QUALITY ASSURANCE 344

ECSA Project Report

GI Potgieter – 23766999

Department of Industrial Engineering

Stellenbosch University

17 October 2022

Abstract

This report goes into extensive analysis and data understanding on sales of a company. This includes at the distributions of the prices and ages of the products and customers. The whole goal is to find ways to in which the company can improve their delivery process by reducing time and costs. Various charts are used to determine if the delivery process is stable and in-control. MANOVA testing is used to see if all the data of the sales are of statistical importance and to find some useful relationships. This data understanding and further analysis will ultimately guide the future decision-making of the company.

Table of Contents

Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

1. Introduction

A lot can be done with data and aid in the decision-making process. Trends and patterns are difficult to identify with the raw data and this is why descriptive statistics and various graphs and charts can make this a lot easier. Relationships between the delivery times and the classes can be identified to help improve the delivery process. Finally, iterations can be utilised to determine minimal costs lost and reduce delivery times. It is very important to be aware of the results and use them to the business's advantage.

2. Part 1

2.1. Data Wrangling

Before the data can manipulated to find trends and patterns the data first needs to be processed. The missing values within the dataset that need to be removed in order to make the dataset complete. All the instances that contain missing values are stored in the invalid dataset and the instances that are complete are stored in the valid dataset. All the instances still have their original primary keys which is why a new column is added with new primary keys that starts at 1 and continues consecutively to the number of rows. The data consists of 11 features regarding the sale of products. The features are: the old and new primary keys; ID; age; class; price; year; month; day; delivery time; and why bought. These features can help understand the market to improve the customer experience. It can also help in terms of where the company needs to focus to maximise profit.

The invalid data contains 17 instances where all the price values are missing. The valid dataset is complete with no missing values and contains 179 983 instances. This dataset will be used to find trends and patterns for the sales of the company.

Finally, before the data can be analysed, it needs to be in chronological order. This will help in seeing trends over time. This will be stored as an ordered dataset.

3. Part 2

3.1. Data Understanding

Using descriptive statistics and various graphs, a lot can be analysed from the data. It is very important to understand the data in depth before analysing the data, because general insights will be very helpful to guide the analysis. Afterall, useful information can be gathered from the valid data that contains information about 179 983 purchases and help with future decision-making processes and within the company.

3.2. Process Capability Indices

There are four process capability indices that are calculated on the delivery times of all the technology class sales. These indices show the variation that the delivery times are experiencing compared to the company's capability and whether the delivery times for technology products are in specification. The upper (USL) and lower (LSL) specification limits are set to 24 and 0 hours respectfully. The company would like to deliver all the technology products within 1 day, which is why the USL is 24 hours. The company should always aim to decrease the delivery time, which is why the LSL is set at 0 hours. This implies that that a delivery time for technology products is in specification when it is less than 24 hours. As seen in figure 1, the delivery time for technology products is approximately normally distributed. The graph needs to be shifted to the left – meaning that the delivery time for technology products need to be decreased to meet the specifications.

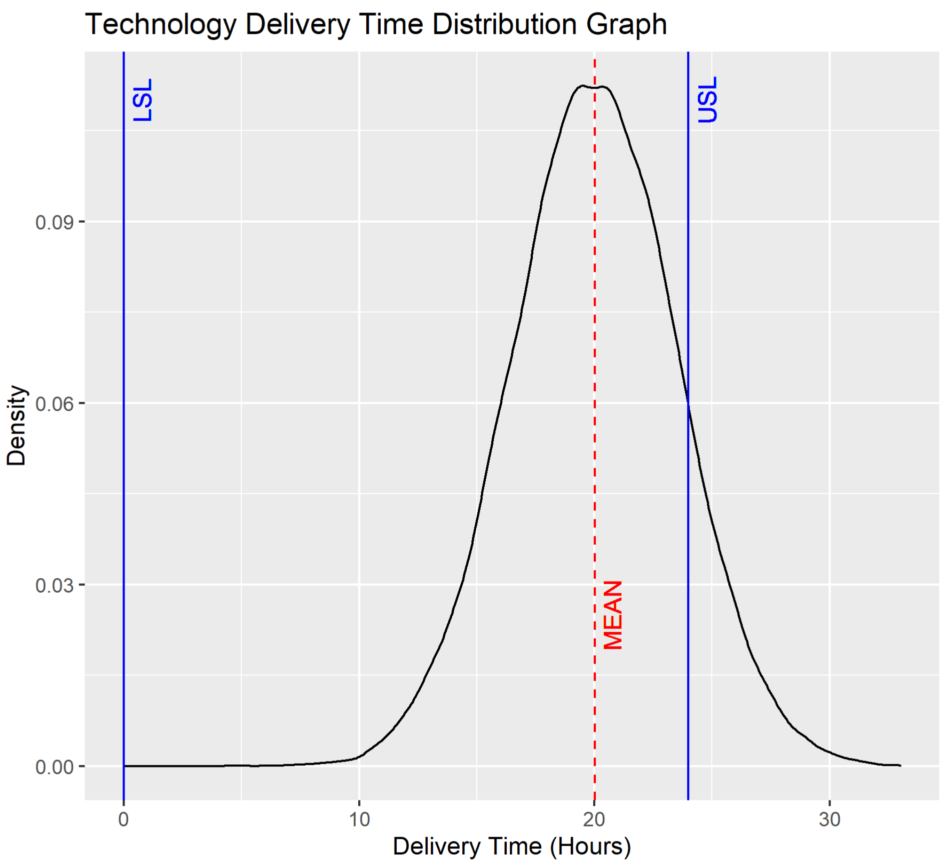


Figure 1: *Graph showing the delivery time of technology class products distribution.*

Capacity Potential			
Cp	Cpu	Cpl	Cpk
1.1422	0.3797	1.9047	0.3797

Figure 2: *Table showing capacity potential values.*

The Cp values as seen in figure 2 indicates how well the delivery times are withing in the specification range and the potential the company has to do so. This means that the Cp value of 1.142 is recommended. The Cpu shows what the delivery time capability is relative to the upper limit. A Cpu value of 0.380 is low meaning that the delivery times still needs improvements. The Cpl on the other hand shows the what the delivery time capability is relative to the lower limit. The Cpu and Cpl need to be calculated to be able to calculate the Cpk, which is the minimum value between the two. A Cpu value close to 0 means that the mean is close to the USL. In this case the Cpu has a value of 0.380 indicating that the delivery times are higher than the specifications (Cp > Cpu). This implies that the company's delivery process needs to be improved for technology products to be able to meet the specifications.

3.3. Price

Price Statistics									
	Count	Min (R)	Q1 (R)	Median (R)	Q3 (R)	Max (R)	Mean (R)	SD (R)	Total (R)
Price	179 983	-588.80	482.31	2 259.63	15 270.74	116 618.97	12 293.74	20 888.97	2 212 664 291.73

Figure 3: *Table showing price statistics.*

As mentioned before and as seen in figure 3 there were 179 983 sales during the course of 9 years. This totalled by to over R2.2 billion. The minimum sale was -R588-80 which indicates that there were refunds during this time. After examining the quartiles, they are significantly lower than the max and the mean is much larger than the median. The mean is also closer to the first quartile which indicates that the price might be skewed to the right. The standard deviation is also high with a value of R20 888-97 indicating that there is a large variation in the price of goods sold.

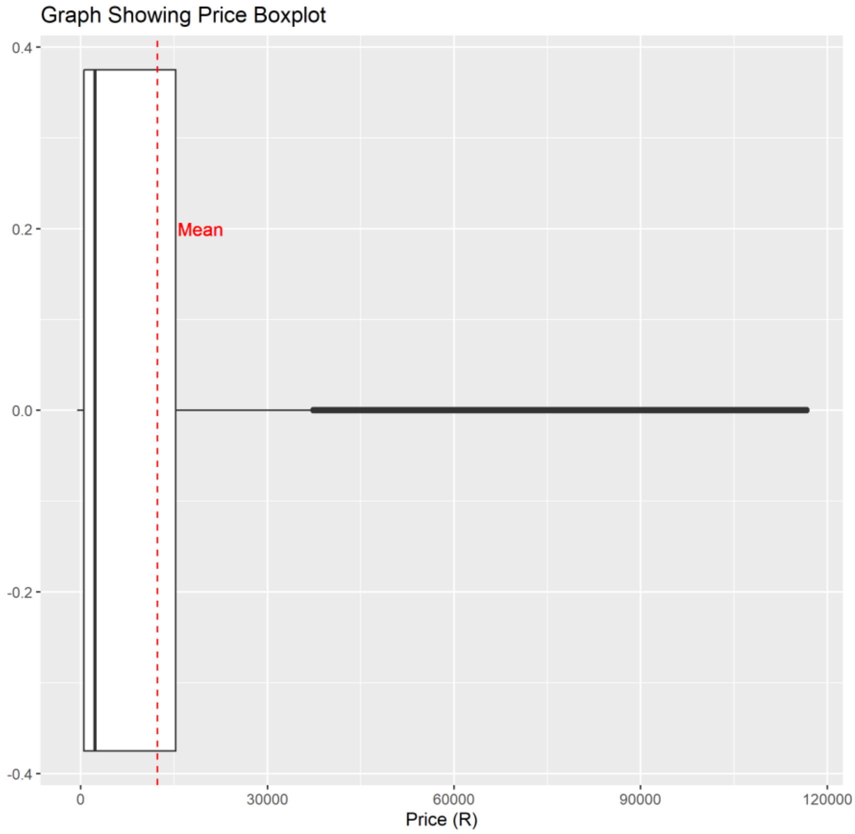


Figure 4: *Graph showing the density distribution of the price of goods sold.*

Looking at figure 4 it is clear that the price of all the sales is skewed to the right because of the long tail towards the higher prices. The dotted line is the mean price. This signifies that most of the prices are relatively low with a few outlying prices that are high. This can be a result of the various types of products that are sold each with a different demand.

3.4. Class

Class Statistics									
Class	Count	Min Price (R)	Q1 Price (R)	Median Price (R)	Q3 Price (R)	Max Price (R)	Mean Price (R)	SD Price (R)	Sum Price (R)
Clothing	26 403	127.76	384.44	642.04	895.56	1 154.02	640.53	296.41	16 911 790.75
Food	24 582	127.76	266.69	408.37	548.37	691.96	407.82	162.94	10 024 915.06
Gifts	39 149	172.61	1 575.50	2 961.59	4 346.77	5 774.49	2 961.84	1 611.12	115 953 129.71
Household	20 065	127.76	5 634.88	10 960.88	16 439.59	21 935.33	11 009.27	6 263.08	220 901 078.50
Luxury	11 868	12 825.37	38 838.52	65 342.14	90 667.15	116 618.97	64 862.64	30 077.55	769 789 795.33
Sweets	21 564	35.65	168.13	303.25	439.64	576.38	304.07	156.53	6 556 973.54
Technology	36 347	935.18	15 340.14	29 653.90	43 651.93	57 735.40	29 508.06	16 368.40	1 072 529 552.84

Figure 5: *Table showing price statistics for each of the various product classes.*

Figure 5 will help understand how the different product classes that are sold influence the price. In this table all the negative price values have been removed, which is a total of 5 sales, therefore this table includes the 179 978 other transactions. Gifts has the highest count seeing as this product type was purchased the most, but has a relatively low mean price, hence has the 4th largest total price value. Technology on the other hand has the 2nd most purchase frequency but has a larger mean price which is why it contributes the most to the total sales price. Luxury items has the lowest purchasing count, nonetheless it has the 2nd highest total price because luxury items are so expensive as seen by the mean price and the maximum price. The standard deviation for luxury items is high and this can be attributable to the wide range of products that can be classified as luxurious. Food and sweets are both the cheapest in comparison which is why they contribute the least towards the total price. Food and sweets do not have a large variety in price and this can be confirmed by the lower standard deviation. All of the class has a mean price relatively close to their respective median price, which indicates that all the classes are possible well distributed.

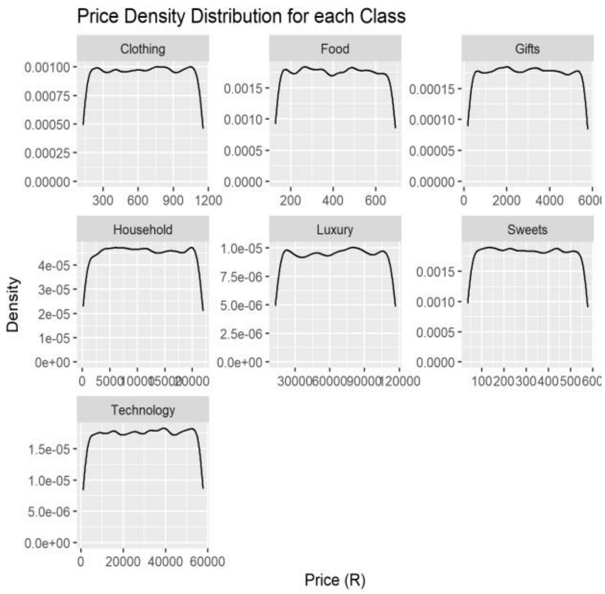


Figure 6: *Graph showing the density distribution of price for all classes.*

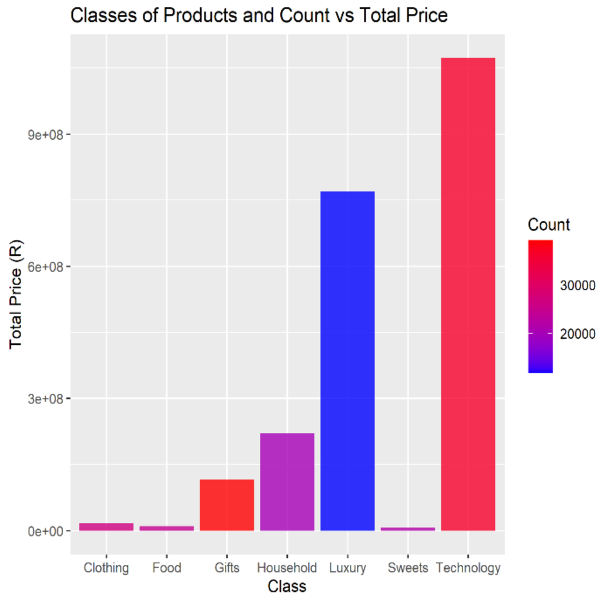


Figure 7: *Graph showing the total price of all classes and comparing counts.*

Figure 6 confirms what was discussed earlier about the price distribution of each product class – the distributions are similar and the graph indicates that it is approximately a uniform distribution. The x-axis is not the same for each distribution, suggesting that the prices are not the same as seen in figure 7. Looking at figure 7 the total price for each class is significantly different. This is the reason why the total price distribution is skewed towards the right.

