Quality Assurance Engineering Counsel of South Africa (ECSA) Project

J.H.L. Strauss

Student number: 22870865

Quality Assurance 344

21 October 2022

Table of Contents

Table of Tables

Table of Figures

Table of Equations

Introduction

In this report an online shopping company would be evaluated on a few key areas and would be advised on areas where we identified possible problems. The client's data analysed is from an online business, the items sold in the online store can be divided into 7 category's that is as follows Clothing, Food, Gifts, Household, Luxury, Sweets and Technology. The grunt of the analysis is done on each category as a whole, this analysis is aided and partly done on a program called R Studio. Using this program, I created certain visual effects to help understand and identify problems. The goal of the report is to give simple and understandable feedback on the company's sales data. In the following contents problematic areas would be addressed and logical and simple explanations will be given on how to improve processes to make a larger profit, as this is the company's primary goal.

Part 1: Data Wrangling

During data wrangling we took the original dataset provided by the online business and analysed it, the dataset contains 180 000 observations of 10 variables. All the rows containing NA and negative values where extracted, this extracted data is stored in a variable called Incomplete data whilst the remaining data were stored in a variable called Valid data. There are 22 observations of Incomplete Rows, the rest is Valid Data. This Step is taken to have a more accurate result when the data analysis is performed.

Part 2: Descriptive Statistics

Delivery Times for the different classes

Taking a closer look at the following graphs we can see that technology and household items would take the longest time to be delivered, these features have mean delivery times of 45 and 17.5 respectively. This could be due to the large fields of technology and household items, meaning these items could vary greatly in size and complexity. Thus, the products do not cater to any one type of transport, this can be one of the aspects that contributes to the large delivery time. Food and other items like sweets have a small delivery time because of them being perishable goods and highly in demand. The graphs all seem to be normally distributed except for the gifts feature that seem to be slightly skewed to the right.

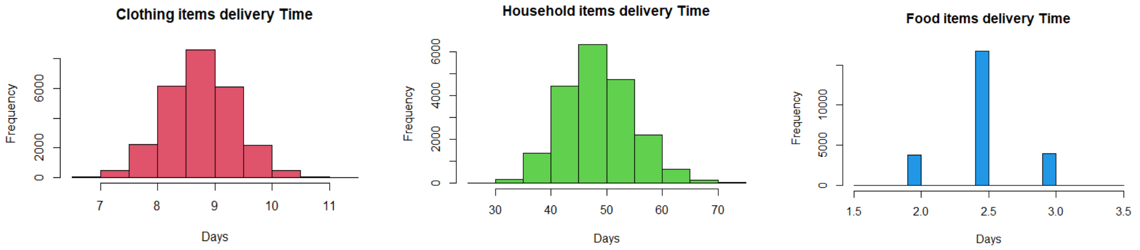


Figure 1 Bar plots of Clothing, Household and Food items with reference to delivery times

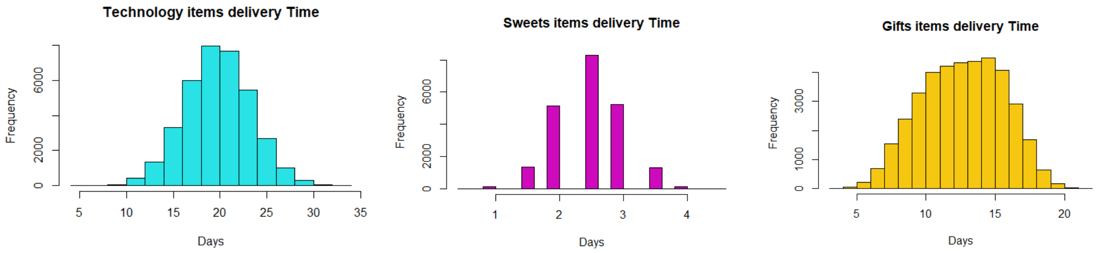


Figure 2 Bar plots of Technology, Sweets and Gift items with reference to delivery times

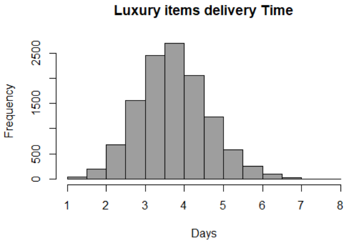


Figure 3 Bar plots of Luxury items with reference to delivery times

The age of customers vs. different classes

Since Age of customers for different classes is a very important and valuable feature within the company, given that it shows us areas for future expansion and improvement. With this information in hand the company can select a target age for each class and advertise and sell accordingly. As seen in the figure below Clothing and technology tends to be bought by younger customers while Food and gifts tends to be bought by the older portion of the customers. The large box of sweets items tells us that it is bought by the largest proportion of the age group.

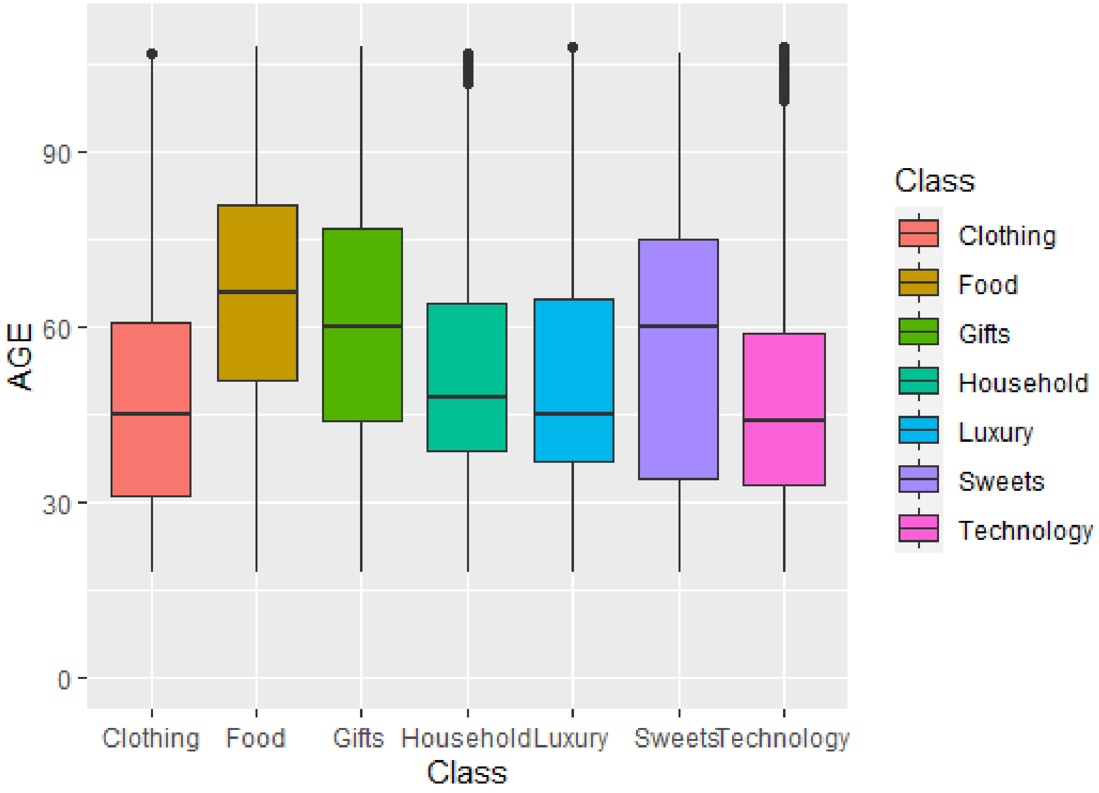
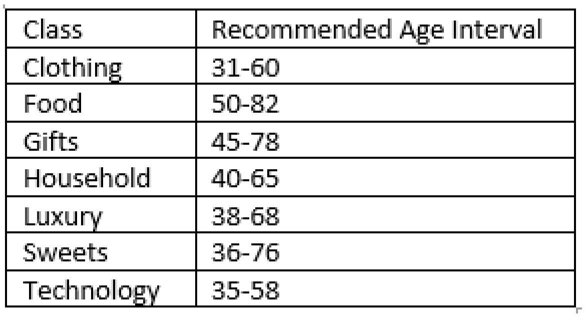


