Consumer online buying trends from 2021-2029

Prepared for: Quality Assurance 344

Prepared by: Fisiwe Hlophe – 24181781

Date: 26 September 2022

Abstract

Retail companies that are built on an online market place must be designed to provide a wide variety of consumer products. It is imperative that such companies conduct trend analysis on a regular basis to not only procure new customers, but to also maintain and/grow the brand image and increase sales. Trend analysis is done by gathering and interpreting data to better understand the behaviours of customers. This report uses statistical tools, combined with the R-Studio coding platform to analyse customer data for an online retail company. The report concludes that the company's delivery times are too high, leading to customer dissatisfaction and thus, need to be optimized.

Table of Contents

1. Introduction

Data analytics enable businesses to conduct market research more quickly and accurately. The insights gained can help a business become more relevant and more competitive in the market. By becoming more efficient, businesses can lower production costs and attract new customers – guaranteeing high profit generation. In this report, data analytic principles were used to study the online buying trends of customers.

Firstly, the data set under study was wrangled to eliminate invalid variables. Next, general data exploration was done to gain an understanding of all the features in the data set and their significance in the study. Statistical analysis was then conducted on the data set, with the effectiveness of different optimization methods computed through the calculation of probabilities. Finally, an analysis of the reliability of the company's delivery times was conducted, as this is the biggest factor constituting to customer satisfaction for the company.

2. Data wrangling

A valid data set was first obtained from the raw data. This was done by removing the rows of data with missing values and invalid values (data with negative prices). The valid data set obtained had 179978 observations and 10 variables (features). Figure 2.1 below shows a sample of the valid client data set.

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

Figure 2.1 Sample of valid client data set.

3. Data exploration using statistics

In order to study the client data set, statistical analysis was performed. The aim of data exploration is to obtain a fundamental understanding of the client data, as well as the understanding of each feature in the data set.

3.1 Analysis of prices clients were charged

Firstly, an analysis of different prices clients were charged was conducted. This was done to find the mean price and modal price to determine the popular prices so that this information can be used to determine the prices of future products. It was gathered that the mean price was R2259.63. The minimum price clients were charged was R35.65 and the maximum price clients were charged was R116 618.97. The modal price, which is the price that was most clients were charged was R567.18.

3.1.1 Determining outlier prices

Prices that are outliers need to be determined as these are the most uncommon prices. Thus, if charged on a product, will result in fewer sales. The lower quartile (LQ), which represents the median of 25% of the prices was R482.31, and the upper quartile (UQ) – which is the median price of 75% of the prices was R15 270.97. Hence, the interquartile range was R14 788.60. The outlier determining constant was thus 1.5*R14 788.60 = R22 182.90. This means that any price greater than R15 270.97 + 22 182.90 = R37 453.87 is an outlier. Prices above R37 453.87 have fewer sales, and thus should be avoided. Figure 3.1.1.1 below shows this distribution, by showing a histogram of the number of clients that were charged different prices.

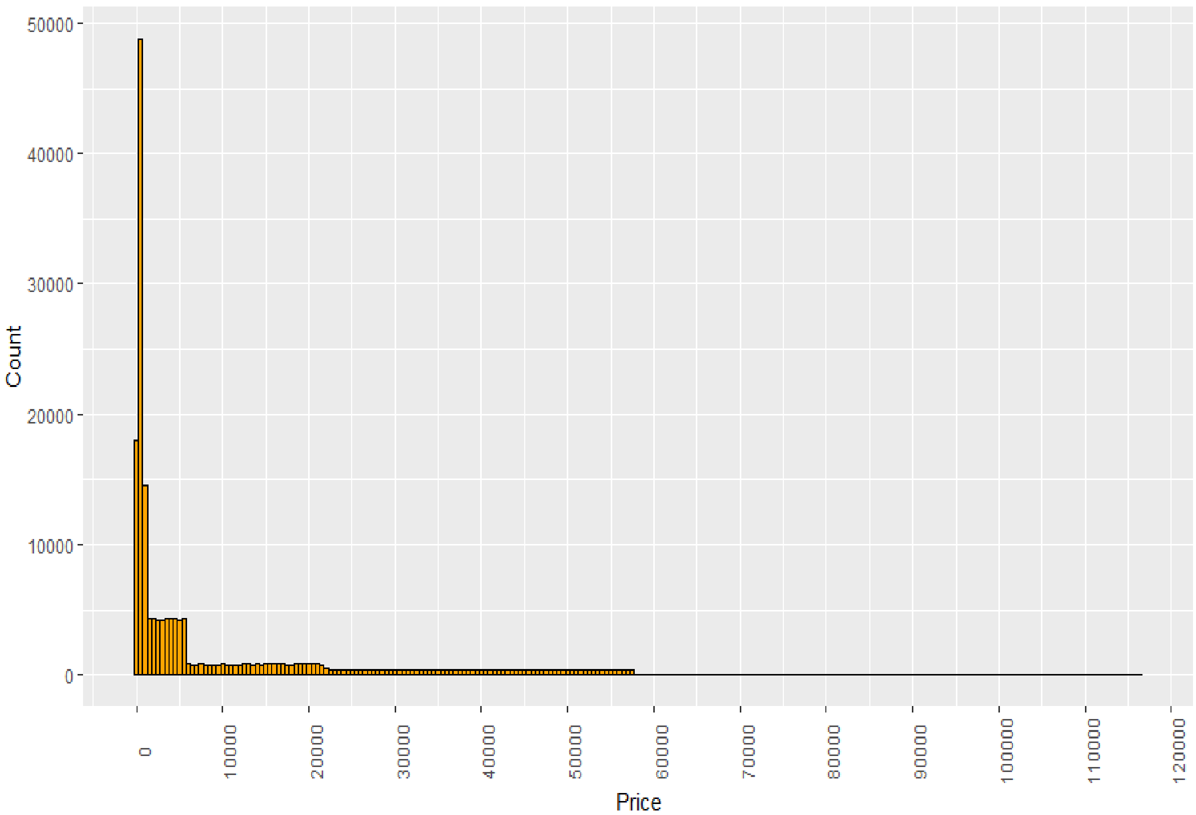

Figure 3.1.1.1 The number of clients charged different prices.

The histogram is right-skewed, showing that most clients were charged lower prices, in the range R0 – R10 000. To maximize sales, prices have to be within this range.

3.2 Analysis of product classes

For further understanding, a bar chart was created showing the mean prices per product class. Figure 3.2.1 below shows the resulting histogram.