3.5. Delivery Time

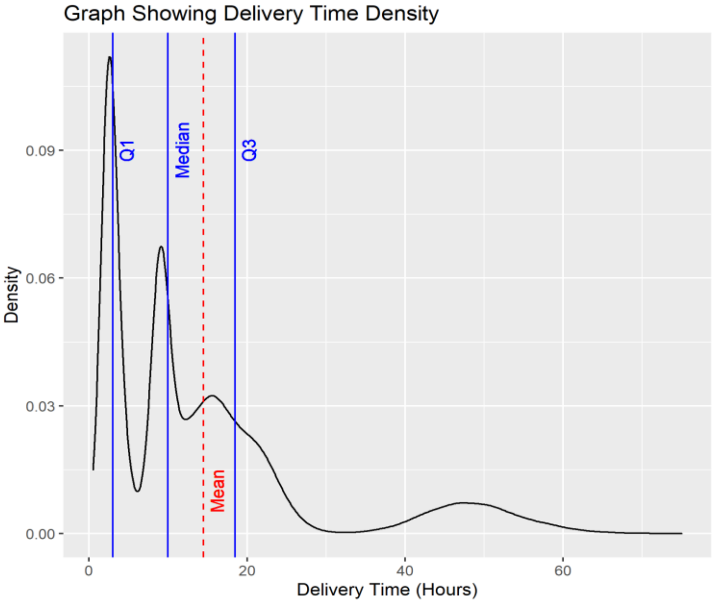


Figure 8: *Graph showing the density distribution of the delivery time.*

Figure 8 shows that the delivery time is trimodal due to the three peaks in the graph. There is also a tail towards the right of the graph as the delivery time increases. This indicates that there are outlying delivery times that are longer than what is expected. This is considered as bad service from a customer's perspective. The company will have to look at the delivering process and make some improvements as seen with the capacity potential above. The trinomial distribution shows that there are some inconsistencies but this might be due to the fact that the company has a variety of different products from which customers can choose from. The delivery times for the different products may differ depending on the size and suppliers of the product. There are three high probabilities for the delivery time, the first and the highest probability being on the first quantile, the second largest probability is on the median of the delivery times, and lastly the third largest probability is near the mean of the delivery times.

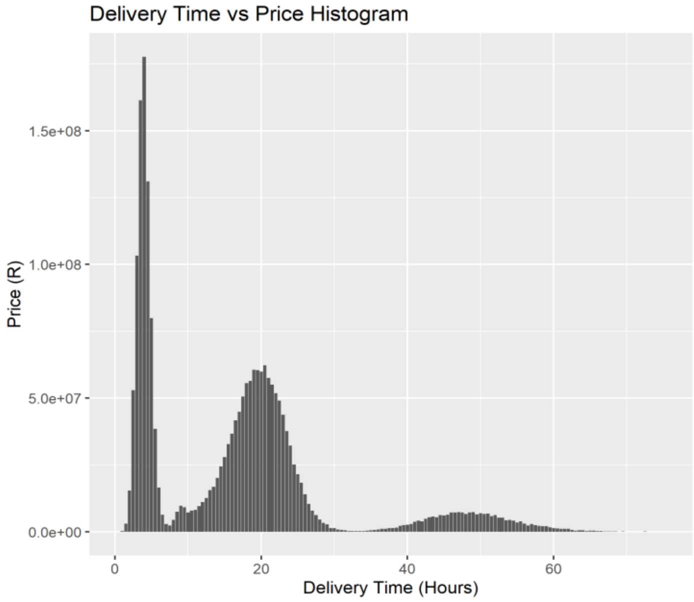


Figure 9: *Graph showing the delivery time vs price of goods sold histogram.*

Figure 9 also has 3 peaks showing that the delivery time influences the price of the product and that the distribution is trinomial. Based off this graph customers tend to pay more for faster delivery, but there seem to be more factors influencing this and more analysis is required.

Delivery Time vs Class Statistics									
Class	Count	Min Delivery Time (Hours)	Q1 Delivery Time (Hours)	Median Delivery Time (Hours)	Q3 Delivery Time (Hours)	Max Delivery Time (Hours)	Mean Delivery Time (Hours)	SD Delivery Time (Hours)	Total Delivery Time (Hours)
Clothing	26 403	6.50	8.50	9.00	9.50	11.50	9.00	0.62	237 614.50
Food	24 583	1.50	2.50	2.50	2.50	3.50	2.50	0.29	61 507.00
Gifts	39 149	3.50	10.50	13.00	15.00	22.00	12.89	2.96	504 652.00
Household	20 067	27.50	44.50	48.50	52.50	75.00	48.72	6.23	977 651.50
Luxury	11 869	1.00	3.50	4.00	4.50	8.00	3.97	0.90	47 137.50
Sweets	21 565	0.50	2.00	2.50	3.00	4.50	2.50	0.52	53 939.00
Technology	36 347	4.50	17.50	20.00	22.50	33.00	20.01	3.50	727 338.00

Figure 10: *Table showing delivery times statistics for each class.*

Figure 10 illustrates the delivery time for each of the product classes. Before looking at density distribution of the delivery times, the standard deviation needs to be examined to see the variation. This will give a better understanding where the company can improve the delivery process to be more consistent and offer the same customer satisfaction to all. There are 3 class of products’ standard deviations that need to be investigated further – Gifts, Household and Technology. These 3 classes are the most inconsistent regarding their delivery time. Household as the largest standard deviation, but this might still be as acceptable seeing as customers usually wait more than 2 days, 6 hours extra might not be too much longer. The delivery process for technology product needs to improve as discussed earlier. The main concern seems to be gifts – there is a possibility that the customer can wait up to 23% longer than expected. Considering the mean delivery time and the class type it is understandable that there is a significant difference. Food and sweets are expected to have the lowest mean seeing as the product has a very short shelf life. The food needs to be served warm and the ice cream needs to be delivered before it melts. The luxury items are generally more expensive, hence faster delivery times are expected. Clothing has a longer shelf life but fast delivery will increase customer satisfaction, hence it is lightly longer than the food and sweets' delivery time. Gifts tend to have a long shelf life, hence the delivery time of these products do not influence the quality of product experience but can influence the customer's experience with the shop. That is fast delivery is not essential but keeping the customer happy is. Lastly, household items have the longest mean delivery time. These items are usually larger and bulky with a very long lifespan. Fast delivery time is not as important and is usually subject to availability of the appropriate vehicles.

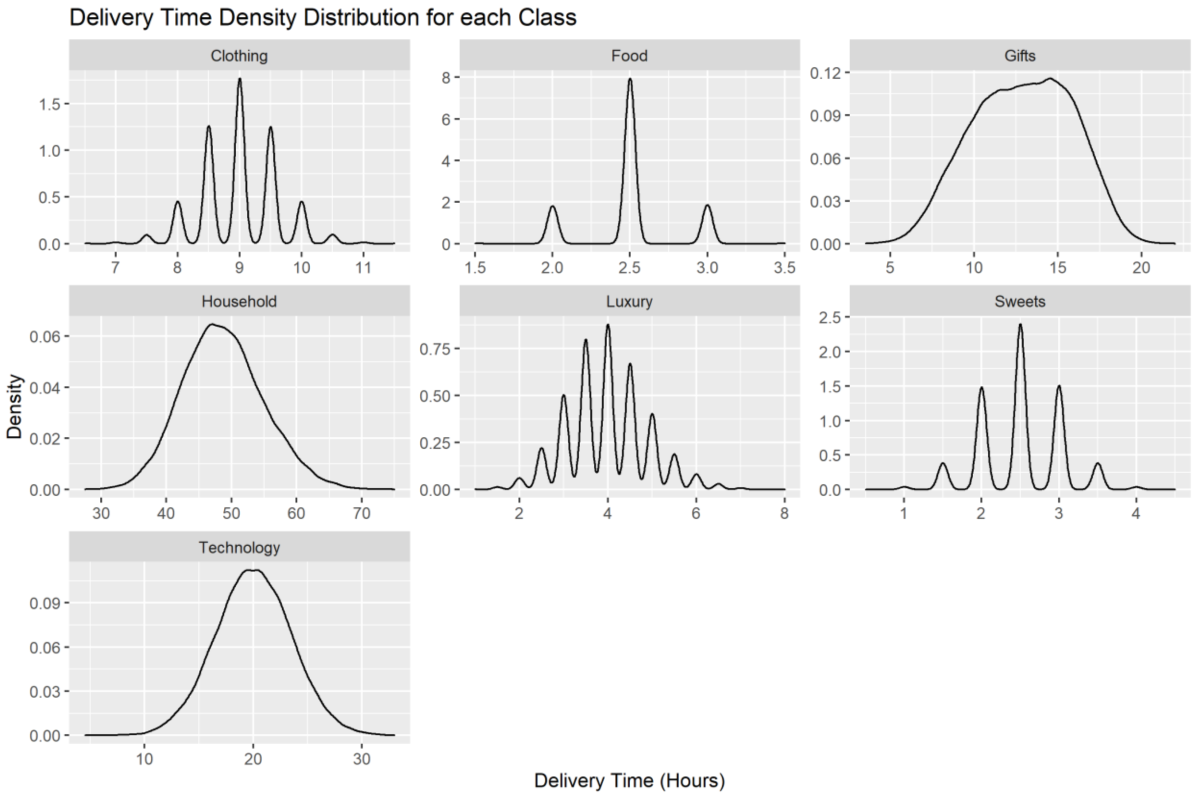


Figure 11: *Graphs showing the delivery time density distribution for each product class.*

The delivery time density distributions are examined in figure 11. It is interesting to note that the 3 classes mentioned with the largest standard deviation are all somewhat normally distributed. Clothing, Food, Luxury and Sweets seems to be delivered in intervals. For example, there are mainly 3 waiting times for food deliveries. The most likely is 2.5 hours and equally less likely is 2 and 3 hours. The other classes have more intervals but follows the same principal. Luxury items have the most intervals (11 intervals) where a delivering time of 4 hours has the highest probability. Sweets has 7 intervals with majority take 2.5 hours with the second highest probability of 3 and 4 hours being equally less regarding the delivery times. Lastly, clothing has 7 intervals where the probability of delivery times is significant with the highest at 8 hours and second highest being 0.5 hours on either side.

3.6. Why Bought

Why Bought Statistics									
Why Bought	Count	Min Price (R)	Q1 Price (R)	Median Price (R)	Q3 Price (R)	Max Price (R)	Mean Price (R)	SD Price (R)	Sum Price (R)
Browsing	18 994	35.65	589.86	4 294.10	26 337.50	116 479.24	16 130.56	22 486.13	306 383 885.06
EMail	7 224	35.77	432.80	1 110.78	5 110.95	114 430.06	6 662.08	13 351.89	48 126 832.60
Random	13 120	35.65	399.59	913.97	4 125.57	116 099.40	4 288.63	9 570.98	56 266 859.59
Recommended	106 985	35.65	473.34	1 972.02	16 974.34	116 618.97	13 440.93	22 762.44	1 437 978 192.84
Spam	4 208	35.65	467.36	2 034.06	11 196.95	107 883.57	9 360.90	15 364.74	39 390 669.23
Website	29 447	35.65	533.48	3 374.67	14 947.53	116 550.89	11 020.50	16 932.59	324 520 796.41

Figure 12: *Table showing the price statistics for each why bought category.*

The same as before, all of the negative price values have been removed in order to create the figure 12. This table will help find the most effective marketing plans to launch in order to maximise revenue. Most customers purchased products because it was recommended to them. This is by delivering an excellent customer experience that is spread by word of mouth, hence it is no surprise that recommended has the highest total price. Inversely email, spam and random has the lowest count and the lowest mean, therefore having the lowest total price. This is a cheap marketing technique but proves to be ineffective. Browsing has the highest mean price, indicating that browsing might be very effective, but it also has the second highest standard deviation meaning it is not a guarantee. Websites are also a good option seeing as it has the second largest count, third largest mean price and second largest total price.