Figure 4 Box Plot of age with reference to Class

Table 1 Recommended Age Range per class



Class	Recommended Age Interval
Clothing	31-60
Food	50-82
Gifts	45-78
Household	40-65
Luxury	38-68
Sweets	36-76
Technology	35-58

Table 1 above represents the recommended age range per feature, deduced from Figure 4 above.

Marketing methods that led customers to purchase products

In figure 5 below we can clearly see that the email and spam is the worst method of advertisement used by the company, while recommendations are by var the most effective way of additional marketing followed by the company's website. This figure is useful since it gives us a representation of the effectiveness of the company's marketing methods and gives an insight into how to improve our overall marketing scheme.

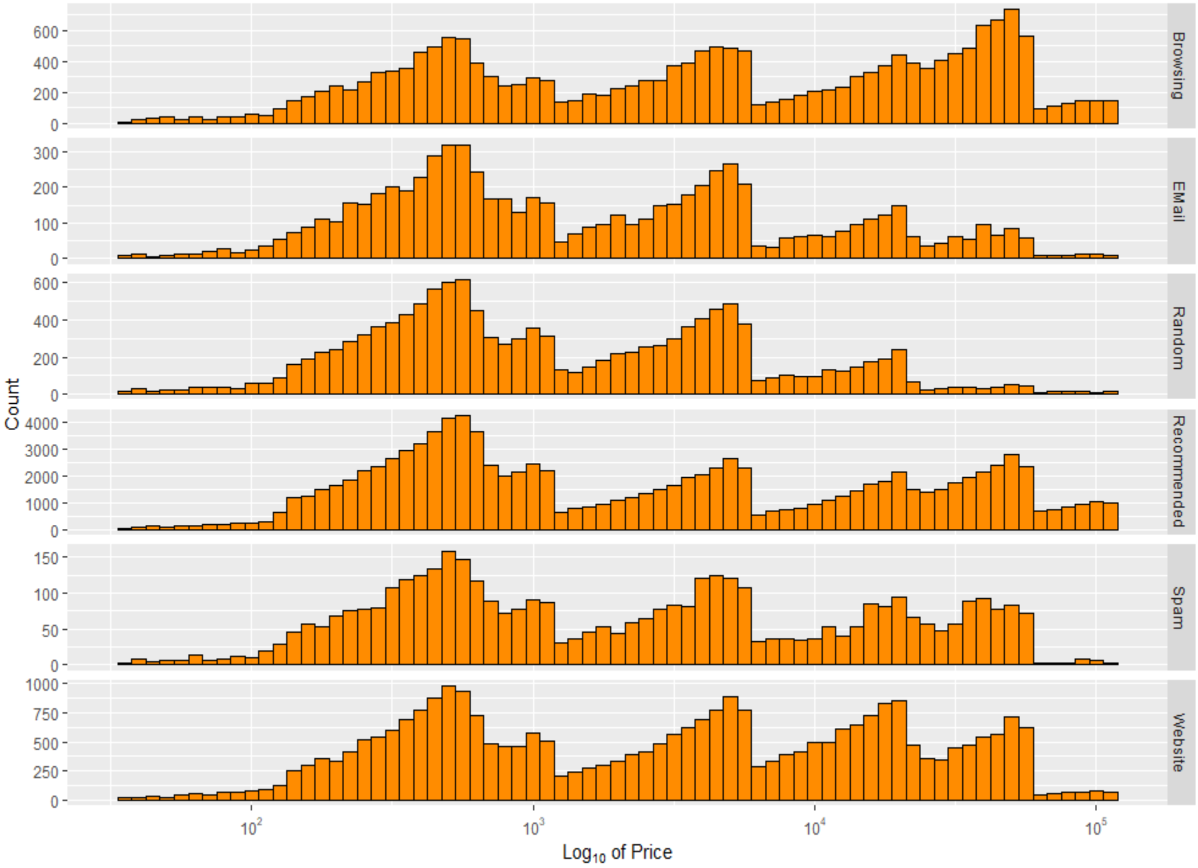


Figure 5 Bar plot of sales according to how costumers heard of the company

Price Range of each Class

In figure 6 below price range of each class of items and the number of units sold is presented. This is valuable information since it gives an indication of what item sold, generates the largest part of our total income. Taking a closer look at the figure we see that Technology and Household class items despite having a large price is also sold in relatively large volumes compared to items such as Clothing, food and Gifts.