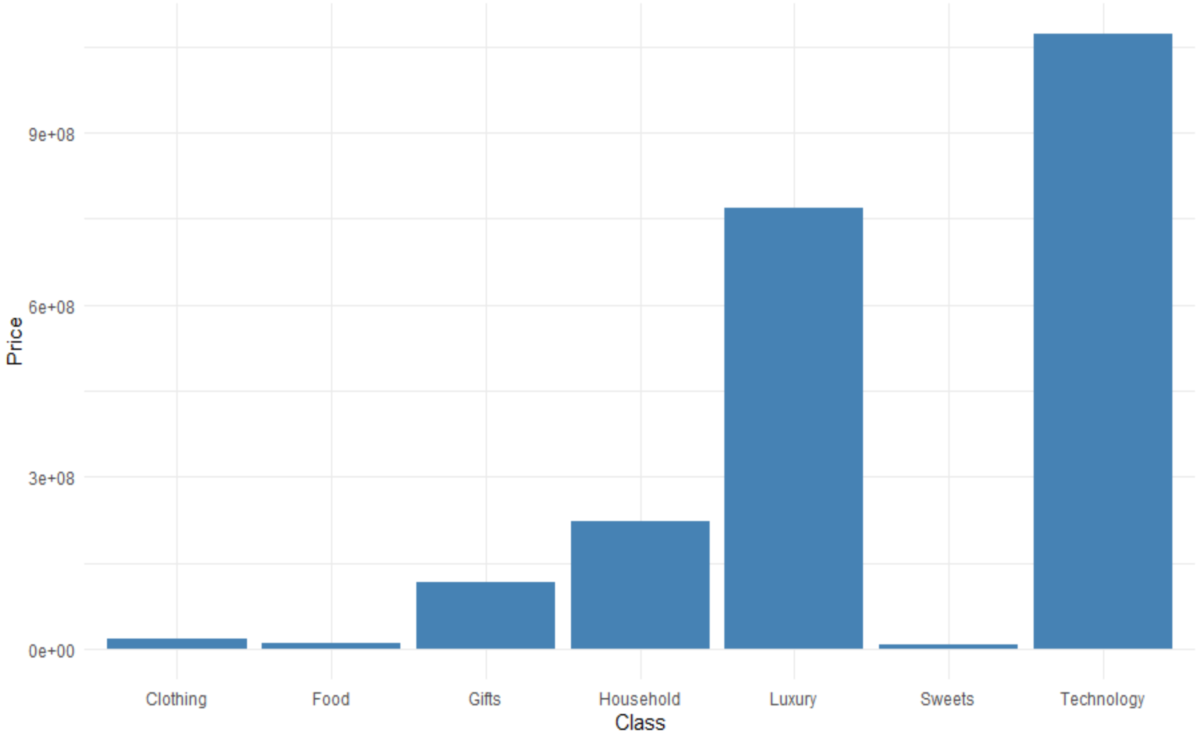

Figure 3.2.1 Price per product class

The bar chart above shows technology products have the highest prices, followed by luxury goods. Sweets have the lowest prices, and food products have the second lowest prices.

The frequency of the different product classes was also computed. Table 3.2.1 below shows the results of this computation.

Class	**Frequency**
Clothing	26 403
Food	24 582
Gifts	39 149
Household	20 065
Luxury	11 868
Sweets	21 564
Technology	36 347

Table 3.2.1 Frequency of the different product classes

The results above show that gifts are the most frequently purchased products, as they have the highest frequency of 39 149 sales. The second most frequently purchased items are technology products. Luxury goods are the least purchased products, as they have the lowest frequency of 11 868 sales.

3.3 Analysis of online marketing strategies

The bar chart in figure 3.3.1 below shows the number of products that were purchased as a result of different online marketing strategies. This allows for the identification of the best marketing strategy.

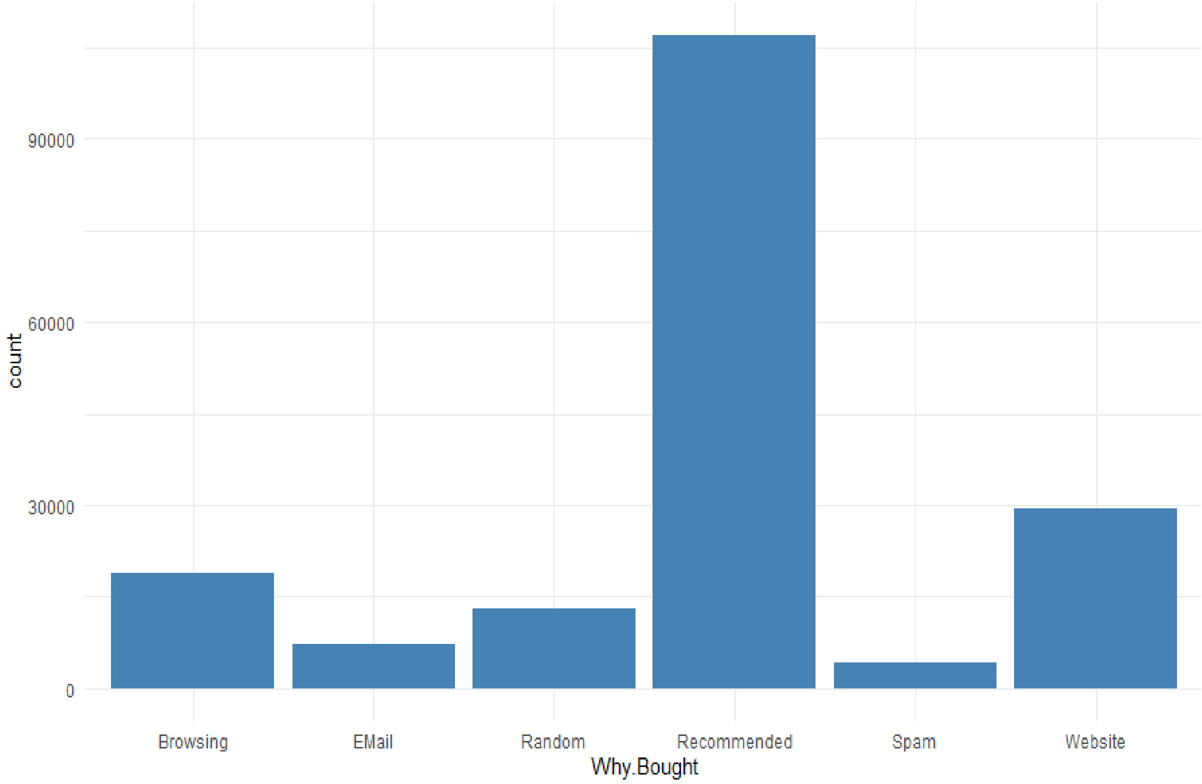

Figure 3.3.1 Number of products bought owing to different online marketing strategies

The figure above shows that clients mostly buy products if they have been recommended to them, showing that recommendation is the most effective marketing strategy. As seen from the graph, clients are least likely to buy products that are marketed via spam. Thus, spam is the least effective online marketing strategy.