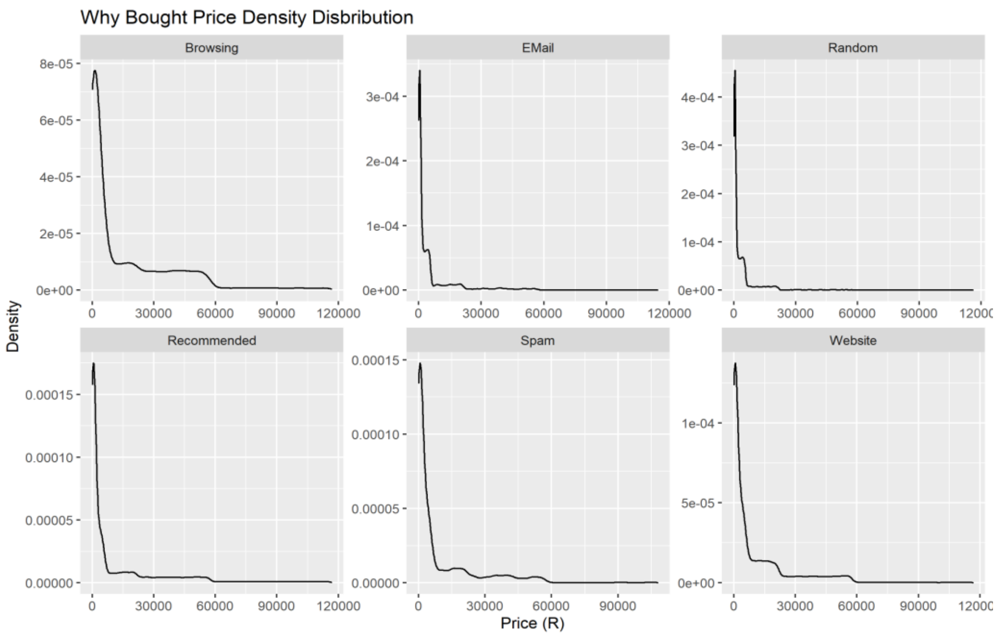


Figure 13: *Graphs showing the price distribution for each why bought.*

Figure 13 shows the price density distribution for each why bought category. This shows that the distribution for all why bought categories are very similar, being skewed towards the right. This means that pushing only one marketing technique will not have a major impact on the price distribution.

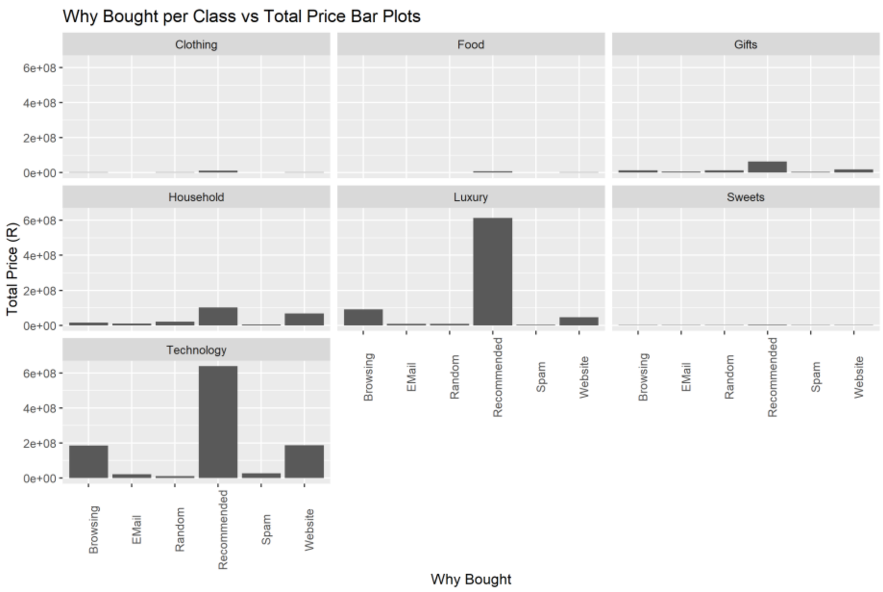


Figure 14: *Graphs showing the why bought for each class vs total price.*

There is however a better marketing technique for each class of product as seen in figure 14. It is clear that recommended is the best marketing technique for all classes, but a combination of techniques will prove to be more effective. The following are the second choices that need to be implemented along with recommended: clothing – websites; food – websites; gifts –

websites; household – websites; luxury – browsing; sweets – websites; and technology – browsing and websites.

3.7. Time

3.7.1. Year

Year Statistics									
Year	Count	Min Price (R)	Q1 Price (R)	Median Price (R)	Q3 Price (R)	Max Price (R)	Mean Price (R)	SD Price (R)	Sum Price (R)
2021	33 443	35.65	529.23	1 083.88	10 799.77	116 192.55	8 232.41	15 023.02	275 316 554.34
2022	15 546	35.65	483.75	2 485.84	15 830.44	116 130.33	12 599.25	21 132.28	195 867 865.30
2023	17 128	35.65	477.94	2 347.15	16 217.18	116 158.97	12 715.32	21 385.75	217 787 949.38
2024	17 698	35.77	475.87	2 711.35	17 665.90	116 332.73	13 406.74	21 938.89	237 272 483.74
2025	17 267	35.77	460.00	2 414.08	18 136.53	116 409.78	13 518.23	22 308.36	233 419 294.10
2026	17 152	35.77	471.73	2 551.39	17 185.37	116 388.85	13 329.08	22 009.43	228 620 403.95
2027	18 656	35.77	470.81	2 628.21	17 279.21	116 523.63	13 311.93	22 006.25	248 347 416.78
2028	20 613	35.77	481.50	2 620.85	17 290.59	116 515.35	13 247.45	21 863.87	273 069 747.33
2029	22 475	35.77	466.67	2 550.47	17 390.99	116 618.97	13 480.11	22 337.36	302 965 520.81

Figure 15: *Table showing price statistics for each year.*

Considering the price for each year is expected to grow due to inflation and as the company expands. Examining the figure 15 it does not follow the convention. The total price starts of very high in the first year and drops in the second. From there onwards it gradually climbs over the total price only in the last year. The price's standard deviation jumps from the first to the second year where it stays relatively stable for the rest of the years. The same goes for the mean price per year. This can be attributed to new products being put on the marker which increases the variation in price.

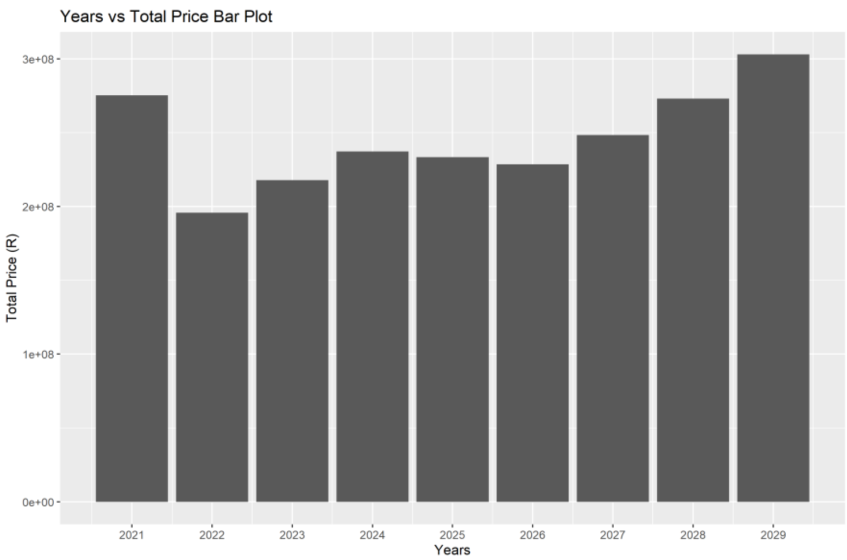


Figure 16: *Graph showing the total price obtained from each year.*

Here the total revenue from each year is visually shown in figure 16. As stated above, the total revenue drops from 2021 to 2022 and the steadily climbs where only in 2029 the total revenue is once again higher than as it was in 2021.

3.7.2. Month

Month Statistics									
Month	Count	Min Price (R)	Q1 Price (R)	Median Price (R)	Q3 Price (R)	Max Price (R)	Mean Price (R)	SD Price (R)	Sum Price (R)
1	14 799	35.65	474.84	2 305.63	15 136.82	116 523.63	12 425.37	21 280.65	183 883 119.50
2	14 905	35.65	477.94	2 105.30	15 125.26	116 550.89	12 178.36	20 814.55	181 518 392.61
3	15 038	35.65	487.60	2 342.55	15 706.78	116 204.74	12 531.76	21 255.46	188 452 543.99
4	14 920	35.77	474.49	2 248.13	15 562.38	116 233.14	12 442.53	21 070.12	185 642 515.66
5	14 865	35.65	482.08	2 217.20	14 530.59	115 801.32	11 951.83	20 510.40	177 663 994.00
6	15 144	35.65	492.97	2 295.86	15 493.29	116 281.67	12 380.76	21 065.04	187 494 291.30
7	14 918	35.65	486.13	2 320.13	15 426.73	116 517.31	12 258.23	20 717.11	182 868 267.26
8	14 859	35.65	486.22	2 262.97	15 573.76	116 549.62	12 569.87	21 379.69	186 775 677.88
9	15 038	35.65	481.73	2 160.97	15 073.13	116 515.35	12 086.68	20 536.97	181 759 431.86
10	15 221	35.77	487.14	2 356.35	15 365.84	116 532.37	12 361.94	20 827.29	188 161 042.58
11	15 046	35.65	478.23	2 193.74	14 259.57	116 618.97	11 902.79	20 287.11	179 089 445.45
12	15 225	35.65	481.50	2 337.95	15 668.40	116 394.83	12 437.34	20 897.18	189 358 513.64

Figure 17: *Table showing price statistics for each month.*

The goal of figure 17 is to see if there is any seasonality with the sales. Based off this table there is no seasonality related to the sales. There is little to no change within any of the months, but no conclusion can be made yet – further analysis will have to follow. The only class that might showcase seasonality is clothing.

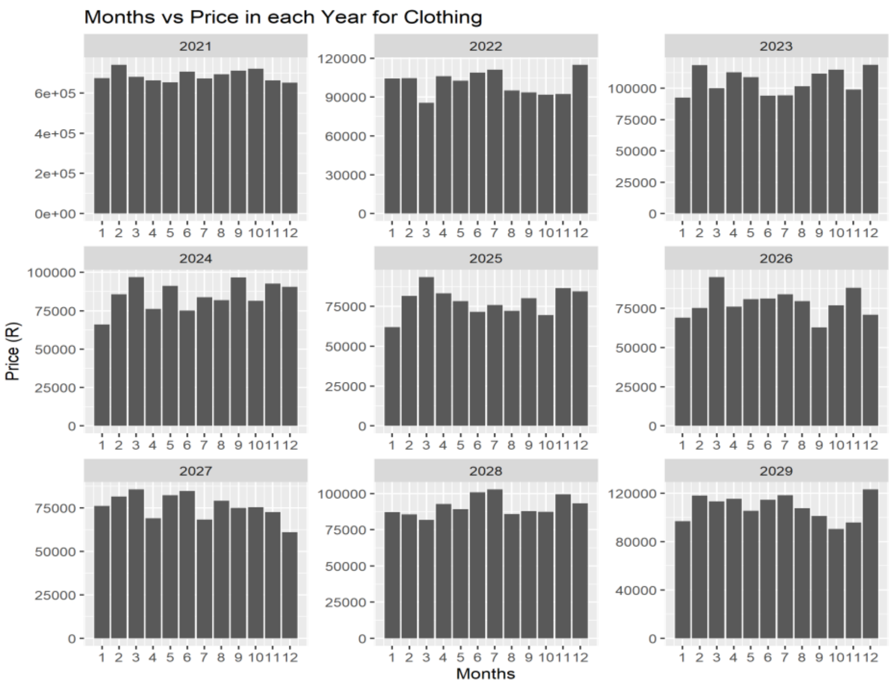


Figure 18: *Table showing the clothing prices for each month in all the years.*

When investigative figure 18, there does seem to be some seasonality with respect to the total revenue of the clothing class. This is to be expected as clothing usually has this trait. There is no definite trend regarding the wave like graphs for each year, but this still classifies as a seasonal item.

3.8. Age

The age of the target market will impact the type of product and the marketing technique used that the company wishes to sell. Theoretically a younger target market will be more susceptible to sweets and would be more drawn by visual marketing or recommendation.

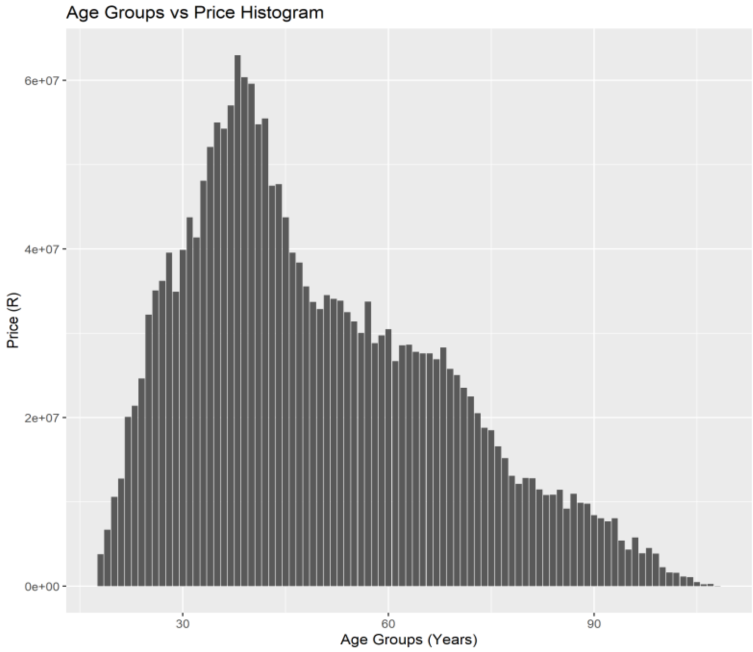


Figure 19: *Graph showing age groups vs price histogram.*

Figure 19 illustrates the distribution of the price over the age groups as a histogram. The price is approximately normally distributed and skewed to the right. This indicates that the most revenue comes from customers around the age of 40 years old. But the price distribution might be different depending on the class of product.