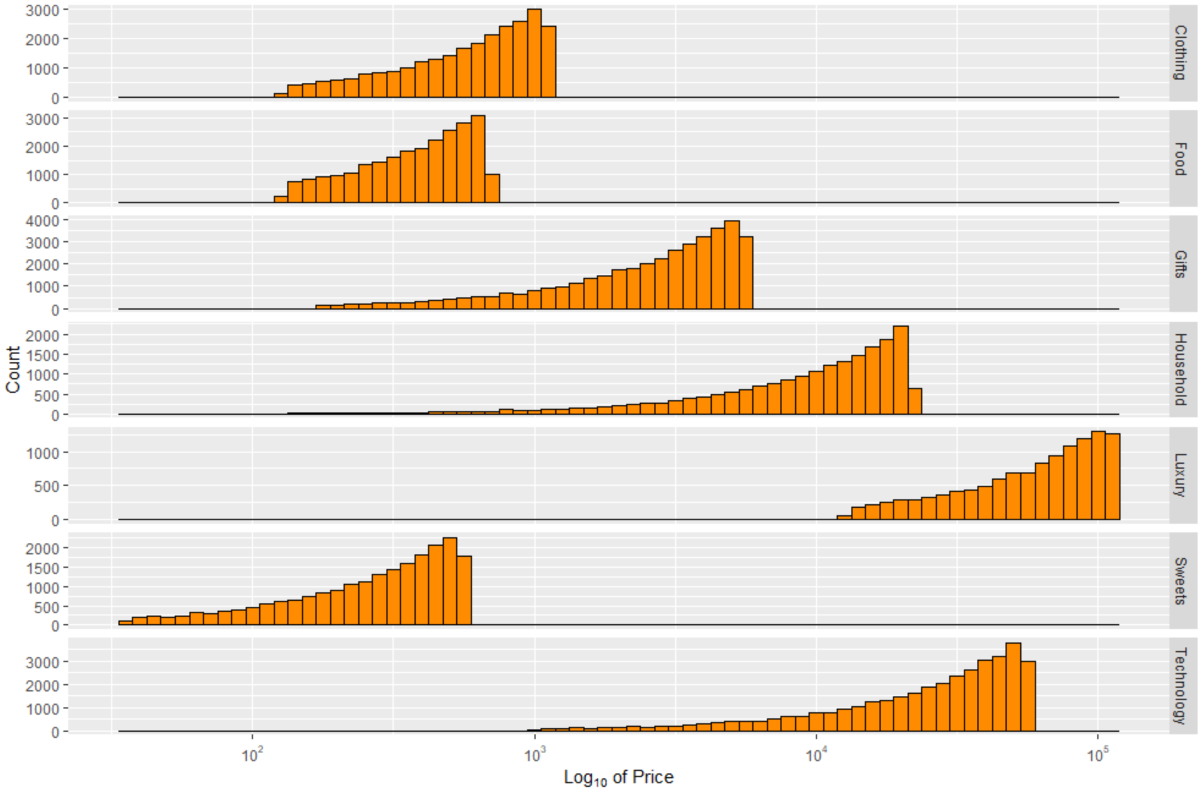


Figure 6 Bar plot of sales according to Class of items

Process Capacity

The formulas in Equation 1 below will be used to calculate the process capacity for the company's delivery times of Technology class items.

Equation 1

$$C_p = \frac{USL - LSL}{6\sigma}$$

$$C_{pu} = \frac{USL - \mu}{3\sigma}$$

$$C_{pl} = \frac{\mu - LSL}{3\sigma}$$

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

The given values form the company for USL (Upper specification limit) an LSL (Lower specification limit) is 24 and 0 Hours, LSL of 0 is logical since delivery times would be able to range from 0-24 hours.

The calculated mean (μ) for technology class items is equal to 20.011.

The calculated standard deviation (σ) for technology class items is equal to 3.502.

<u>Now we can calculate Process capacity values as follows</u>

Cp = 1.142204

Cpu = 0.3796878

Cpl = 1.904721

Cpk = 0.3796878

The Cp of 1.142204 indicates that the process in not perfectly centred but still considered process compatible.

The Cpu and Cpl indicates the process capability based on upper and lower specification limits respectively.

The Cpk value of 0.3796878 a small value but not quite negative, where a negative Cpk value would indicate that Technology class products are not being delivered in the specified time, whilst a Cpk of 1 would indicate that marginally or barely capable of delivering products in the specified time, while Cpk of 2-3 would be capable of on time deliveries with low to no exceptions. This is very useful since the company can address the issue at hand.

Part 3: Statistical process control (SPC)

Analysis of the first 30 Samples.

The fist step is to use our first 30 samples of sample size 15 to determine the centre lines, outer control limits, 2-sigma-control and 1-sigma-control limits for X&S-charts to aid in plotting X&S-charts.

In the two plots below, we can see that there is one sample in sweets and food respectively that is above the UCL line, following the SPC process we remove these values from the calculations of UCL and LCL.

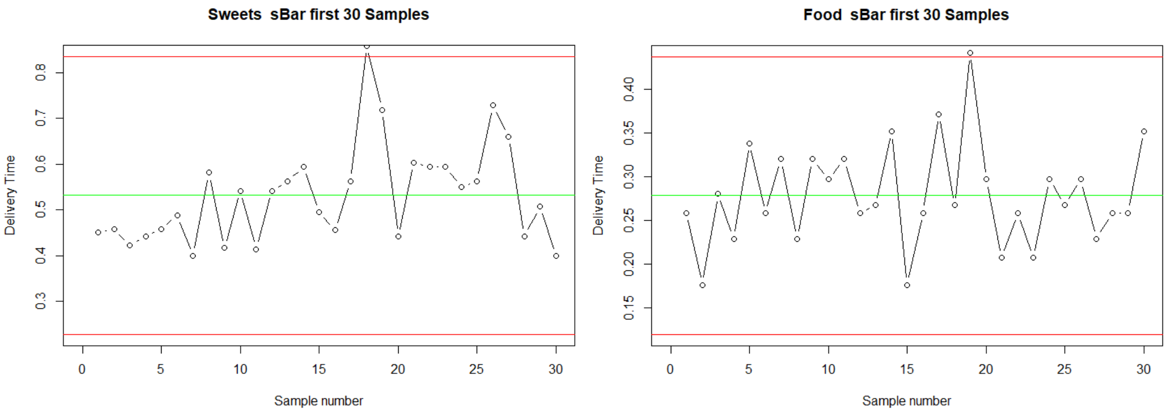
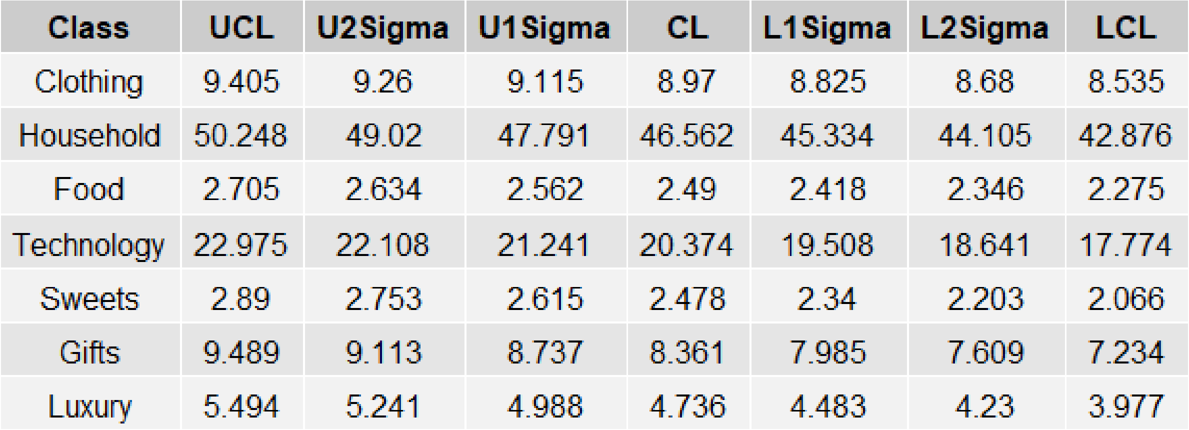