3.4 Process capability for the delivery times

The delivery times of technology products were examined through the use of process capability indices. The delivery times were expressed in hours in the data set. The process capability indices computed whether technology products were delivered within 24 hours. Thus, the upper specification limit (USL) for the delivery times was 24 hours, which is the maximum acceptable number of hours a client can wait for a technology product to be delivered to them. Consequently, the lower specification limit (LSL) was 0 hours – the minimum number of hours a client can wait to receive their technology purchase. The USL and LSL are based on the range of hours in a day, which is 24 hours.

Firstly, the normality of the delivery times had to be tested. This is because the process capability indices can only be used if the data under consideration is normally distributed.

A Q-Q plot (quantile-quantile plot) was used to visually check for the normality of the delivery times for technology products. This plot shows the correlation between the delivery times and a normal distribution, plotted along a 45° reference line. Figure 3.4.1 below shows this plot.

FiguF

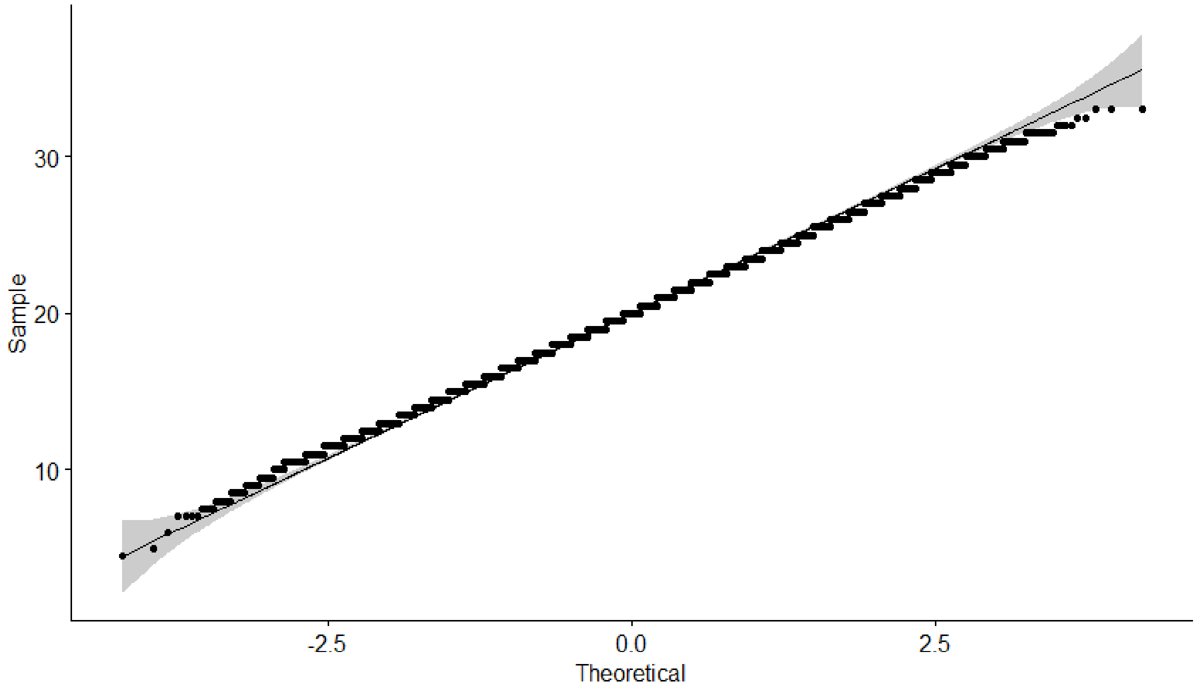

Figure 3.4.1 Q-Q plot of delivery times for technology products

In the Q-Q plot above, the points lie along the 45° reference line. This shows that the delivery times can be assumed to be normally distributed. Hence, process capability calculations will yield valid results.

The process capability indices were determined using the following formulae:

3.4.1 CP = (USL- LSL)/6σ
3.4.2 CPU = (USL - μ)/3σ
3.4.3 CPL = (μ – LSL)/3σ
3.4.4 CPK = min(CPL, CPU)

Where,

σ = the standard deviation of the delivery times of technology products, which was found to be 3.50199

μ = the mean of the delivery times of technology products, which was found to be 20.01095 hours.

The results of the process capability indices are shown in table 3.4.1 below.

Process capability index	**Value**
Cp	1.412
Cpu	0.380
Cpl	1.905

Cpk	0.380

Table 3.4.1 Process capability indices

A process is ideal if its Cpk value is higher than 1. In this case, the Cpk value is very low, showing that the delivery times are too long and thus, may lead to customer dissatisfaction. Reasons for this will have to be investigated in order to optimize the delivery time of these products.

4. Statistical process control (SPC)

An SPC test was carried out by drawing 30 samples of each product type and calculating their mean and standard deviation. The SPC test was carried out on the delivery times for each product class. SPC allows the sample means to form their own normal distribution. X-bar and S-charts were then created so that patterns in the data set could be visualized. X-bar charts indicate the mean of the variables, whilst S-charts indicate the standard deviation. The visualization of these charts allows for the identification of trends in the delivery times of the product classes, so that areas of waste can be identified and corrected.

The data set was first rearranged in ascending order with respect to the year, month and day. Thereafter, for each product type, a sample of 30 products was drawn and an X-bar and S-chart was created. For each chart, the centre line (CL) was calculated. The centre line represents the mean of those first 30 samples. The upper control limit (UCL) and lower control limit (LCL) were also calculated. The UCL gives the value of “mean + 3*sample error”, whilst the LCL gives the value of “mean-3*sample error”.

A sigma (σ) divides the area between the CL and the control limits into six equal parts – three above and below the CL. This gives an indication of how the samples are spread from the mean.

Figure 4.1 and 4.2 below show an example of X-bar and S-chart plots for the technology class after initialization.