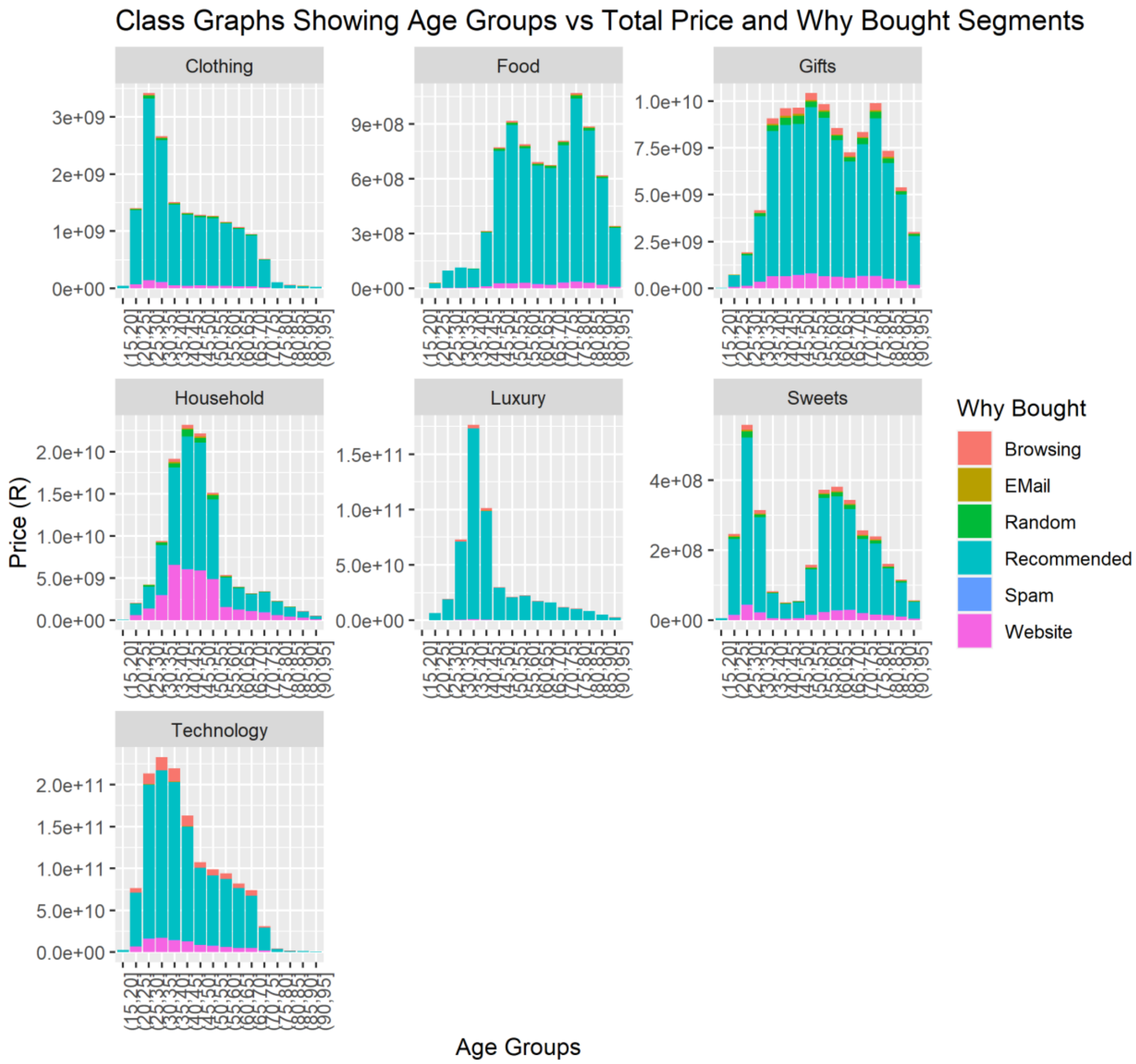


Figure 20: *Graph showing class bar plots of age groups vs price and reason for purchase.*

The distribution of revenue according to the age groups for the different classes can be better understood with figure 20. All of the distributions are unique. Food, gifts, and sweets are bimodal distributions where food is skewed to the left and sweets skewed to the right. This shows that there are 2 main age groups that contribute to the majority of the total price for these 2 products. For food the main two groups are ages 50-55 and 70-75 years. For sweets the main 2 age groups are 25-30 and 60-65 years. The household items also have two main age groups being 50-55 and 75-80 years of age. Clothing, household, luxury, and technology are approximately normally distributed and all are skewed towards the right. This means that the majority of the revenue for these classes come from the younger age groups.

4. Part 3

4.1. S & X-Bar Chart Initialisation

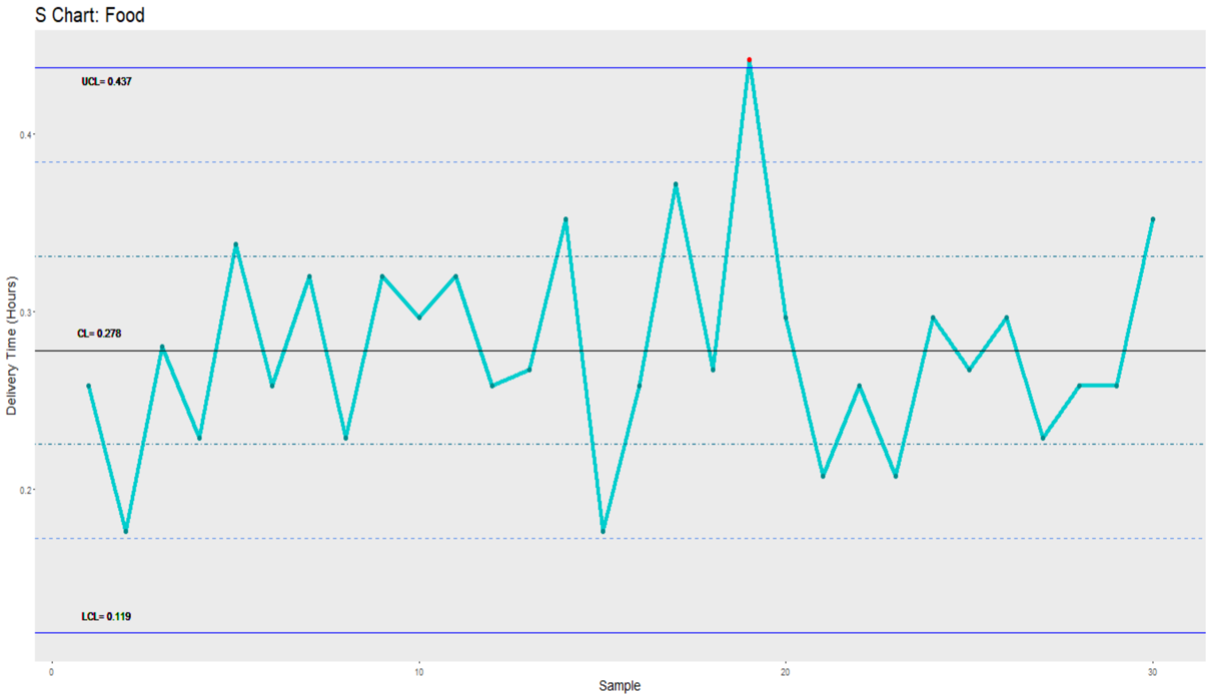


Figure 21: *Graph showing S Chart of initializing samples for food for the first 30 samples (15 sales per sample).*

This graph showed in figure 21 show the first 30 samples of sales, where there are 15 individual sales within a sample, that will be used to initialise the rest of the S Charts. The goal of this initialisation step is to find any out-of-control samples that are above the upper centre line (UCL) or below the (LCL) and remove them as this will have an impact on assessing the rest of the samples. In this graph there is a red point that exceeds the (UCL) and this sample 19 needs to be removed before calculating the critical values. The graph explores the mean standard deviation of the food sample delivery times and assesses whether the consistency of the delivery time is under control. The solid black line indicates the centre line (CL) where the target value for the mean standard deviation of the clothing delivery time is 0.278. The outer most lines are the control limits that are 3 standard deviations above (UCL) and below (LCL) the CL – the lines that the chart needs to stay within. The second most outer lines are 2 standard deviations above (U2Sigma) and below (L2Sigma) away from the CL. These are the warning limits and if the chart exceeds these limits action needs to be taken to ensure that the company does not loose even more control by the next sample exceeding the control limits. The inner most lines is 1 standard deviation above (U1Sigma) and below (L1Sigma) from the CL. This is the range that the company should aim to stay withing. This is achieved by having control of the process to ensure that the sample's standard deviation of delivery times stays near the CL.

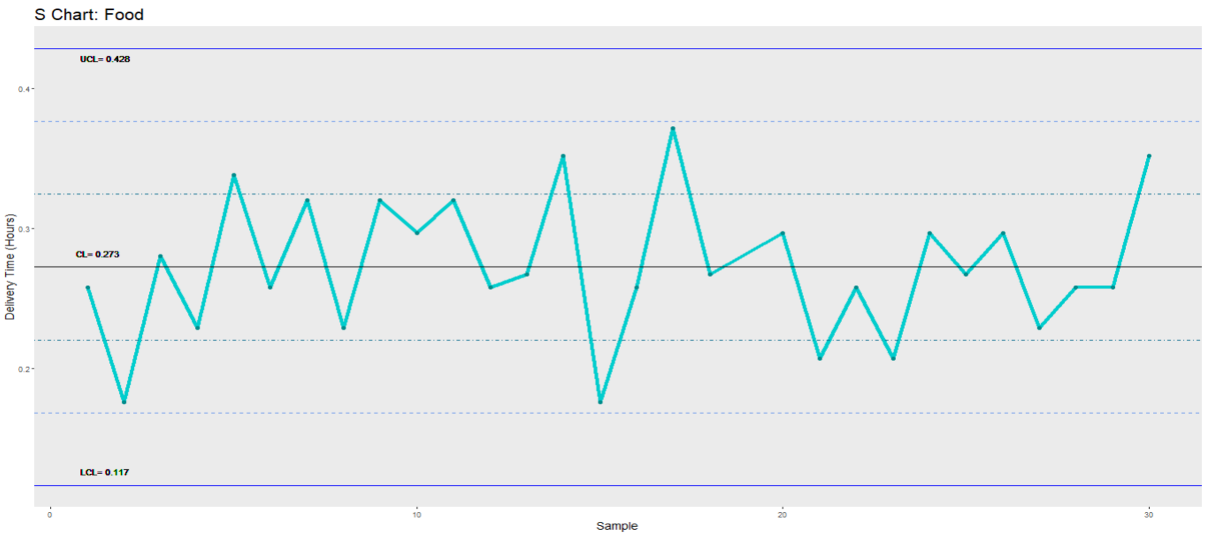


Figure 22: *Graph showing S Chart of initializing samples for food without sample 19 for the first 30 samples (15 sales per sample).*

Figure 22 shows the new initialised S Chart for the food delivery times without sample 19, that as seen on figure 21 above, exceeds the UCL. The solid grey CL indicates that the process is under control. Take note of the new critical values that will be used to evaluate the rest of the food standard deviation delivery times.

This Control Chart in figure 23 is used to calculate the standardised values for the X-Bar

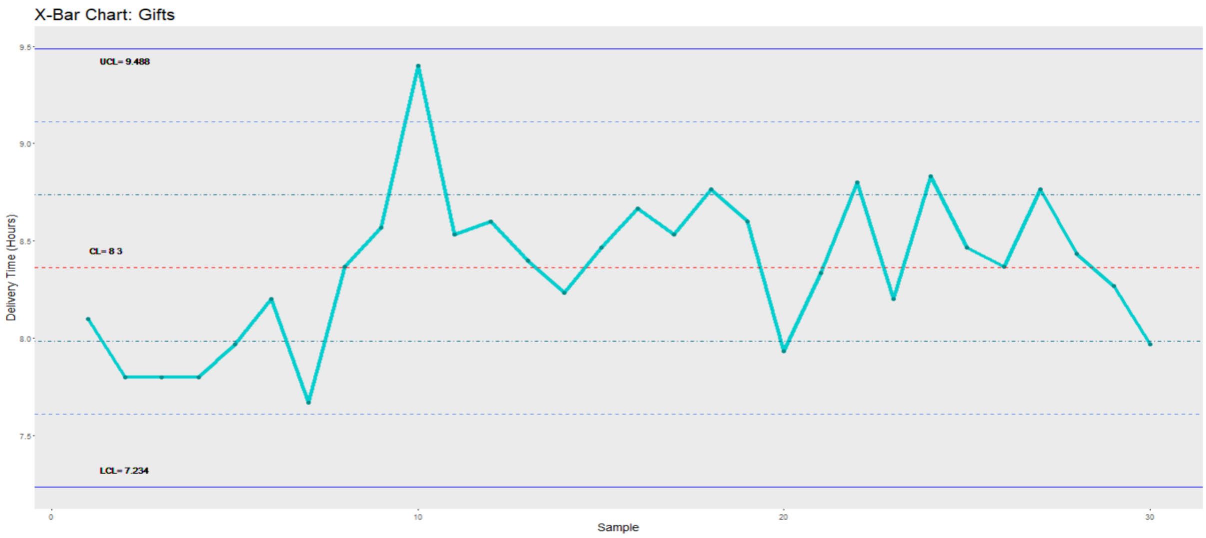


Figure 23: *Graph showing X-Bar Chart of initializing samples for gifts for the first 30 samples (15 sales per sample).*

Charts of all the sales and is similar to the previous graph in figure 21 in terms of the different regions. This particular graph indicates that the mean delivery time for gifts is out-of-control by a red dotted line as the CL. This is caused by 7 consecutive points being under the CL followed by 6 consecutive points being above the CL. There are no points above of below the UCL or LCL but there is one point that exceeded the U2Sigma line, but the company has seen this warning and kept the delivery time under control from there on out.