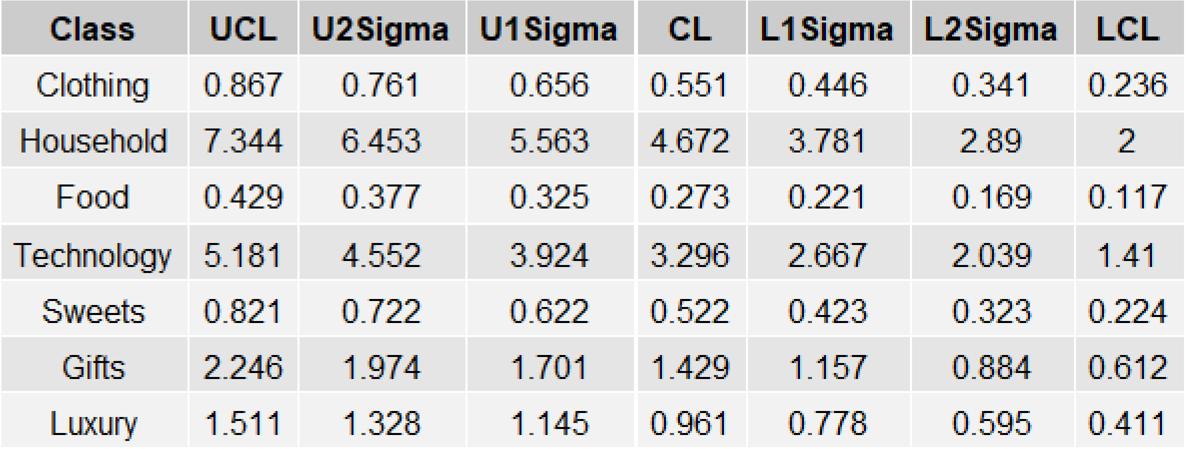
Figure 7 Identifying Outliers in X&S chart of first 30 Samples

Table 2 SPC values for X-charts



Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	9.405	9.26	9.115	8.97	8.825	8.68	8.535
Household	50.248	49.02	47.791	46.562	45.334	44.105	42.876
Food	2.705	2.634	2.562	2.49	2.418	2.346	2.275
Technology	22.975	22.108	21.241	20.374	19.508	18.641	17.774
Sweets	2.89	2.753	2.615	2.478	2.34	2.203	2.066
Gifts	9.489	9.113	8.737	8.361	7.985	7.609	7.234
Luxury	5.494	5.241	4.988	4.736	4.483	4.23	3.977

Table 3 SPC values for S-charts



Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	0.867	0.761	0.656	0.551	0.446	0.341	0.236
Household	7.344	6.453	5.563	4.672	3.781	2.89	2
Food	0.429	0.377	0.325	0.273	0.221	0.169	0.117
Technology	5.181	4.552	3.924	3.296	2.667	2.039	1.41
Sweets	0.821	0.722	0.622	0.522	0.423	0.323	0.224
Gifts	2.246	1.974	1.701	1.429	1.157	0.884	0.612
Luxury	1.511	1.328	1.145	0.961	0.778	0.595	0.411

Because of the S-charts using sample standard deviations, it creates an illusion of the features being more in control. Therefore, we prefer to make conclusions of the X-charts. The S-charts are still given and given a can be used for further investigation but are not as problematic as the X-charts.

X-Bar SPC charts for all the Samples

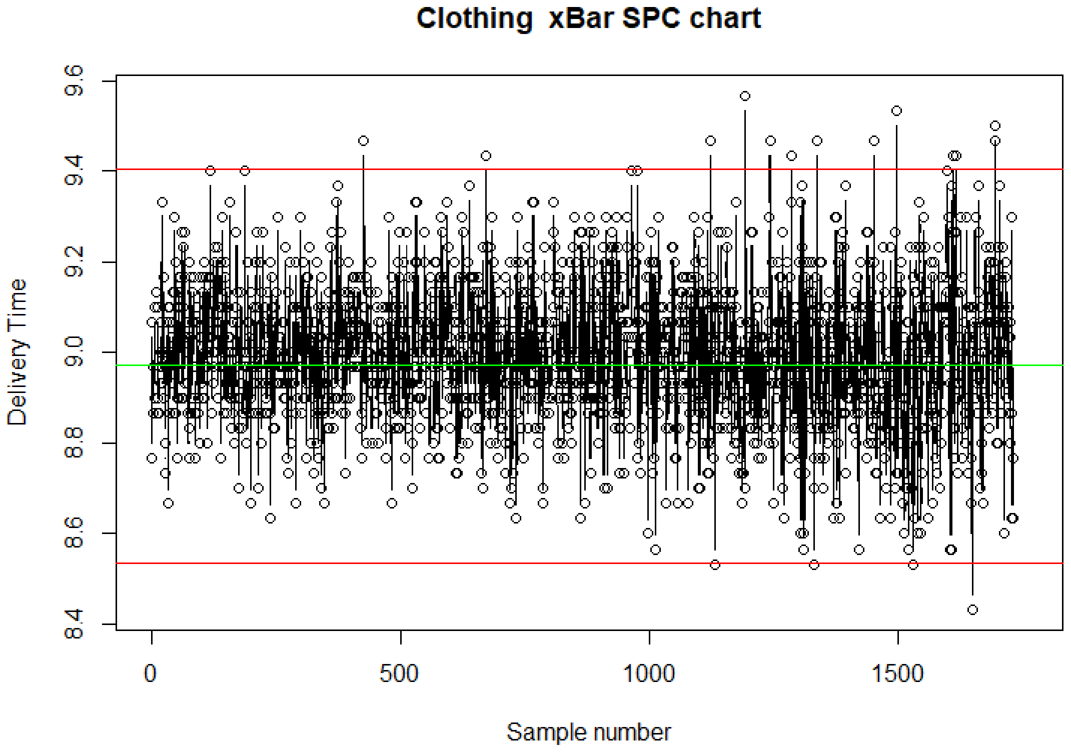


Figure 8 X-chart for Clothing Items

Figure 8 above represents x bar chart for “Clothing” feature, according to the graph this feature seems to be in control, except a few outliers.