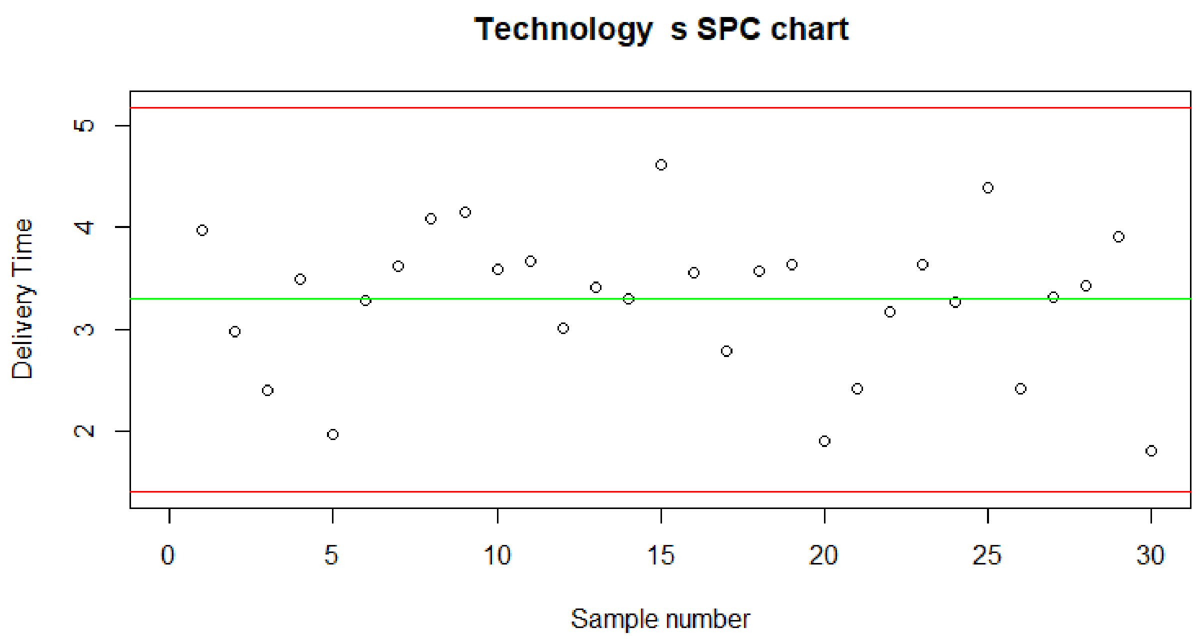

Figure 4.1 Technology S-chart for the first 30 samples

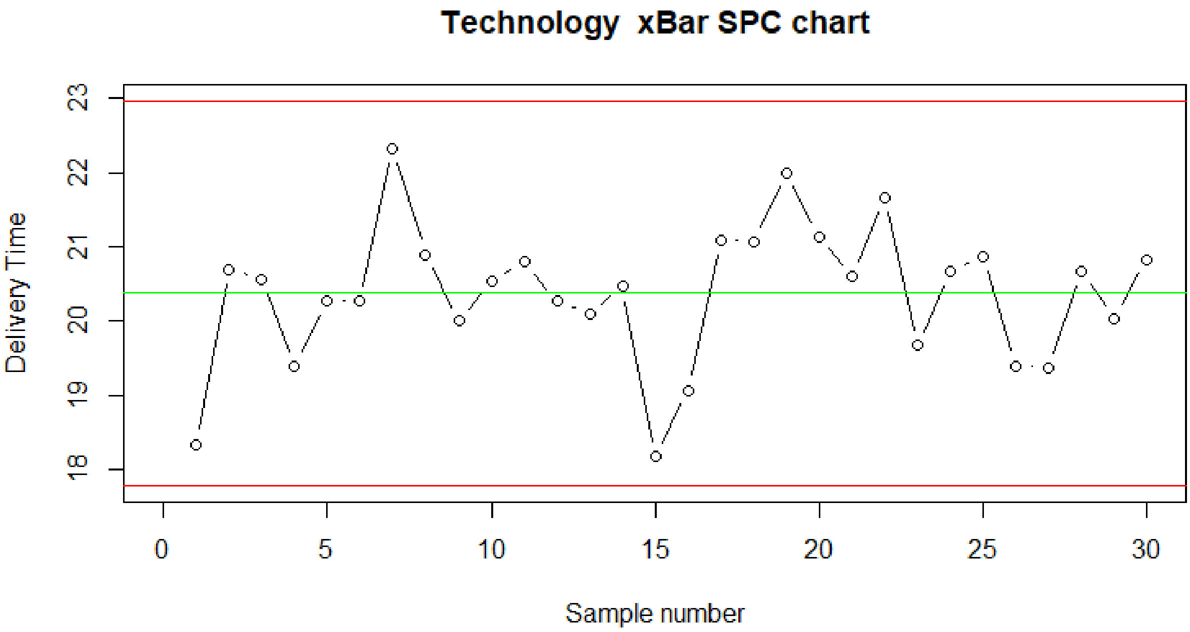

Figure 4.2 Technology X-bar chart for the first 30 samples

From both the X-bar and S-chart plots of the first 30 samples, it can be observed that all he samples used during initialization are in control, since they lie within the control limits. This means that all calculations of the sample mean and sample standard deviations can be based off the control limits calculated during this initialization phase since all the samples are in control. Had they not been in control, the erroneous samples would have to be removed, and the CL and control limits recalculated. An example of this case can be observed in the S-chart of food items, shown in figure 4.3 below. The sample whose value lies outside the UCL will have to be eliminated, and the control limits recalculated.

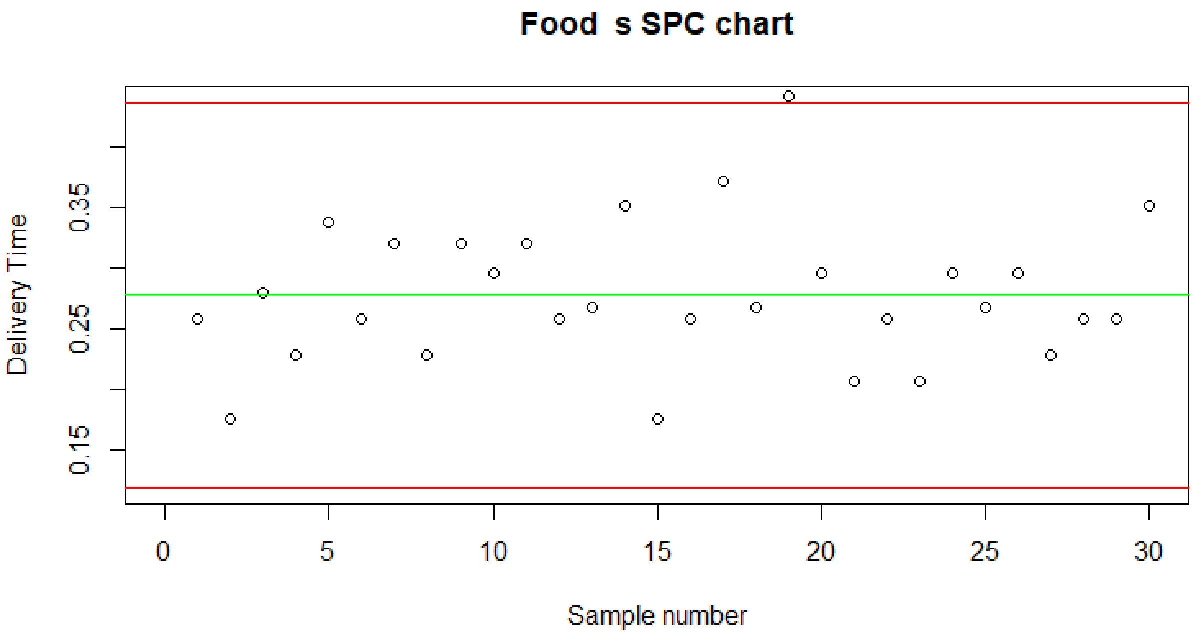

Figure 4.3 Sample value outside UCL.