X-Bar Table: Delivery Time							
Class	UCL (Hours)	U2Sigma (Hours)	U1Sigma (Hours)	CL (Hours)	L1Sigma (Hours)	L2Sigma (Hours)	LCL (Hours)
Technology	22.9731	22.1069	21.2407	20.3744	19.5082	18.6420	17.7758
Clothing	9.4047	9.2598	9.1149	8.9700	8.8251	8.6802	8.5353
Household	50.2462	49.0182	47.7902	46.5622	45.3342	44.1062	42.8783
Luxury	5.4935	5.2409	4.9882	4.7356	4.4829	4.2302	3.9776
Food	2.7093	2.6362	2.5631	2.4900	2.4169	2.3438	2.2707
Gifts	9.4879	9.1123	8.7367	8.3611	7.9855	7.6099	7.2343
Sweets	2.8968	2.7571	2.6175	2.4778	2.3381	2.1984	2.0588

Figure 24: *Table showing X-Bar Chart values of delivery time per class for the first 30 samples (15 sales per sample).*

Figure 24 displays all the critical values for the X-Bar Chart that is used for Statistical Process Control. This will help the company to see which class of product's delivery process requires a more hands on approach in order to keep this process under control. The X-Bar Chart specifically examines the mean of the delivery time as sales are recorded in a chronological order. The sales are grouped into samples of 15 sales per sample which is used to initialise the charts for the rest of the sales. For this specific control chart initialisation step the first 30 samples are considered, which adds up to a total of 450 sales. This is essentially the standard to which the rest of the sales need to conform to. Examining figure 24, it is evident that the household items have the highest centre line CL of 46.56 hours for the delivery time. This is considerable higher than the lowest which is sweets taking 2.48 hours, but it is all relative which is why further analysis is required in order to compare the rest of the sales to the initialising samples. When the samples move above or below the UCL or LCL respectfully, the delivery process is deemed to be out of control.

S Table: Delivery Time							
Class	UCL (Hours)	U2Sigma (Hours)	U1Sigma (Hours)	CL (Hours)	L1Sigma (Hours)	L2Sigma (Hours)	LCL (Hours)
Technology	5.1799	4.5518	3.9237	3.2955	2.6674	2.0393	1.4111
Clothing	0.8664	0.7614	0.6563	0.5512	0.4462	0.3411	0.2360
Household	7.3432	6.4528	5.5623	4.6719	3.7814	2.8910	2.0005
Luxury	1.5109	1.3276	1.1444	0.9612	0.7780	0.5948	0.4116
Food	0.4283	0.3764	0.3244	0.2725	0.2206	0.1686	0.1167
Gifts	2.2460	1.9737	1.7013	1.4290	1.1566	0.8842	0.6119
Sweets	0.8352	0.7340	0.6327	0.5314	0.4301	0.3288	0.2275

Figure 25: *Table showing S Chart values of delivery time per class for 30 samples (15 sales per sample).*

The S table seen in figure 25 is similar to the X-Bar table seen in figure 24, but the S table is used to examine the standard deviation of the delivery times of each product class. Figure 25 contains all the centre lines, and the upper and lower sigma values that are used to initialise the S Charts. Again, these values are the standard to which the rest of the sale samples will be evaluated against. The first 30 samples are used in this initialisation of the s chart also containing 15 sales per sample. The reason for it being 15 sales per sample is because the charts are produced in real time. Every time these charts are updated a further 15 sales have occurred. The CL is effectively the target at which the company aims to deliver their products by, UCL and LCL are the control limits and the Sigma's are the warning limits.

4.2. S & X-Bar Charts for all Samples

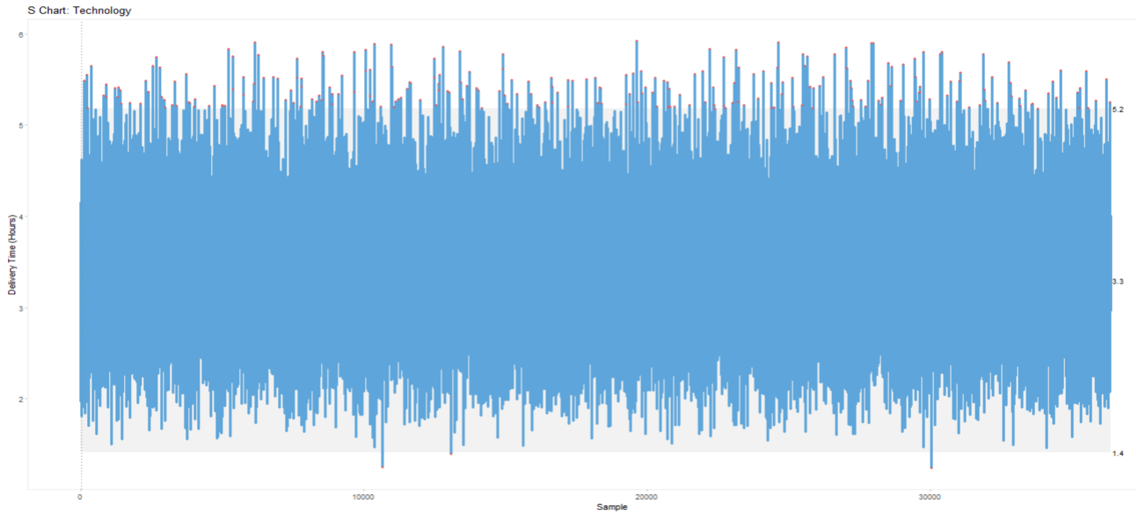


Figure 26: *Graph showing S Chart of delivery time for technology for all samples (15 sales per sample).*

The variation in the standard deviation of delivery time for technology is higher than it should be for some samples as seen in figure 26. There are multiple samples that has a standard deviation that exceeds the UCL. This indicates that the process is out-of-control.

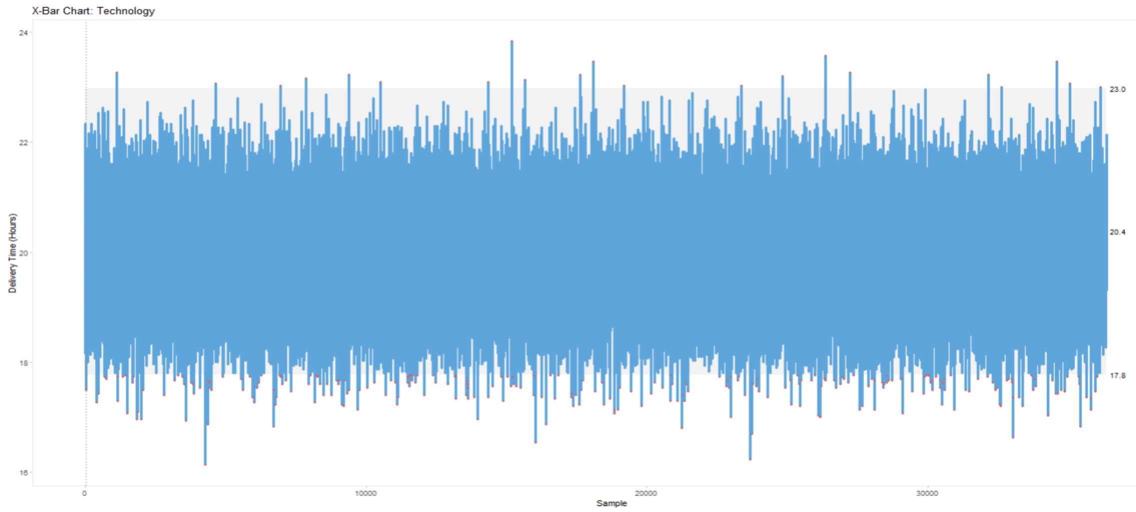


Figure 27: *Graph showing X-Bar Chart of delivery time for technology for all samples (15 sales per sample).*

The mean of all the samples is showed in figure 27 and indicates that the delivery time for technology is mostly within the UCL and LCL, however there are multiple points below the LCL and a couple above the UCL. This means that if the company improves on the variation of the technology delivery times, then this process will be in control.

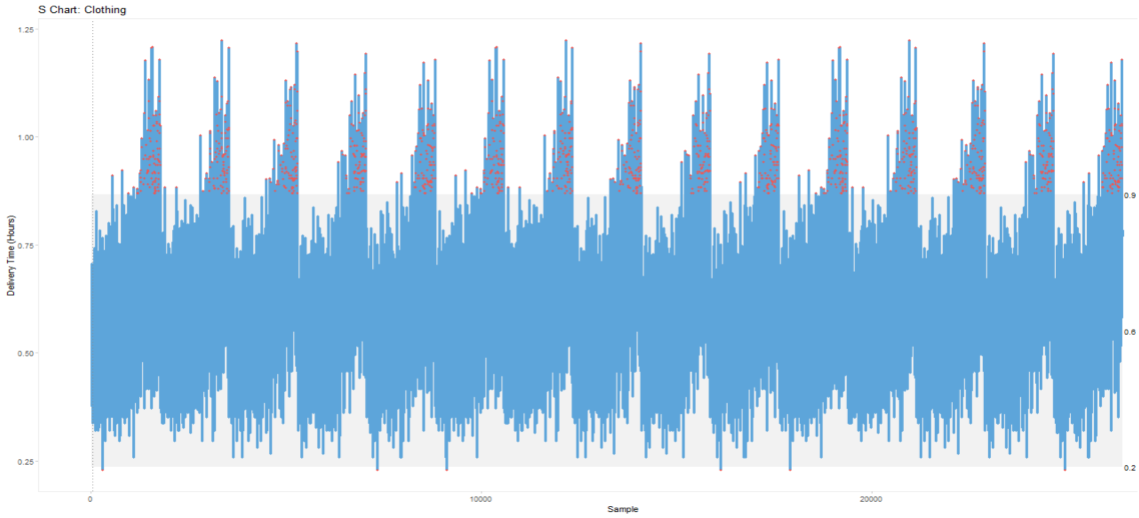


Figure 28: *Graph showing S Chart of delivery time for clothing for all samples (15 sales per sample).*

The standard deviation of the clothing delivery time has a higher variation than the technology class as seen in figure 28, but there is a definite pattern. The cause of this needs to be addressed as this can affect the reliability of the delivery times. This chart indicates that the standard deviation of the clothing delivery times is out-of-control.

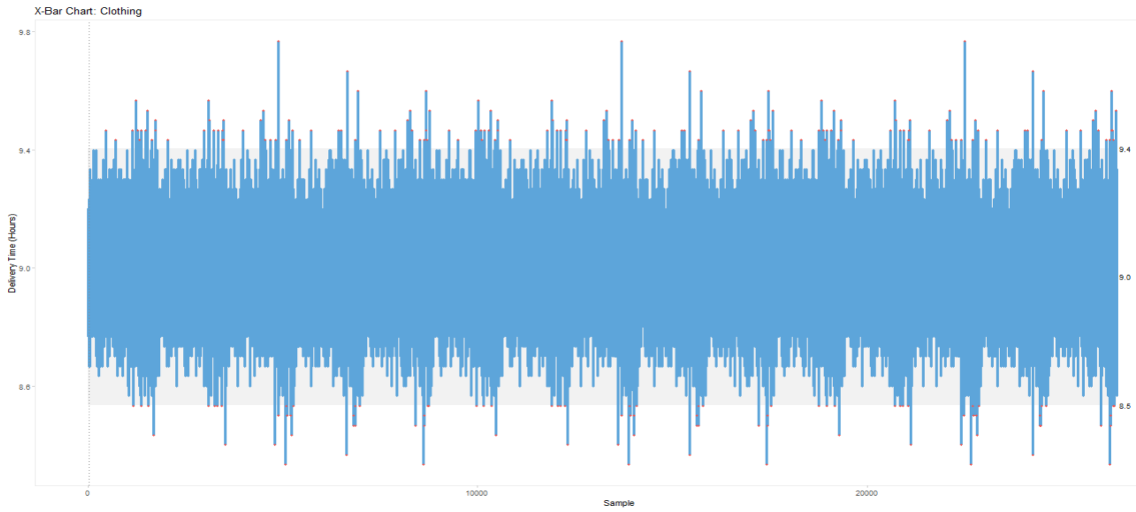


Figure 29: *Graph showing X-Bar Chart of delivery time for clothing for all samples (15 sales per sample).*

Figure 29 examines the mean delivery time for clothing items for all samples. This chart indicated that the process is somewhat under control with a couple of samples being above the UCL and below the LCL. These samples need to be identified and prevented in the future.