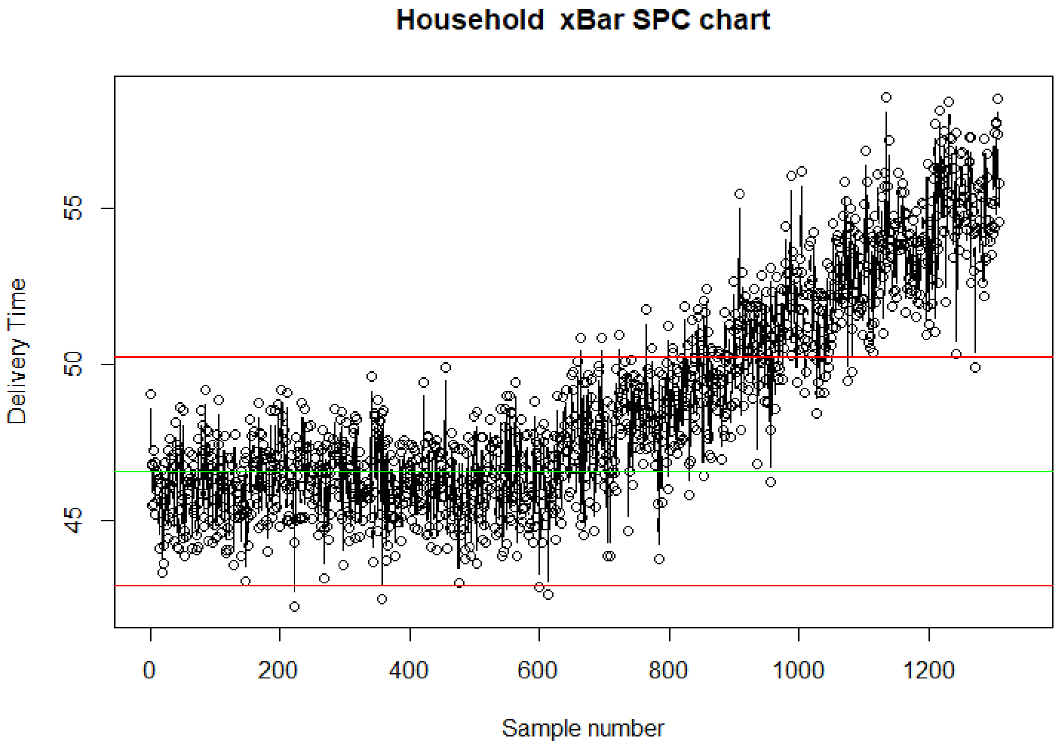


Figure 9 X-chart for Household Items

Figure 9 above represents x bar chart for “Household” feature, according to the graph this feature is definitely out of control. The tendency of the samples gradually increase in delivery time is easily noticeable, this should be an area of concern for the company since they are not delivering in there promised time and thus. The problem could be caused by the large variety in sizes of household products, ether way this is a problem that needs to prioritised.

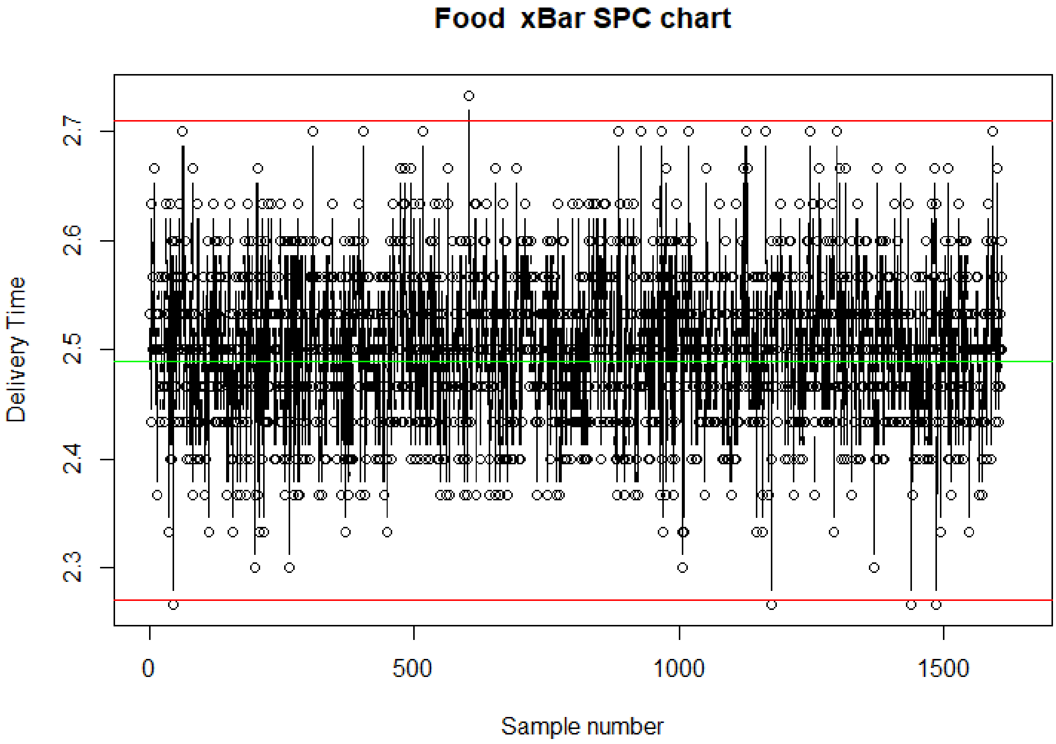


Figure 10 X-chart for Food Items

Figure 10 above represents x bar chart for “Food” feature, according to the graph this feature seems to be in control, except a few outliers.

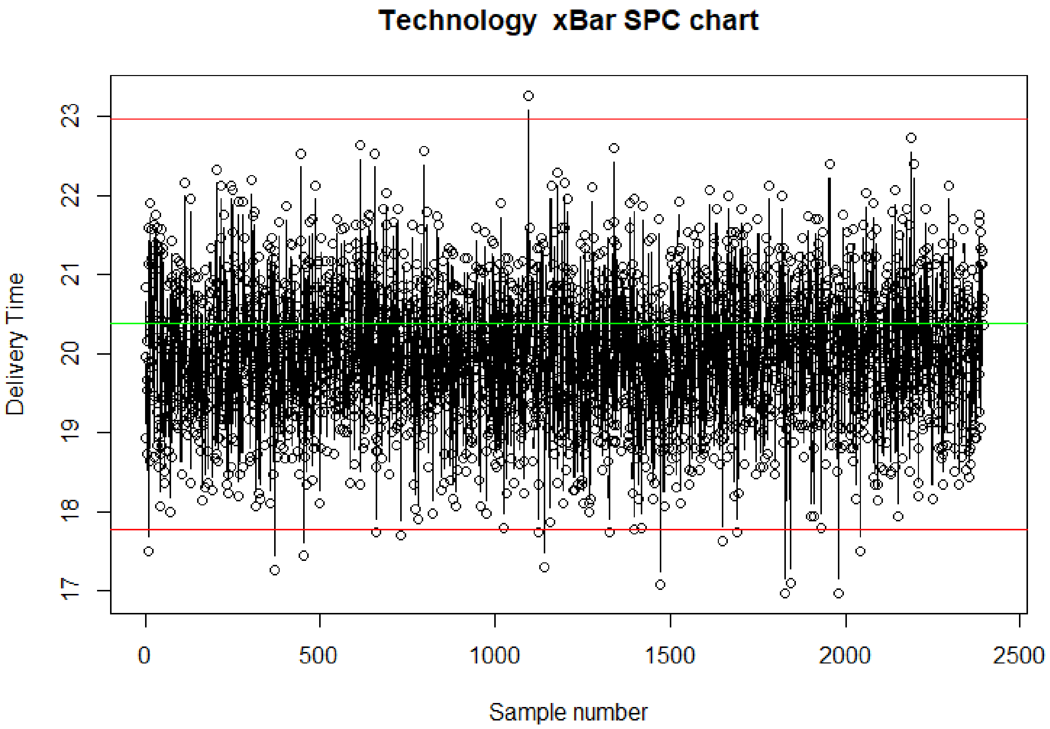


Figure 11 X-Chart of Technology class items

Figure 10 above represents x bar chart for "Technology" feature, according to the graph this feature seems to be in control, except a few outliers.