X-bar and S-charts will then have to be created for the remainder of the delivery times for each product after initialization. Using technology items as an example, the X-bar and S-charts of all the samples are shown in figures 4.4 and 4.5 below.

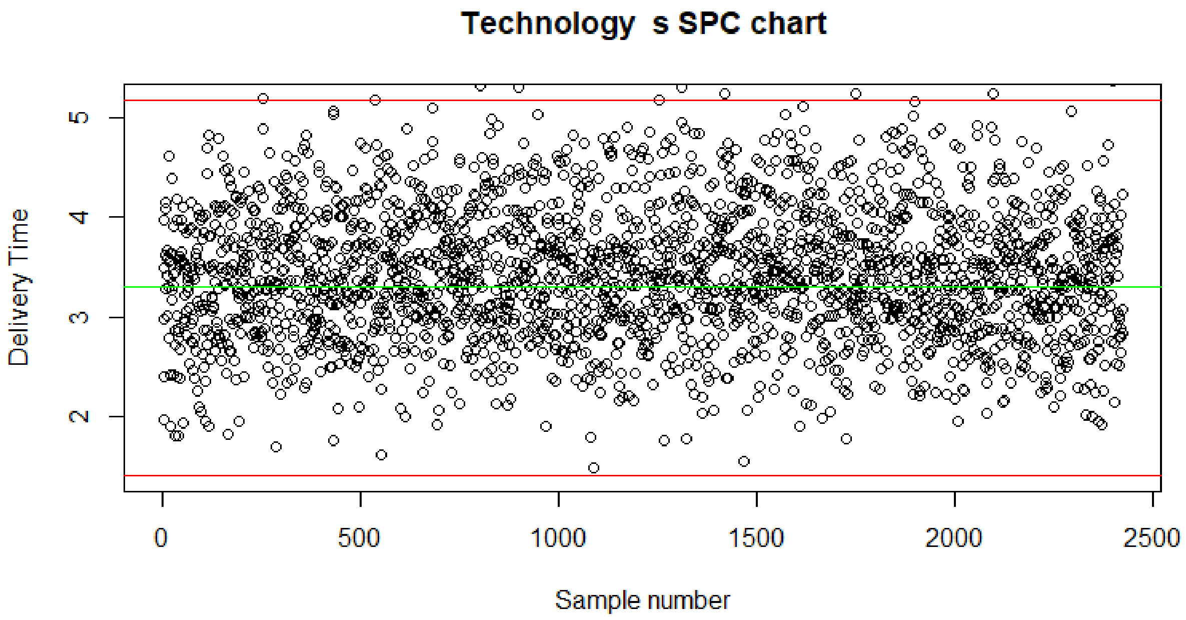

Figure 4.4 Technology S-chart for all the samples

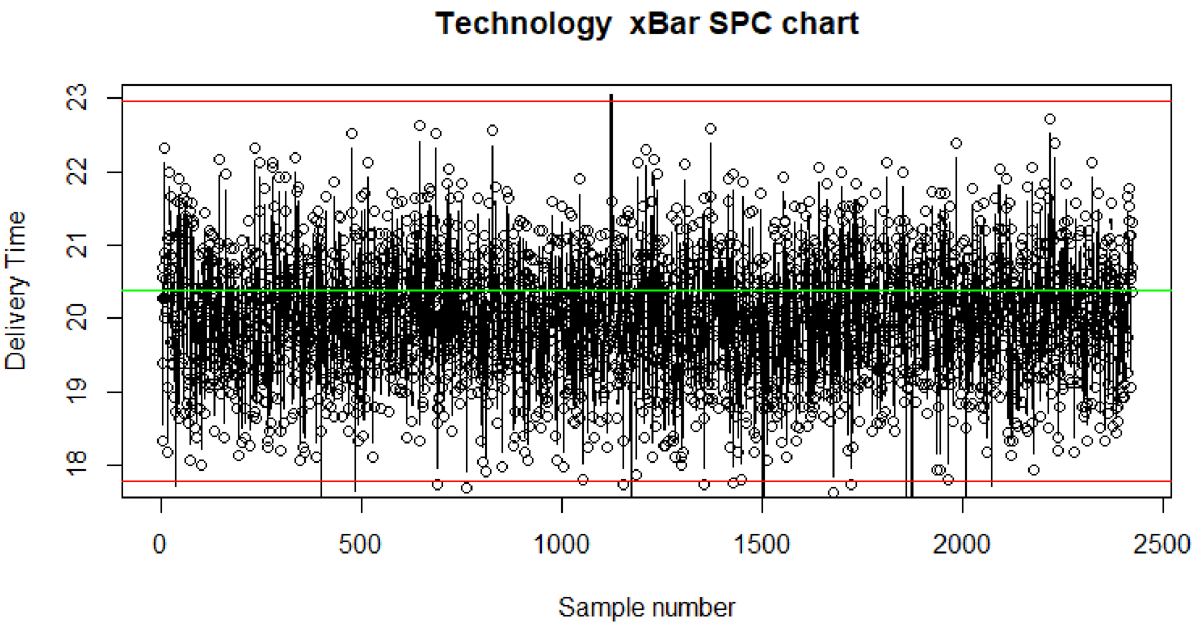

Figure 4.5 Technology X-chart for all the samples

It can be observed from the charts that some of the values are out of control, i.e. they lie outside the control limits. This shows that in conclusion, the process is not stable. Root cause analysis has to be conducted to find the causes of the variable delivery times, which in turn can result to a decrease in the reliability of the company's delivery of technology products, which could decrease sales as technology products are the second most frequently bought products as was seen in table 3.2.1.

The company must aim to ensure that the delivery times for all the products lie within the control limits. If they lie within the control limits, it means mean that the delivery times are stable (shown by the S-chart), hence, the process can remain as is – i.e., the process does not need to be optimized. An X-bar plot with values all lying within the control limits shows that the means of the delivery times can be predicted. Customers can have an idea of when they can expect their package. If the company's delivery date estimates are reliable, this increases customer satisfaction and potentially increases sales as current customers spread the word of the company's good service. This is known as word-of-mouth marketing and is highly effective, as was seen from the graph in figure 3.3.1, where recommendation was seen as the best marketing strategy (Hayes, 2022).