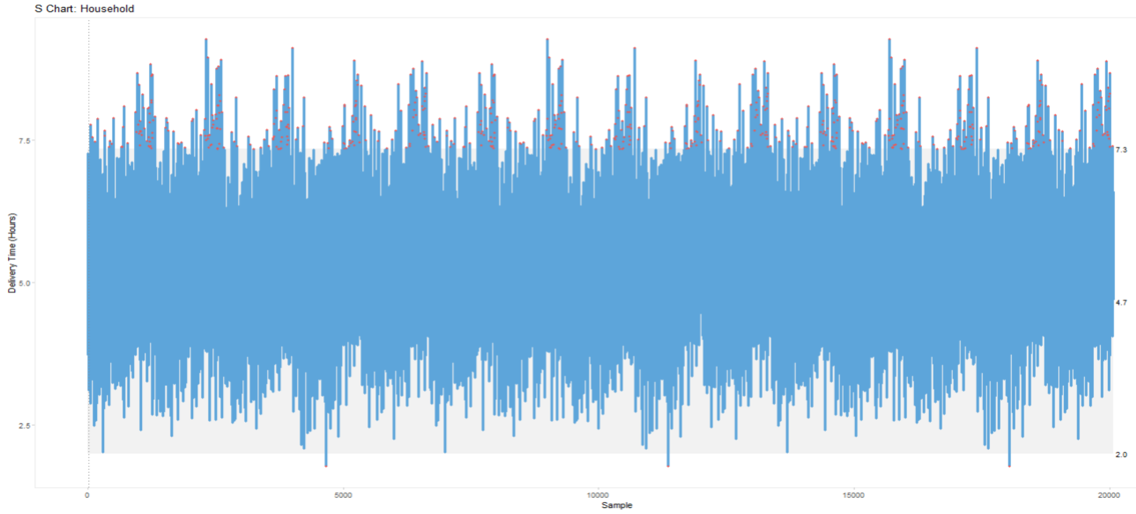


Figure 30: *Graph showing S Chart of delivery time for household for all samples (15 sales per sample).*

The standard deviation for the household delivery time is fairly stable, but there is a pattern where the chart becomes unstable where the sample points are above the UCL as indicated by the red points. This can be seen in figure 30.

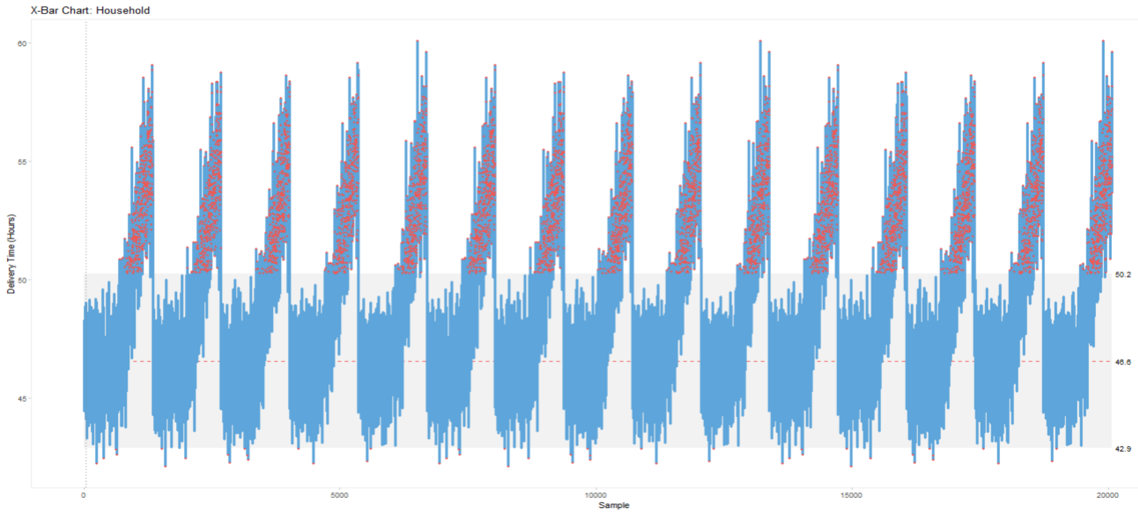


Figure 31: *Graph showing X-Bar Chart of delivery time for household for all samples (15 sales per sample).*

The mean delivery time chart for household items is displayed in figure 31. Like the S Chart seen in figure 30 there is an even more distinct pattern regarding the out-of-control points. This is very concerning and the cause of this out-of-control pattern needs to be identified and rectified.

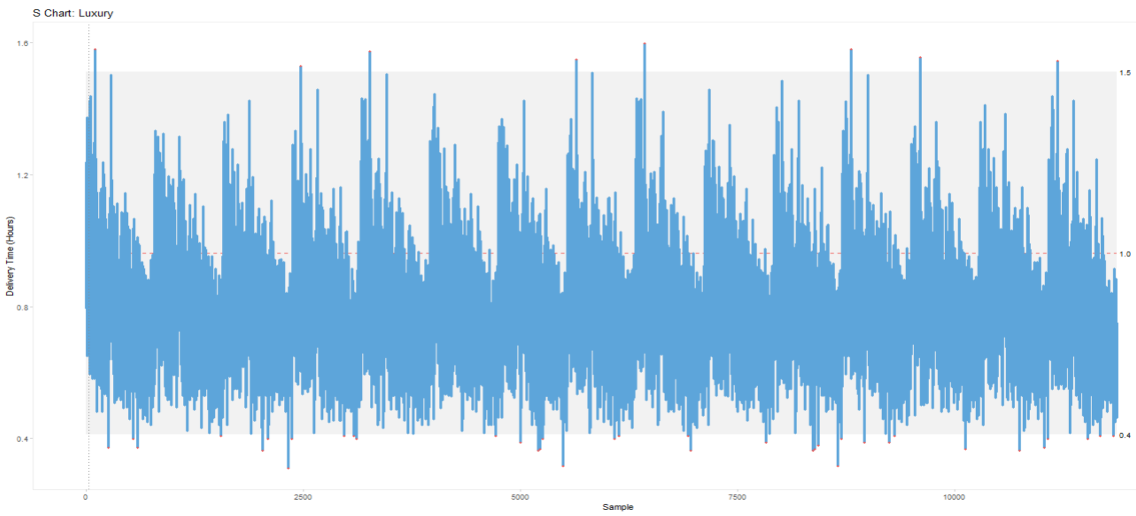


Figure 32: *Graph showing S Chart of delivery time for luxury for all samples (15 sales per sample).*

Figure 32 is the S Chart of the luxury delivery times. The process is out-of-control as indicated by the red dotted line on the CL even though most of the points are within the UCL and LCL. The cause of this out-of-control line is the fact that the majority of the sample points are below the CL and not oscillating over the CL as an in-control chart would.

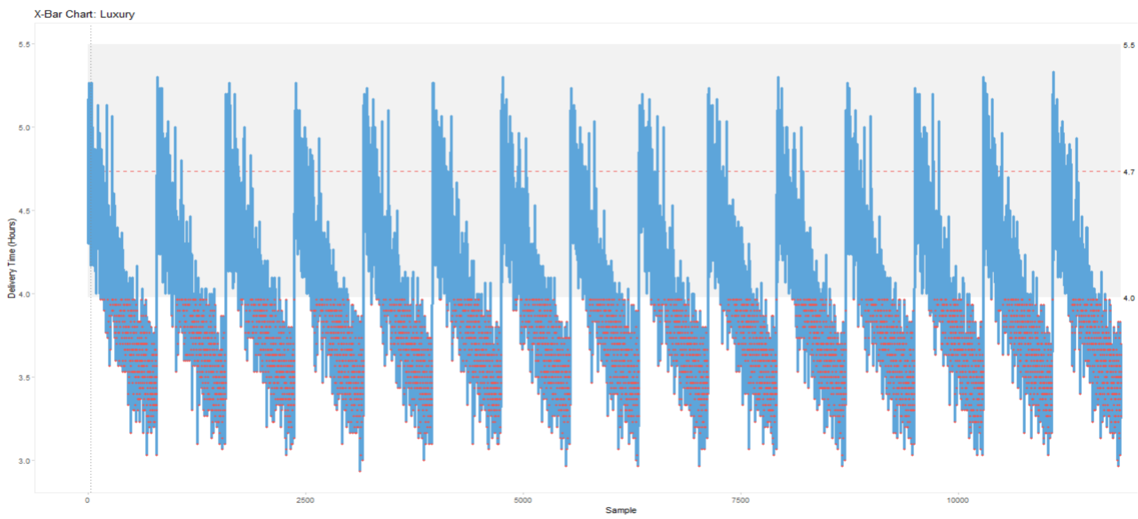


Figure 33: *Graph showing X-Bar Chart of delivery time for luxury for all samples (15 sales per sample).*

The X-Bar Chart displayed in figure 33 should result cause a lot of concern for the company regarding the mean delivery time of luxury items. The process is out-of-control as indicated by the red dotted line. Majority of the sample points are below the LCL meaning that the company is delivering the sales faster than what the customer expects. This is not a concern for the customer, but the concern for the company is that they could potentially save money on deliveries by increasing the delivery time without affecting the customer experience negatively. The variation in the delivery times will have to be a priority before fixing this process.

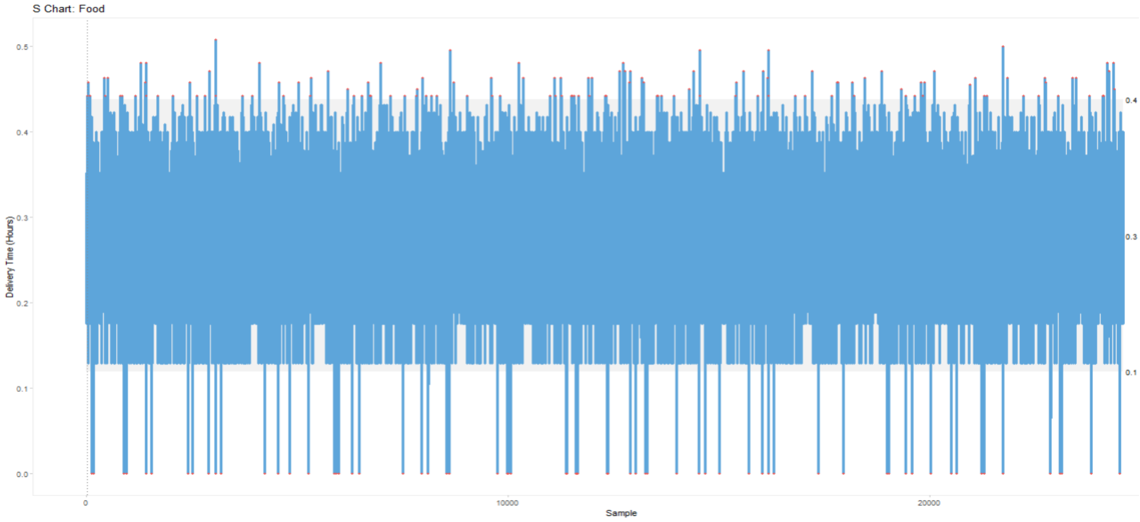


Figure 34: *Graph showing S Chart of delivery time for food for all samples (15 sales per sample).*

The delivery times for food as a relatively under control mean standard deviation, but there are some outlying samples that are above and below the UCL and LCL respectfully as seen in figure 34.

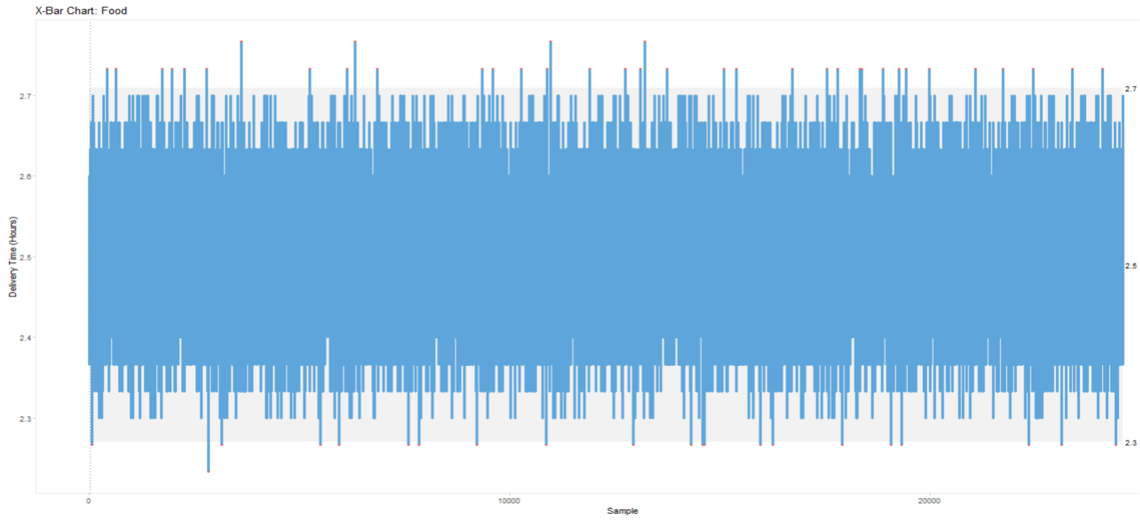


Figure 35: *Graph showing X-Bar Chart of delivery time for food for all samples (15 sales per sample).*

The same goes for the X-Bar Chart in figure 35. The majority of the sample's mean delivery time for food is withing the UCL and LCL with a few exceptions. This indicates that the process is stable.