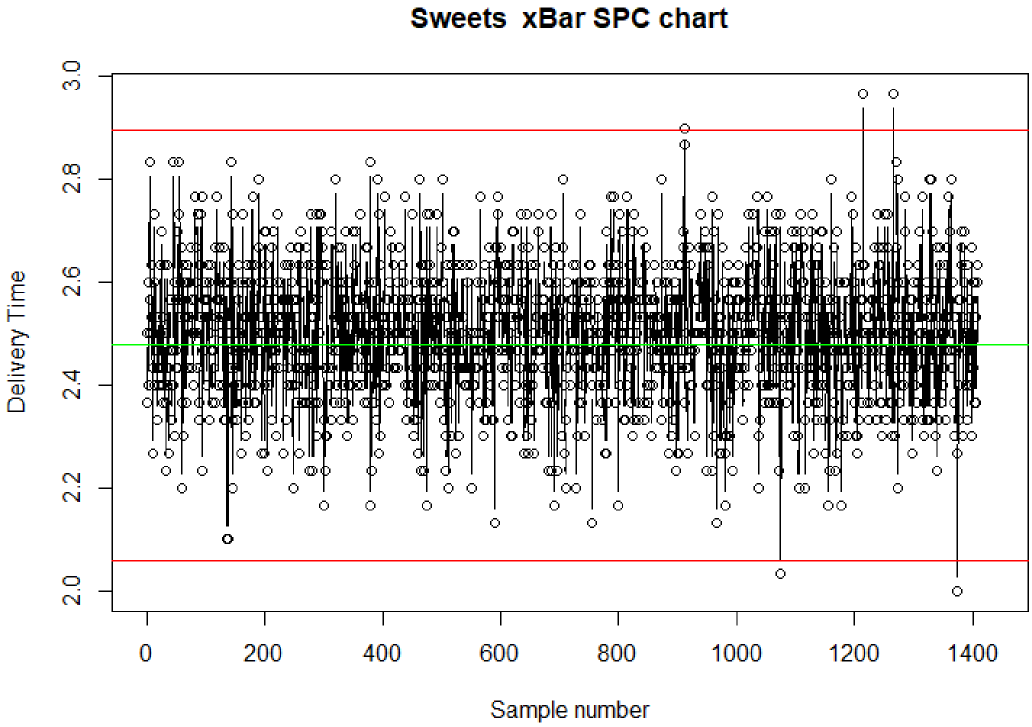


Figure 12 X-Chart for Sweets

Figure 11 above represents x bar chart for "sweets" feature, according to the graph this feature seems to be in control, except a few outliers.

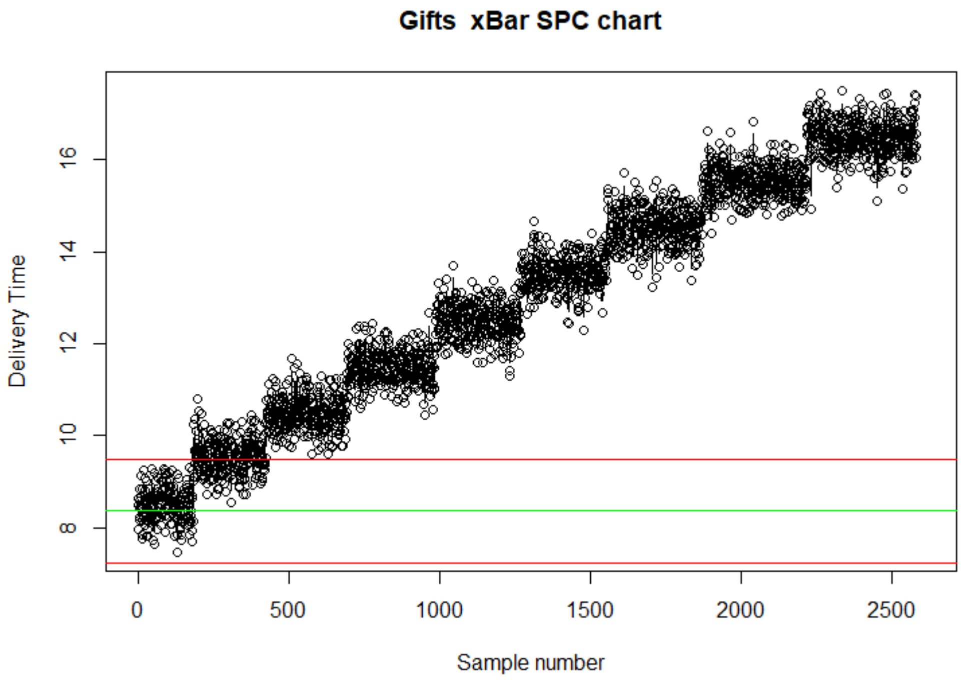


Figure 13 X-Chart for Gifts

Figure 13 above is a representation of X-bar or mean delivery times for the “Gifts” feature, this is also n feature of great concern. This feature is clearly out of control with an incremental growth in delivery times from a mean value of 8.5 hours to a staggering 16.5 at 2500’th sample.

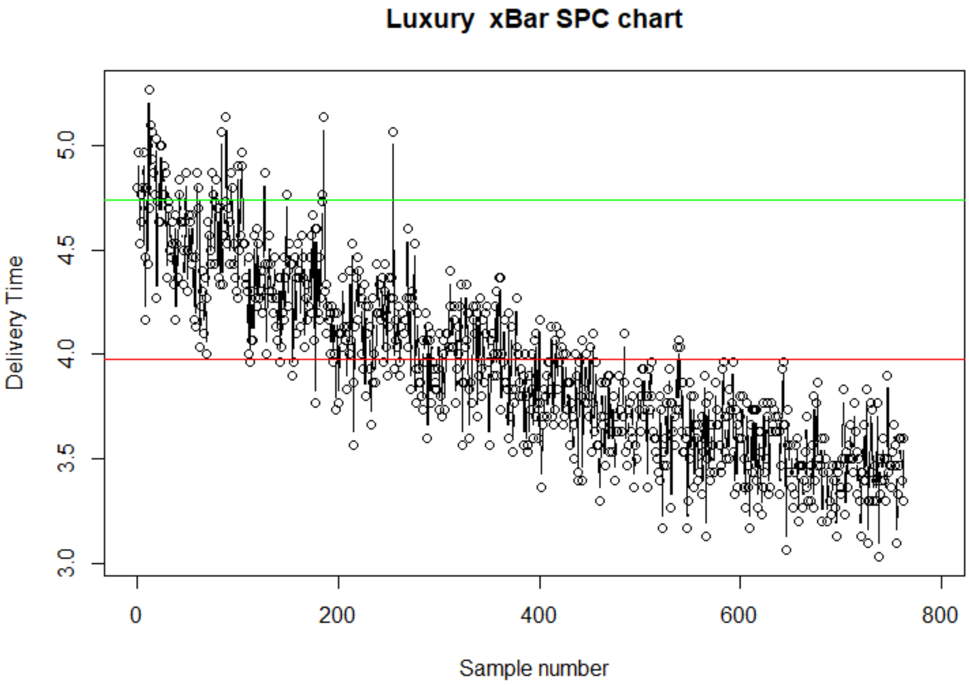


Figure 14 X-chart for Luxury Items

Figure 14 represents the mean delivery time of the “Luxury” feature, indicating a large number of points below the LCL this feature is out of control. But compared to the previous features this the delivery time of Luxury items reduced.