The statistical interpretations of the X-bar plots are summed in the table below.

		X-chart					
Class	**UCL**	**U2σ**	**U1σ**	**CL**	**L1σ**	**L2σ**	**LCL**
Technology	22.9996	22.1245	21.2495	20.3744	19.4994	18.6244	17.7493
Clothing	9.3939	9.2526	9.1113	8.9700	8.8287	8.6874	8.5461
Household	50.2880	49.0460	47.8041	46.5622	45.3203	44.0784	42.8365
Luxury	5.4718	5.2264	4.9810	4.7356	4.4901	4.2447	3.9993
Food	2.6982	2.6288	2.5594	2.4900	2.4206	2.3512	2.2818
Gifts	9.4915	9.1147	8.7379	8.3611	7.9843	7.6075	7.2307
Sweets	2.8868	2.7504	2.6141	2.4778	2.3414	2.2051	2.0688

Table 4.1.1 X-bart statistics

The statistical interpretations of the S-chart plots are shown in the table below.

		S-chart					
Class	**UCL**	**U2σ**	**U1σ**	**CL**	**L1σ**	**L2σ**	**LCL**
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	0.4372	0.3842	0.3312	0.2781	0.2251	0.1721	0.1191
Gifts	2.2460	1.9737	1.7013	1.4290	1.1566	0.8842	0.6119
Technology	0.8352	0.7340	0.6327	0.5314	0.4301	0.3288	0.2275

Table 4.1.2 S-chart statistics

5. Optimising the delivery process

In order to optimize the delivery process, run tests were carried out on the X-bar and S-charts. These tests are used to determine unnatural patterns in the delivery times which are as a result of assignable causes. These causes can then be investigated to optimize the delivery times.

5.1 Delivery times outside the control limits

An out-of-control run test was carried out on the sample means. To do this, the control charts were each divided into six equal zones (three on each side of the centre line). Since each control limit (UCL or LCL) is three sigma limits (three standard deviations of the mean) in width, it means that each zone is one sigma wide.

Table 5.1.1 below shows the sample means of the delivery times outside the upper zone (UCL).

Class	Number of samples outside the outer control limits	First 3 samples above UCL			Last 3 samples above UCL		
Technology	19	31	179	377	1856	1961	2298
Clothing	17	129	157	316	1296	1475	1691
Household	192	6	12	14	1305	1306	1307
Luxury	442	3	5	6	753	757	758
Food	13	45	213	261	1130	1356	1505
Gifts	2579	1	2	3	2577	2578	2579
Sweets	5	336	486	625	1051	1328	

Table 5.1.1 Out-of-control test results

The out-of-control test indicates a shift in the delivery time average, since it was carried out on the X-bar charts. In the table above, the samples outside the upper zone are summarised, with the first consecutive 3 samples and the last consecutive 3 samples in the zone listed.

It can be noted that the product type 'gifts' has the highest number of samples outside the upper zone, followed by luxury products. This means that the delivery times of these products are too high – customers have to wait very long hours to receive their package. Investigations have to be carried out on these products so that bottlenecks in the production process can be identified and rectified to shorten their delivery times.

Sweets have the least number of samples outside the upper zone. This means their delivery times still need optimizing but to a lesser extent than gifts and luxury items.

5.2 Sample standard deviations

Table 5.2.1 below shows the most consecutive samples of sample standard deviations between -0.3 and +0.4 sigma-control limits

Class	Most consecutive samples between -0.3 and +0.4	Last sample given in the range
Technology	11.0	518.7
Clothing	16.2	600.9
Household	8.2	750.9
Luxury	1	2.7
Food	12.2	1564.9
Gifts	0	0
Sweets	14.2	698.9

Table 5.2.1 Sample standard deviations between -0.3 and 0.4 sigma control limits

5.3 Estimating the likelihood of getting a Type-I error (Manufacturer's error)

A Type-I error is the probability that in the SPC calculations, a sample was classified as erroneous (delivery time falling outside the control limits) when there was nothing wrong with the process. Thus, the predictions were wrong about the delivery time being unstable. In contrast, a Type-II error (Consumer's error) is the probability that a sample is classified as in control (delivery time lying within the control limits) and yet the sample is actually out of control. Type-I and Type-II errors allow the accuracy of the model used in the SPC calculations to be determined. Figure 5.3.1 below shows a summary of Type-I and Type-II errors.

	Process is fine	Process is not fine
SPC indicated the Process is not fine	Type I Error or Manufacturer's Error	Correct to fix process
SPC indicated the Process is fine	Correct to do nothing	Type II Error or Consumer's Error

Figure 5.3.1 Type-I and Type-II errors

The probability of making a Type-I error was found to be 0.002699796. This reveals that the samples classified as out-of-control the SPC analysis are reliable, as there is only an approximately 0.27% chance that they were erroneously classified.

5.4 Estimating the likelihood of getting a Type-II error (Consumer's error)

The probability that the delivery time of a technology item was erroneously classified as lying within the control limits of the X-bar plot whilst it was actually out-of-control was calculated and found to be 0.9973002. This is a high error, showing that there may have been systematic and random errors in the data set being analysed, and this needs to be investigated further. Increasing the sample size during the SPC analysis could also help reduce this error.

5.5 Loss in sales (case study – technology items)

An analysis was conducted on the delivery times of technology items to calculate the number of delivery hours that will result in maximum profit generation. This calculation is based off the potential loss in sales if technology items are delivered slower than 26 hours. To do this analysis, the following were assumed:

- 5.5.1 R329.00 is lost per late-hour delivery for each item
- 5.5.2 It costs R2.50 per hour to reduce the delivery time for each item by one hour.

The results of this analysis are shown in figure 5.5.1 below.