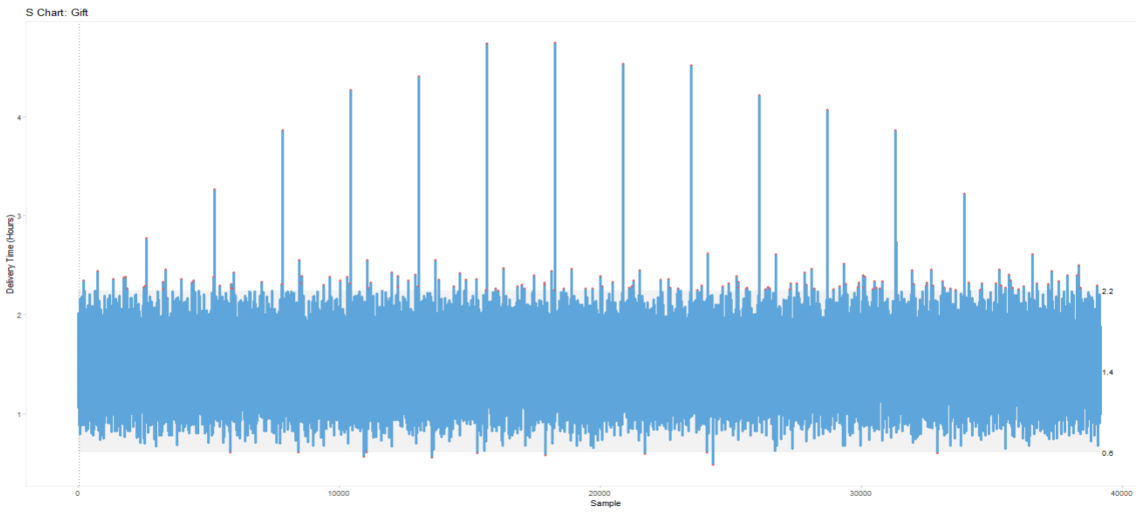


Figure 36: *Graph showing S Chart of delivery time for gifts for all samples (15 sales per sample).*

The S Chart for the gifts shows that the process is somewhat stable except for a couple of points exceeding the UCL and LCL. There are however extreme outliers as seen in figure 36 that have exceeded the UCL by a very big margin which is very concerning. There is a pattern regarding these extreme outliers meaning that there is a specific cause that needs to be identified and managed better by the company.

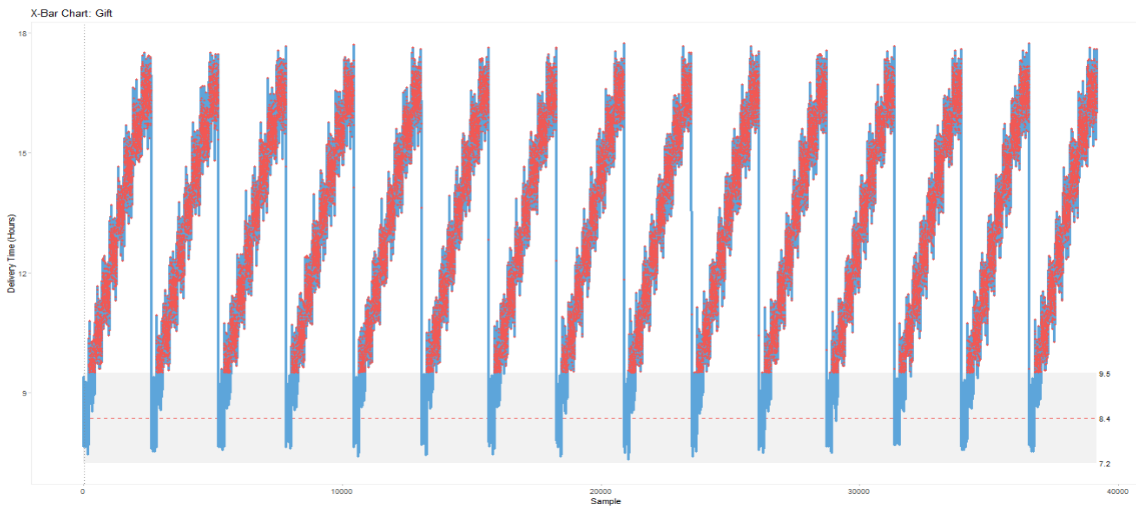


Figure 37: *Graph showing X-Bar Chart of delivery time for gifts for all samples (15 sales per sample).*

The mean delivery times for the gift sale samples presents a very distinct pattern as seen in figure 37. The corresponds to the extreme outliers seen in the previous chart (Figure 36). The process is out-of-control and very unstable as indicated by the red dotted line caused by the majority of the sample points being above the UCL.

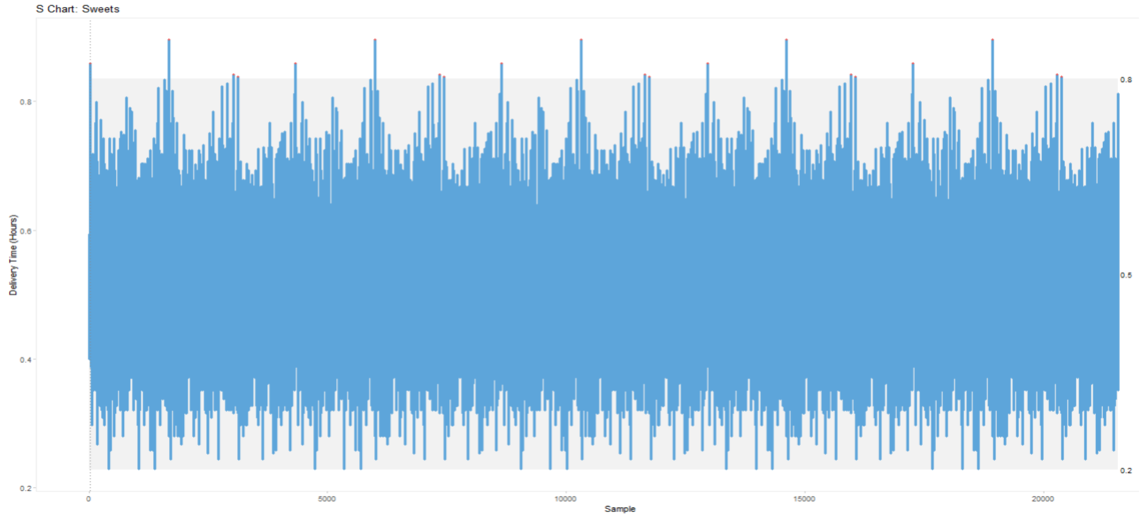


Figure 38: *Graph showing S Chart of delivery time for sweets for all samples (15 sales per sample).*

Figure 38 shows the mean standard deviation for all the delivery time samples for sweets sales. This chart indicates that the process is stable with a few points being above the UCL and majority of the sample points being close to the CL.

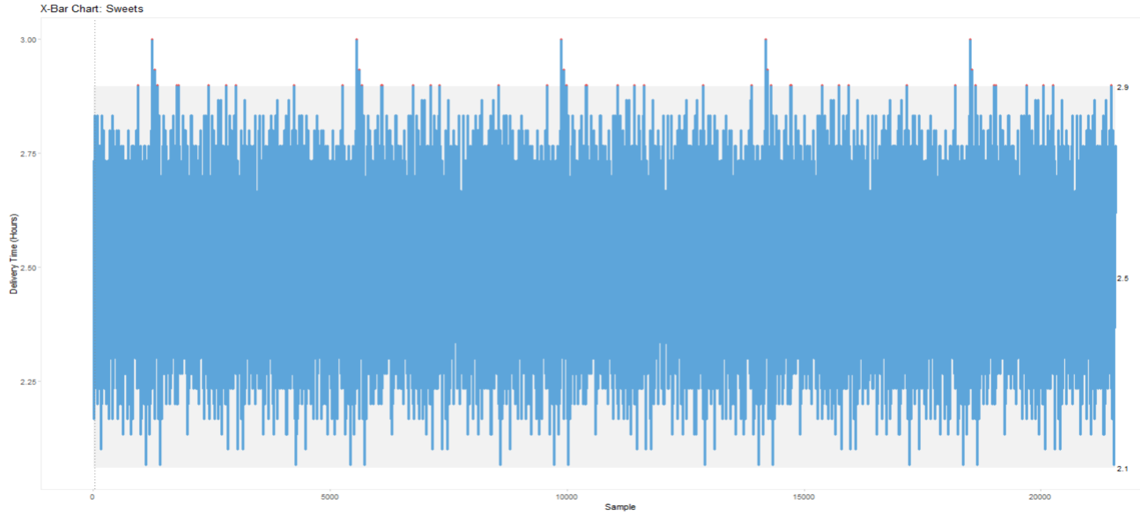


Figure 39: *Graph showing X-Bar Chart of delivery time for sweets for all samples (15 sales per sample).*

The X-Bar Chart displayed in figure 39 indicates that the mean delivery time for sweets is stable and in-control. The majority of the sample points are well within in the UCL and LCL, but there are a few exceptions where the points exceed the UCL.

5. Part 4

5.1. Consecutive Count

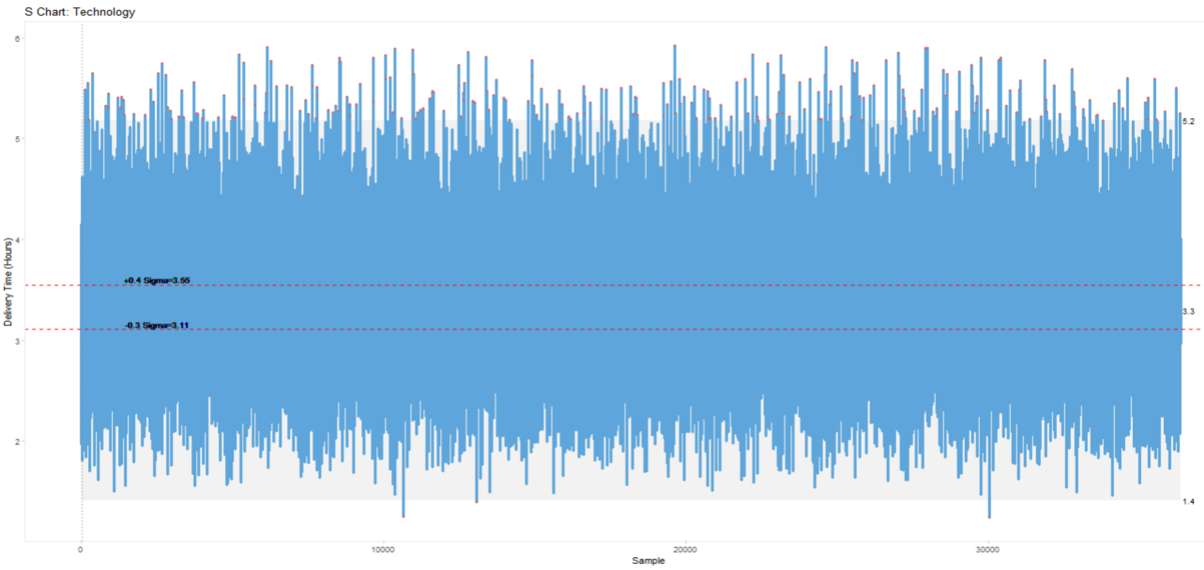


Figure 40: *Graph showing S Chart of delivery time for food for all samples (15 sales per sample).*

Figure 40 presents the range that mean standard deviation of the samples needs to be in to be considered withing range (+0.4 σ and -0.3 σ). This is visualized by the two red dotted lines. The same concept is used for all the other classes and the results are shown below.

Consecutive S Chart Table		
Class	Max	Last Within
Technology	6	36 334
Clothing	5	26 394
Household	6	20 064
Luxury	5	11 842
Food	6	24 559
Gifts	9	39 145
Sweets	6	21 555

Figure 41: *Table showing the maximum consecutive count where samples are within +0.4σ and -0.3σ and the last sample within range.*

Figure 41 shows the maximum consecutive count what all the class's sample standard deviation mean is within the range of +0.4 σ and -0.3σ as seen in figure 40. Gifts has the most consecutive samples withing this range sitting at a maximum of 9. The fact that the range is different from UCL and LCL range – closer to the CL – means that the class with a higher consecutive count should potentially be more stable.

5.2. Type I Error A

It is very important to know how accurate the results are from the S and X-Bar Charts in order not to be bias in the findings. The Type I Error is the risk of finding that the delivery process is out-of-control even though it is not. Searching for the cause of something that is not there can be a big waste of time and money, hence being aware of this is very important.

Type I Error:

$$\alpha = \left(1 - pnorm(3)\right) \cdot 2$$

$$= 0.0026998$$

This means that there is a 0.026998% chance that the charts indicate that the process is out-of-control even though it is not.

5.3. Type I Error B

Type I Error:

$$\alpha = pnorm(0.4) - pnorm(-0.3)$$

$$= 0.2733332$$

This 27.33% is the probability that the mean standard deviation for the delivery time samples is shown to be outside of range as indicated in figure 40 above even though it is within range. This probability indicates that the chart should not be relied on alone, but further investigation will need to be done.