S-Bar SPC charts for all the Samples

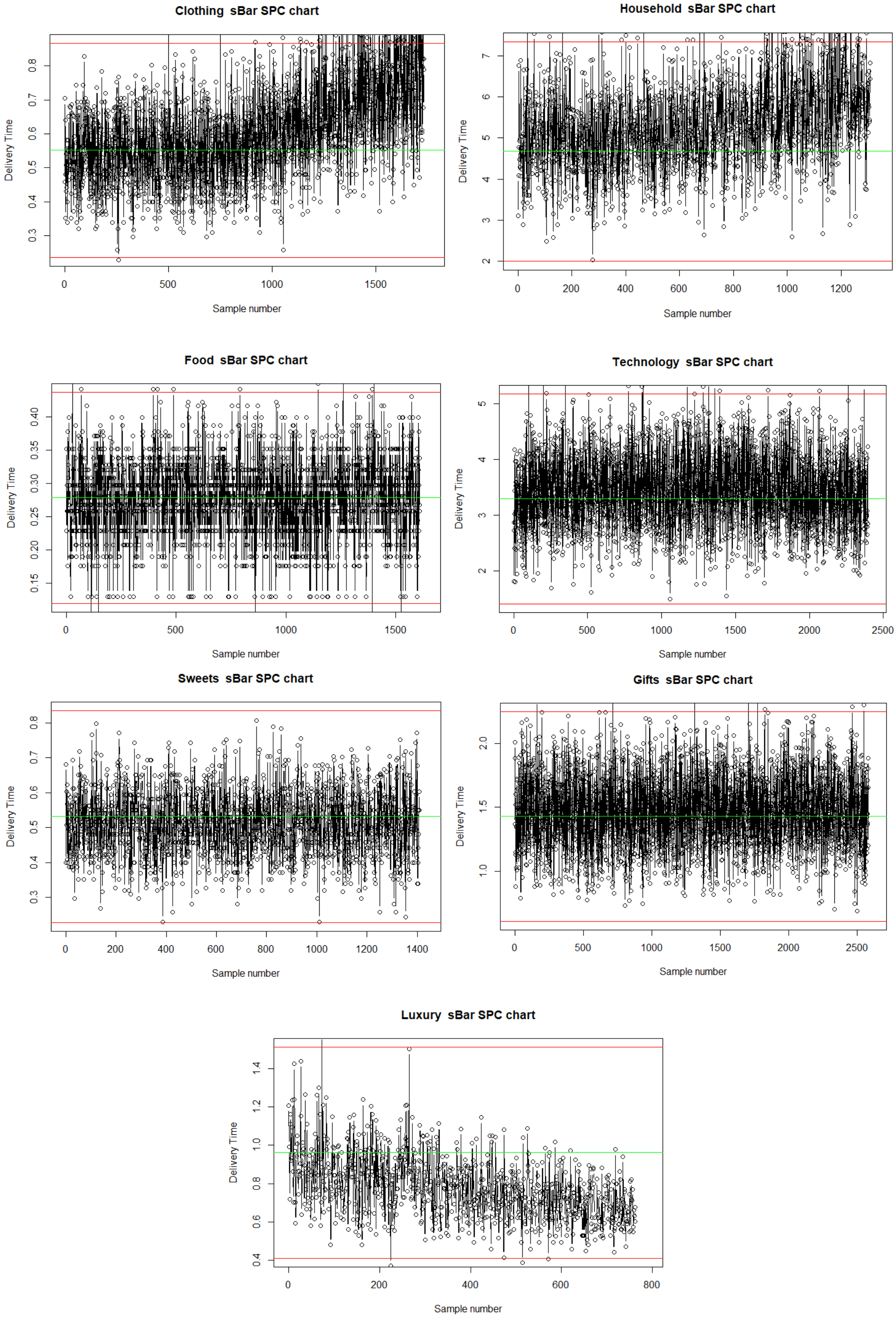


Figure 15 S-charts for Clothing, Household, Food, technology, Sweets, Gifts and Luxury

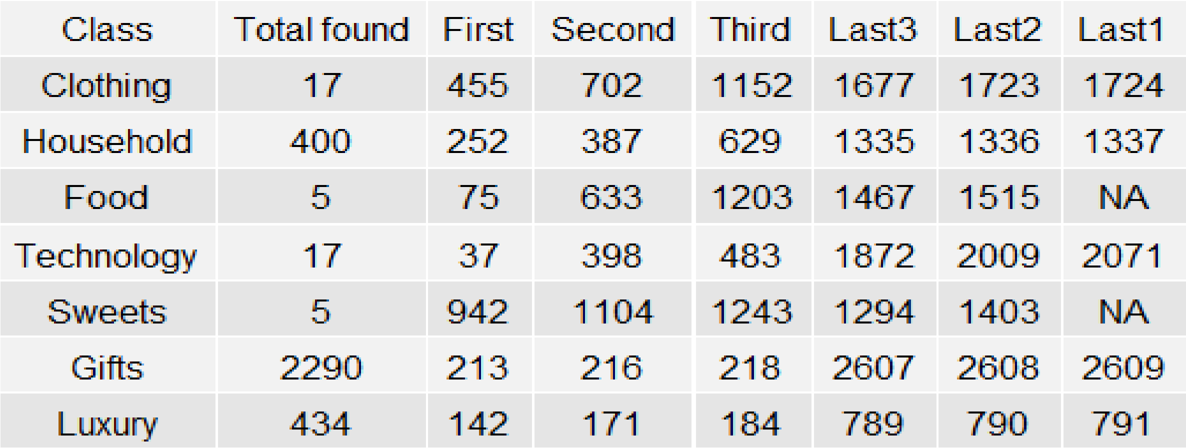
If S-charts is considered, feature like Household and clothing seem to be out of control. While the rest of the features are in control, which seems odd since gifts and luxury are presenting out of control in X-charts.

Part 4: Optimising the delivery Process

4.1 A List sample means (X-bar) outside of Control Limits

The following table (table 4) indicates the number of samples means outside of the Control limits, this table is useful because it gives us a better indication of classes that could be out of control and validates topics discussed in part 3.