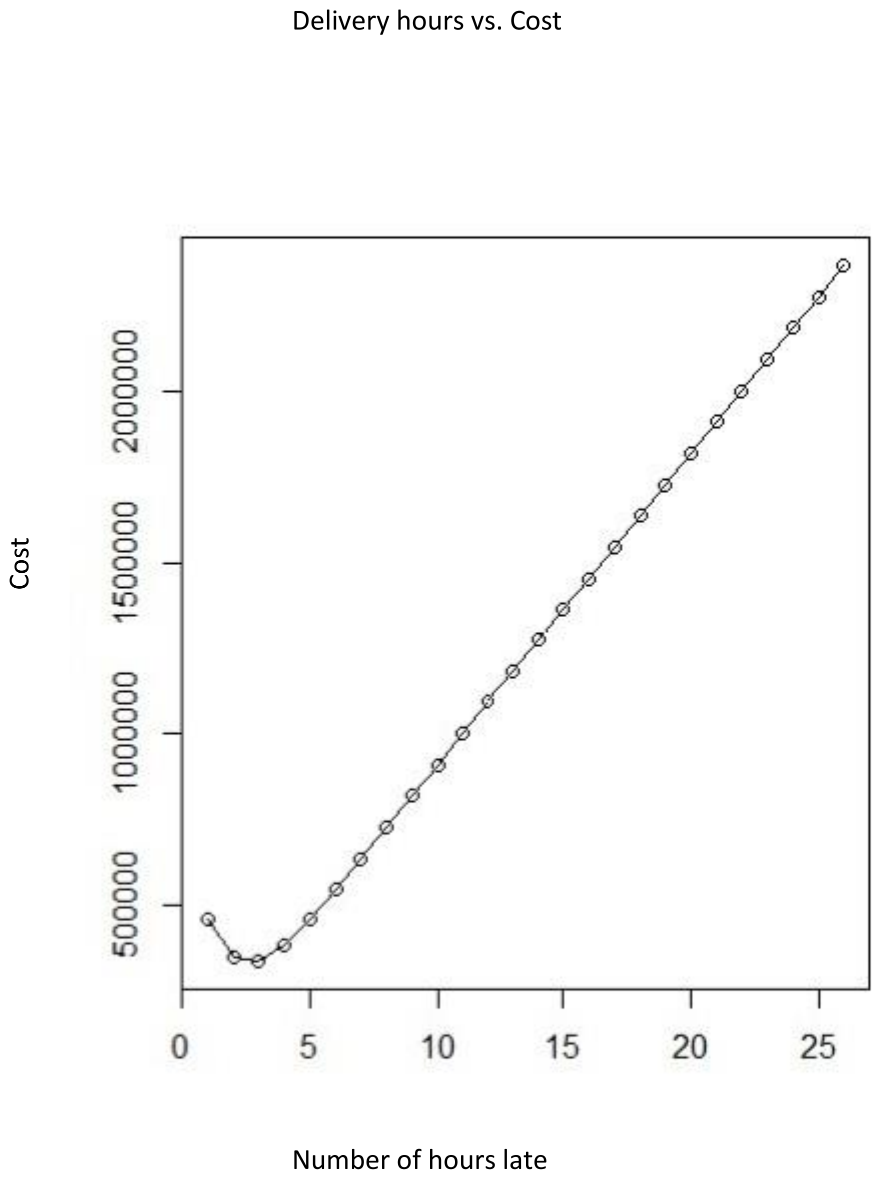

Figure 5.5.1 Loss of sales for technology items due to late delivery

The graph above shows that there is generally a proportional relationship between the delivery hours and the cost. As the number of delivery hours increases, the loss of sales (increase in costs of production) also increases, reducing profit – which can be seen in the delivery time range of 5 – 26 hours. The lowest production cost is found to be about R125 000.00, and it occurs when the delivery time is about 2.5 hours. This means that in order to ensure maximum profit, the company should aim to deliver technology items within 2.5 hours from when an order is generated from a customer.

6. MANOVA analysis

A multivariate analysis of variance (MANOVA) was carried out in order to relate marketing strategies that the company uses to the age of the customers. This will allow the company to identify the marketing strategy that best suits each age group, to streamline sales for maximum profit. The MANOVA analysis was used since multiple dependent variables were investigated, i.e., there are 6 different marketing strategies that the company uses. Figure 6.1 below shows a box-plot that is a result of this analysis.

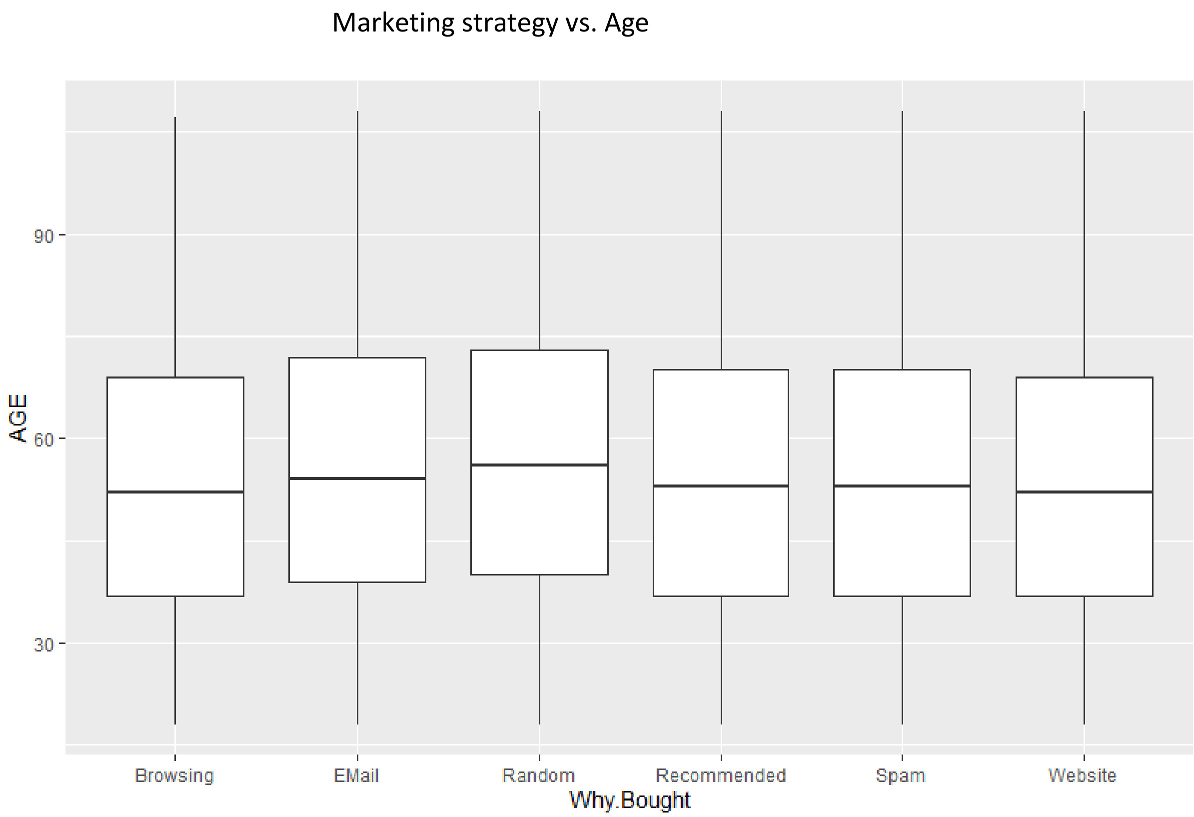

Figure 6.1 How the marketing strategy correlates to each age group

From the plot above, it can be observed that the data set has a perfect normal distribution, as shown by symmetry of the box-plots (each quartile has the same length). Almost all the marketing strategies correlate to a median age of about 52.5 years (rounded up to 53 years to the nearest whole number). The minimum age of a customer is about 7.5 years (rounded up to 8 years to the nearest whole number); and the maximum age is about 112.5 years (rounded up to 113 years to the nearest whole number).

