues. For example, if vehicles keep breaking down due to mechanical failure, then the business should use different mechanics and car service companies to increase the reliability of their cars.

The probability that there is one vehicle unavailable is 0.1288543, or 12.89%. The total reliability is thus 0.990363, or 99.04%. This means that 99.04% of the time, the business is meeting the requirements of having 19 or more vehicles available. The total reliability probability for vehicles is 0.9903954, or 99.04%. Thus, the number of days the delivery process is reliable is 361.49. This is very good because it shows strong control over service reliability in terms of vehicles. The business is, therefore, doing well in managing their vehicles.

In analysing the historical driver information, it is found that the probability of having zero drivers unavailable (i.e., all the drivers are available) is 0.9344269, or 93.44%. The number of days that zero drivers are unavailable is 341.068. This is a good indication because it shows that for most days of the year, all the drivers are available to work. However, management can still investigate ways to further improve the reliability of drivers.

The probability of one driver being unavailable is 0.06347701, or 6.35%. The number of days that one driver is unavailable is 23.17. This is also a good indication because the business's service reliability will not be adversely affected by unavailable drivers for a large portion of the year. The total reliability for drivers is 0.9979039, or 99.79%. The number of days that drivers are reliable are 364.2349. This is very good, and it shows strong staff management and high service reliability. Despite it being hard to achieve practically, management should attempt to have all drivers available for all days of the year.

The total reliability of the delivery process is 0.9883195, or 98.8315 %. The total number of days per year that the delivery process is reliable is 360.74 days. This is a very good figure which indicates that there are only about 4 days per year where the delivery process won't meet the required operational levels. In order to increase this number for the delivery process to meet the requirements for all days of the year, the business should aim to maximise its vehicle and staff management to ensure that all the vehicles are available and in good working condition and that all the drivers can work every day.

If the number of vehicles are increased by one to 21 vehicles, we can expect the delivery process to be reliable for 364.06 days. This is a 3.32 day increase on the previous delivery process reliability of 360.74 days. Therefore, management should strongly consider increasing the number of vehicles to 21 vehicles due to the highly positive increase in the number of days the delivery process is reliable.

8. Conclusion

It has been identified that the delivery process for luxury items is performing above expectations, and the business should take advantage of this to boost their profits. The business should also promote bulk purchases for sweets by offering discounts to bulk buyers. The consistent increase in delivery time over time for gifts, as well as the relatively low average price of items in the gifts class, is a major concern – the manager needs to work on improving the performance of the sales and delivery processes for gift items.

The process capability of the delivery process for the technology class also needs to be investigated and fixed as it is currently incapable and this is very bad for customer satisfaction and for general business performance. This class is also susceptible to a high chance of a type II error, so management as well as workers on the production floor need to be aware of this to ensure that they do not suffer greatly from this type of error.

The first MANOVA confirmed that price and age have effects on the reasons for purchase, and the second MANOVA confirmed that delivery time and price have large influences on the class of items. The business needs to investigate how they can use the results of the MANOVA to their advantage to grow their customer base.

The Taguchi loss analysis has revealed the costs of not meeting specifications for refrigerator parts, and management as well as the production staff should be aware of this.

The reliability analysis showed the various reliabilities for different system configurations, and designers of work processes should let this aid them when deciding on machine allocations.

Lastly, the delivery process reliability has revealed that vehicles and staff are mostly reliable for a large portion of the year, but there is still room for improvement.

9. References

Evans, J.R. & Lindsday, W.M. 2020. *Managing for Quality and Performance Excellence*. 11th edition. Boston: Cengage.

Huecker, M.R & Shreffler, J. 2022. *Type I and Type II Errors and Statistical Power*. [Online]. Available: Type I and Type II Errors and Statistical Power - StatPearls - NCBI Bookshelf (nih.gov) [15 October 2022].

McNeese, B. n.d. *Interpreting Control Charts*. [Online]. Available: Interpreting Control Charts (spcforexcel.com) [7 October 2022].

Process Capability Analysis. n.d. [Online]. Available: A Guide to Process Capability (Cp, Cpk) and Process Performance (Pp, Ppk) | 1factory [5 October 2022].

Taguchi Loss Function. n.d. [Online]. Available: Taguchi Loss Function – Lean Manufacturing and Six Sigma Definitions (leansixsigmadefinition.com) [17 October 2022].

Stellenbosch Faculty of Engineering, 2022. *Engineering Academic programmes and faculty information.* [Online]. Available: https://www.sun.ac.za/english/Documents/Yearbooks/Current/2021-Engineering.pdf [22 September 2022].

van Schalkwyk, T.G.D. 2022. *Statistical Methods in Quality Assurance Part 1 Summary* [Online]. Available: https://learn.sun.ac.za/mod/resource/view.php?id=1999132 [3 October 2022]