5.4. Technology Delivery Time Reduction Cost

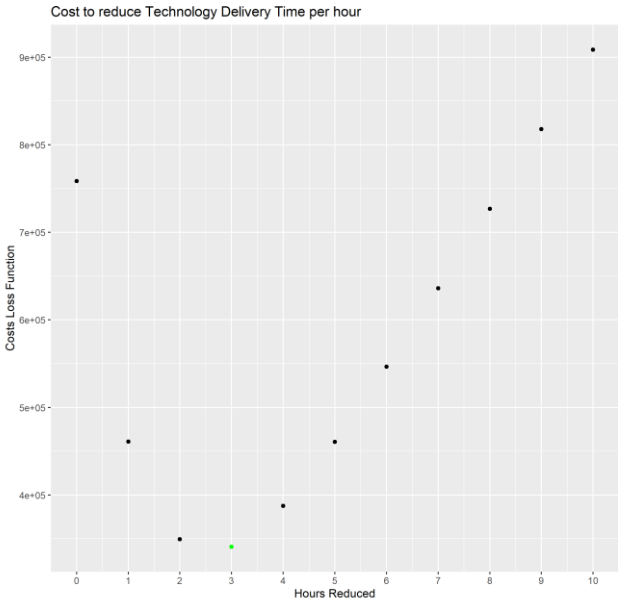


Figure 42: *Graph showing 10 iterations of the cost loss function to reduce the delivery time for technology*

The delivery of technology items is considered late if it takes longer than 26 hours. There is a R329 cost associated with each item that is late for an hour. This means that if there are two items that are both two hours late, the costs lost is these two items is:

$Cost\ Lost = R329(2)(2) = \text{R1 316}$

There is however an option to reduce the average delivery time by an hour at a cost of R2-50. This means that the total costs loss is:

$$Costs\ Loss\ Function =$$

$$(Total\ hours\ late)(R329) + (R2.50)(\#of\ technology\ sales)(hours\ reduced)$$

In order to find the optimal time to reduce the average technology delivery time, 10 iterations were done as in figure 42. The optimal time is visualised by the green dot. This graph indicates that the average delivery time for technology items needs to be reduced by 3 hours making the Costs Loss Function a total of R340 870 whereas it would be R758 674 if the company did nothing.

5.5. Type II Error

It is once again to understand the results for the all the graphs and tables when analysing the data, because these results will impact the decision making of the company. These results may be misleading and it will be very helpful to know what the chances are that this is the case. Type II Error takes in consideration the probability that the X-Bar Charts indicates that the technology delivery process is stable even though it is not given that the mean has shifted to 23 hours.

$$Type\ II\ Error = pnorm(UCL, 23, \sigma) - pnorm(LCL, 23, \sigma)$$

$$= 48.76147\ \%$$

UCL - 22.9731

LCL - 17.7758

σ - 0.86621

This is concerning, indicating that the control charts might not be the best method to see if the delivery process for t

6. Part 5

6.1. MANOVA Testing

It is important to understand that a company can only control the controllables, but it becomes difficult to find what these controllables are because everything is intertwined. This is why MANOVA testing is very useful to find which dependant variables are statistically significant to some independent variables. It would be very significant to know if class, why bought and price has an impact on the age of the customers as well as the delivery time. All 3 of these dependant variables are known before the sale is made, hence the independent variables can then be predicted if there is a statistical significance. Another important test to do would be to know if the price of a product is statistically significant to the class of the product, why it was bought and the delivery time.

Response Age:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Class	6	8.42E+06	1.40E+06	3 805	2.00E-16
Why.Bought	5	4 625	925	2.51E+00	0.028
Price	1	41	41	1.10E-01	0.740
Residuals	179 970	6.64E+07	369		

Response Delivery Time:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Class	6	3.35E+07	5.58E+06	6.30E+05	2.00E-16
Why.Bought	5	107	21	2.41	0.034
Price	1	0	0	3.87E-02	0.844
Residuals	179 970	1 594 356	9		

Figure 43: *Table showing MANOVA summary of independent (class, why bought and price) vs dependent (age and delivery time) variables.*

H_0: Class, Why Bought and Price is not statistically significant to the combination of age of customer and delivery time.

H_a: They are of statistical importance.

As indicated by the small Pr(>F) values in figure 43, class is very significant and can potentially be used to predict delivery time and age (Reject H_0). Why bought is less significant but still significant as the Pr(>F) value is <0.05 (Reject H_0). The price however is not statistically significant to the age of customers and delivery time, because the Pr(>F) is >0.05 (Accept H_0).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Response Class:					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Price	1	189 384	189 384	55 795	2.20E-16
Residuals	179 981	610 909	3		
Response Why Bought :					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Price	1	182	182	99	2.20E-16
Residuals	179 981	332 224	2		
Response Delivery Time :					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Price	1	304 242	304 242	1 576	2.20E-16
Residuals	179 981	3.48E+07	193		

Figure 44: *Table showing MANOVA summary of independent (price) vs dependent (class, why bought and delivery time) variables.*

H_0: Price is not statistically significant to the combination of class, why bought and delivery time.

H_a: They are of statistical importance.

Here it is clear to see in figure 44 that the combination of class, why bought and delivery time is statistically significant to price as all the Pr(>F) values are very small meaning that all the H_0 hypothesis are rejected.

7. Part 6

7.1. Taguchi Loss

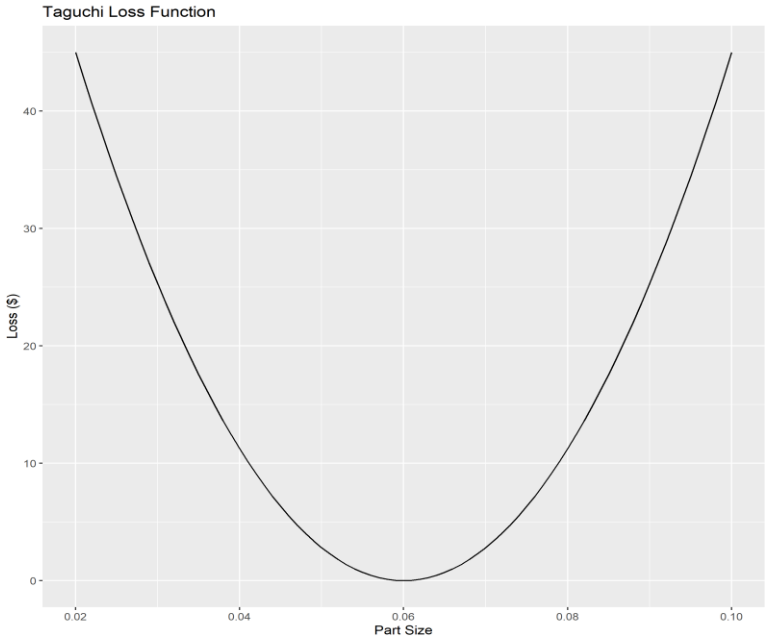


Figure 45: *Graph showing the Taguchi Loss Function.*

The Taguchi Loss Function showed in figure 45 describes the cost that is associated with deviating from the product specifications. This is not a linear relationship as seen in the graph but rather a parabola where no cost is lost if the part size meets the specifications exactly. As the size deviates this cost grows exponentially until the part can no longer be sold and has to be scrapped at a maximum cost of $45/part. The parabola has the following equation:

$$Loss\ Function\ (\$) = 28\ 125 * (y - 0.06)^2 \qquad y\ \epsilon\ [-0.4; 0.4]$$

Where y is the deviation from the part size specification.

If the cost to scrap a part were to reduce to $35 the Taguchi Loss Function will look as follows as seen in figure 46. The new parabola was the following equation:

$$Loss\ Function\ (\$) = 21\ 875 * (y - 0.06)^2 \qquad y\ \epsilon\ [-0.4; 0.4]$$

Where y is the deviation from the part size specification.

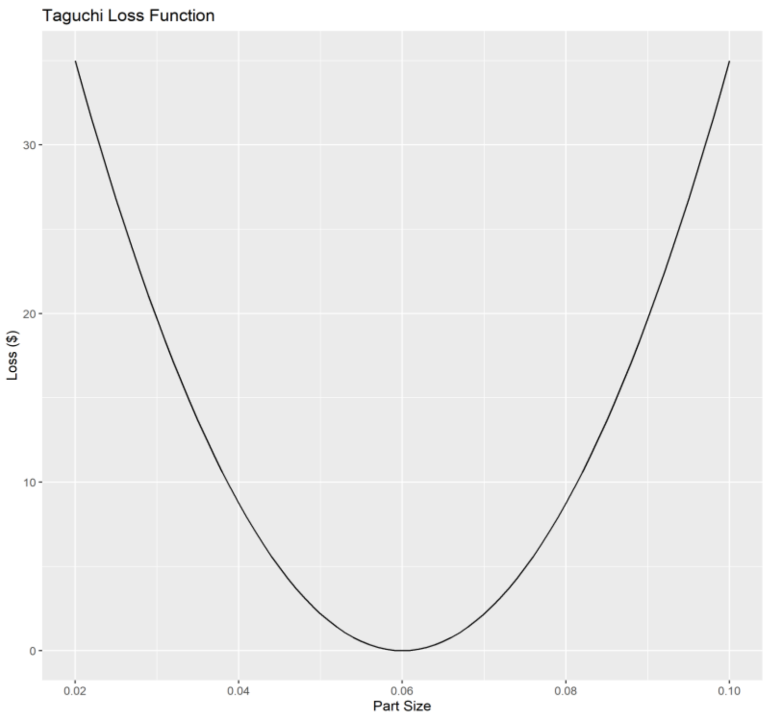
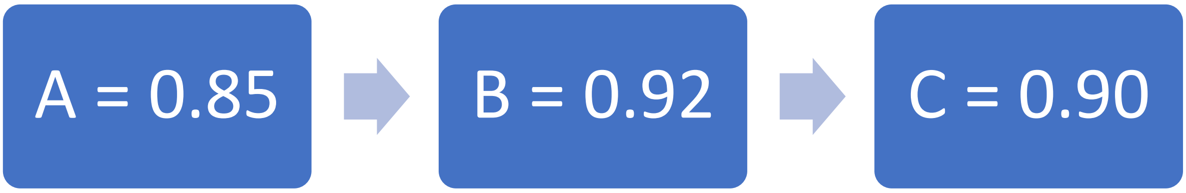


Figure 46: *Graph showing the Taguchi Loss Function.*

7.2. System Reliability

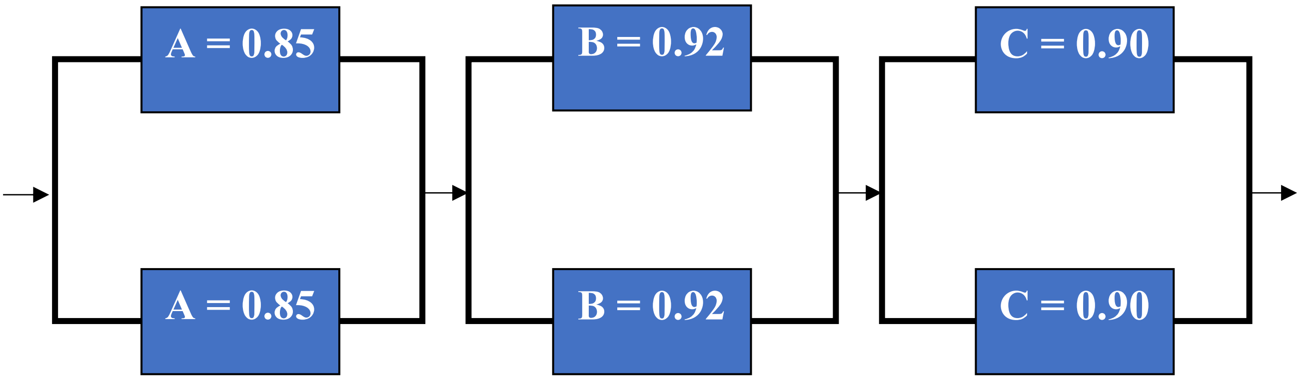
In series system:



$$Reliability = 0.85 * 0.92 * 0.90$$

$$= 0.7038$$

In parallel system:



$$Reliability = [1-(1-0.85)^2][1-(1-0.92)^2][1-(1-0.90)^2]$$
$$= 0.96153$$

The adding of the back-up machines improves the system's reliability by 25.77%. This is very significant and was definitely a good decision made by the company.

7.3. Delivery Process Reliability

If there were to be 21 vehicles and 21 drivers the probabilities look as follows:

$P(Vehicles\ reliable') = 9.137794e-06$

$P(Drivers\ reliable') = 0.01329$

$P(Delivery\ process\ is\ reliable) = 1 - P(Vehicles\ reliable') * P(Drivers\ reliable')$

$\#Days\ in\ year\ to\ expect\ reliable\ delivery\ times =$

$P(Delivery\ process\ reliable) * 365 = 360.15 = 360\ days$

There is little to no change in the reliable delivery days when there is an increase from 21 to 22 vehicles.

8. Conclusion

Raw data by itself is not useful at all, but after some processing and analysing there is a lot to learn from it. The descriptive statistics and graphs have led the analysis in the right direction to find patterns and trend that can give the company a competitive advantage and improve the customer experience whilst still reducing costs. The delivery process needs to be examined as there is a lot of room for improvements.

9. References

(no date) *Index of /web/packages/manova.rm/vignettes*. Available at: https://cran.r-project.org/web/packages/MANOVA.RM/vignettes/ (Accessed: October 24, 2022).

Binomial probability example (video) (no date) *Khan Academy*. Khan Academy. Available at: https://www.khanacademy.org/math/ap-statistics/random-variables-ap/binomial-random-variable/v/probability-of-making-2-shots-in-6-attempts (Accessed: October 24, 2022).

Manova - Gusta Me (no date) *Google Sites: Sign-in*. Available at: https://sites.google.com/site/mb3gustame/hypothesis-tests/manova (Accessed: October 24, 2022).

MANOVA test in R: Multivariate analysis of variance (no date) *STHDA*. Available at: http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance (Accessed: October 24, 2022).

Theuns (no date) *QA344 Statistics, SUNLearn*. Available at: https://learn.sun.ac.za/ (Accessed: October 24, 2022).