The plot shows that there is little correlation between the marketing strategy and the age – a customer in any age group in the range given is likely to buy a product from the company regardless of the marketing strategy used. However, customers that are about 57 years old are more likely to randomly buy products, as the median of the “Random” marketing strategy is slightly offset from the rest.

7. Taguchi Loss

The Taguchi loss function defines products that are no longer acceptable because they fall outside of specification. The Taguchi function was used to determine the quality that food deliveries need to be kept at during transit, i.e., food deliveries need to be kept cool during transportation.

Taguchi loss was used to determine the tolerance for the thickness of a refrigerator that is used to transport food deliveries. The blueprint specification for the thickness of a refrigerator part was stated to be 0.06 ± 0.04 cm (Evans et al., 2020). It also costs $45.00 (R816.33) to scrap a part that is outside the specifications (Evans et al., 2020). The quality of the food deliveries decreases from the moment the refrigerator thickness deviates from this specification. The Taguchi loss function was found to be: $L = 28125(y-0.06)^2$, where L = Taguchi loss and y = actual thickness of the refrigerator.

If the scrap cost is reduced to $35.00 (R635.82), the new Taguchi loss function becomes $L = 21875(y-0.06)^2$. From this loss function, if the process deviation from the target can be reduced to 0.027cm, the Taguchi loss is found to be $L = 21875(0.027)^2 = 15.94687$ (Evans et al., 2020).

8. System reliability

Magnaplex – a subsidiary company, manufactures some of the technology items that the company sells. Analysis was done to determine the reliability of their current production process in the case of failures, as Magnaplex currently uses identical machines as back-up. The reliability block diagram in figure 8.1 below demonstrates the interconnection between the different machines in Magnaplex's current production process.

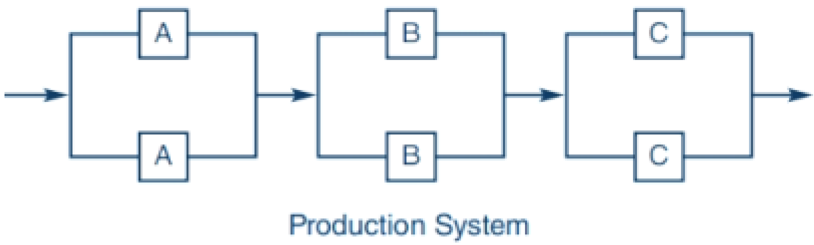

Figure 8.1 Magnaplex's current production process

Table 8.1 below shows the reliability of each machine in the production process (Evans et al., 2020).

Machine	Reliability
A	0.85
B	0.92
C	0.90

Table 8.1 Reliability of each machine

Two scenarios were analysed. The first case was if only one machine in each of A, B, and C is working. In this case, the reliability is $R_A*R_B*R_C = 0.85*0.92*0.90 = 0.7038$. The 70.38% reliability is high, which means that Magnaplex's current production process still functions even in the event that only one machine works at each stage.

The second case is if there are two machines working at each stage. Since the components are parallelly connected, the reliability of the process at each stage is 1 – probability that both machines fail. The reliability of the production process in this case was calculated as follows:

Reliability = $(1-(1-R_A)^2)*(1-(1-R_B)^2)*(1-(1-R_B)^2) = (1-(1-0.85)^2)*(1-(1-0.92)^2)*(1-(1-0.90)^2) = 0.9615$

The calculation shows that the system reliability increases significantly if both machines operate at each stage. It can thus be concluded that Magnaplex is a reliable supplier for the company.

9. Reliability of the delivery times

The reliability of the delivery times depends on the number of delivery vehicles available, as well as the number of drivers available. The reliability of the delivery times and the drivers was calculated based on 1560 working days. Table 9.1 below shows the availability of delivery vehicles during this time frame. The company currently has 20 delivery vehicles, of which 19 are required to be operating at any time to give reliable service (Project Description, 2022).

Number of vehicles available	**Number of days (out of 1560) that specified number of vehicles were available**
17	1
18	3
19	22
20	190

Table 9.1 Vehicle availability

Table 9.2 below summarises the availability of drivers in the 1560 days. The company currently has 21 drivers, who each work an 8-hour shift per day (Project Description, 2022).

Number of drivers available	**Number of days (out of 1560) that specified number of drivers were available**
18	1
19	6
20	95

Table 9.2 Driver availability

Using binomial distribution, it was computed that in a year (365 days), the number of days where the service will be reliable because of vehicle availability is 361.49 day. In addition, out of 365 days, drivers will be available in 364.24 days. Thus, the total number of days where the service can be expected to be reliable is 360.737 days, rounded up to 361 days (to the nearest integer).

An analysis was further conducted to compute the number of days when reliable service can be expected if the number of vehicles is increased by one. Using binomial distribution, it was computed that there would now be 364.84 days of vehicle reliability.

Combined with the availability of drivers, the number of days when reliable service can be expected overall is 364.08 days, rounded to 364 days (to the nearest whole number). Thus, increasing the number of vehicles by one greatly increases the number of days where reliable service can expected. Thus, the company should purchase an additional delivery vehicle.

10. Conclusion

This report showed that statistical analysis can enable a company to identify bottlenecks in their supply chain and find ways to optimize them. In the report, the reliability of different statistical methods of analysis were also computed to determine the accuracy of the results being interpreted.

From the report, it can be concluded that the company needs to optimize the delivery times. Currently, the delivery times are too long and this causes loss of sales. It is recommended that the company further investigates the reasons behind the lengthy delivery times so as to better streamline sales.

11. References

Evans J. and Lindsay W.M. 2020. *Managing for Quality and Performance Excellence*. 11th Edition. Cengage Learning Inc., Boston, USA.

Hayes A. 2022. Word-of-Mouth Marketing: Meaning and Uses in Business. [Online]. Available: https://www.investopedia.com/terms/w/word-of-mouth-marketing.asp#:~:text=Word%2Dof%2Dmouth%20marketing%20is,products%20and%20share%20positive%20experiences. [20 October 2022]

Project Description. 2022. Preamble to the Engineering Counsel of South Africa (ECSA) report that proves graduate attribute 4 (ECSA GA4) in 2022. [Online]. Available: https://learn.sun.ac.za/course/view.php?id=70723 [20 October 2022]