Table 4 X Samples outside of CL's

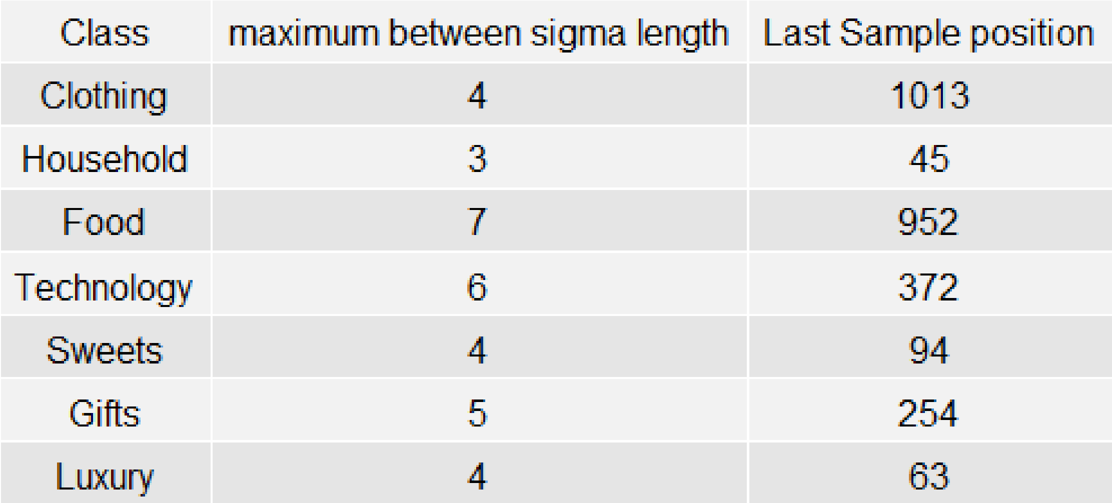


Class	Total found	First	Second	Third	Last3	Last2	Last1
Clothing	17	455	702	1152	1677	1723	1724
Household	400	252	387	629	1335	1336	1337
Food	5	75	633	1203	1467	1515	NA
Technology	17	37	398	483	1872	2009	2071
Sweets	5	942	1104	1243	1294	1403	NA
Gifts	2290	213	216	218	2607	2608	2609
Luxury	434	142	171	184	789	790	791

4.1 B Most consecutive Sample standard deviations between 0.4 and -0.3 Sigma

Indicated in the table 4 below is the number of consecutive points outside of +0.4 and -0.3 Sigma and the sample position of the last sample in the consecutive group.

Table 5 Consecutive S Samples +0.4-0.3Sigma



Class	maximum between sigma length	Last Sample position
Clothing	4	1013
Household	3	45
Food	7	952
Technology	6	372
Sweets	4	94
Gifts	5	254
Luxury	4	63

4.2 Type I error (manufacture's error)

The probability of making a type 1 error in a general sense UCL and LCL cover 99.74% of the data

This can be represented by the following:

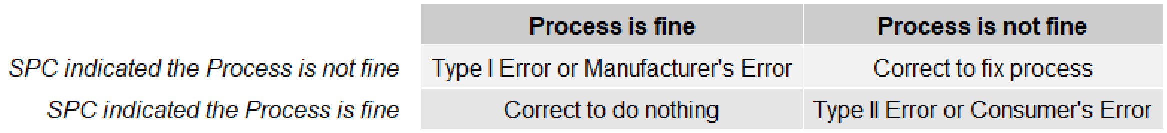
P (Xbar < LCL) + P (Xbar > UCL) - Xbar chart

P (Sbar < LCL) + P (Sbar > UCL) - S chart

This equation can then be written as -> P (z<-3) +P (z>3)

Resulting in the Probability of making a type I error being equal to 0.27 %

Table 6 Types of 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

4.3 Optimizing delivery times

If we want to optimize the delivery times for Technology Class items, we will need to shift the dataset by a set number of hours each way and calculate the resulting total cost, where the total cost is equal to cost of late items plus the cost of reducing the delivery times.

The calculated Current mean for Technology class items = 20.01095

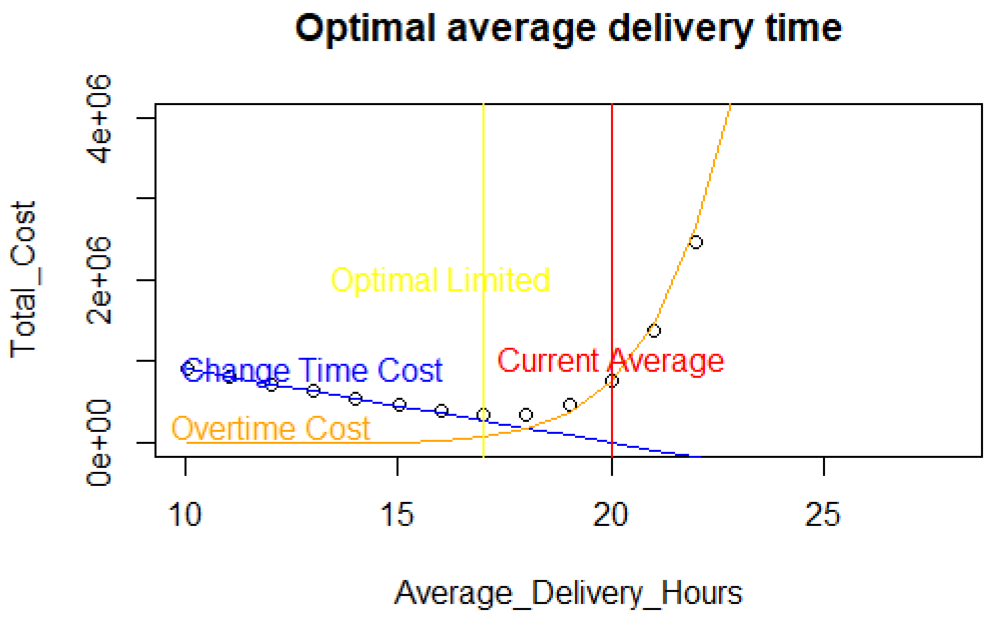


Figure 16 Optimal Average delivery Time

Finding the minimum value of total cost resulted in a position value of -3 hours, thus the best solution would be to shift our dataset by -3 hours, this should be done while keeping in mind that our minimum value in delivery times tech is equal to 4.5. Thus, we are limited to -4.5 since a negative delivery time is not possible.

Therefore, as indicated in the figure above the new average time is equal to 17.01095 resulting in savings of R 336610,2.

4.4 Type II error (Customer's error)

The following information would represent the likelihood of a type II (Consumer's) Error in the Technology Class to be made. The vales for UCL an LCL used in Equation is 22.97 and 17.77 respectively.

Standard deviation = (UCL-LCL)/6 = 0.8667

P of type II error = P (22.97-23/0.8667)-P (17.77-23/0.8667)

Solving the equation results in a Probability of 48.83 % chance of making a type II error.

Part 5: MANOVA & DOE

Looking the results of Part 2-4, we would like to know if there is a correlation between the reason the product was Bought and the Age, Delivery time and Price.

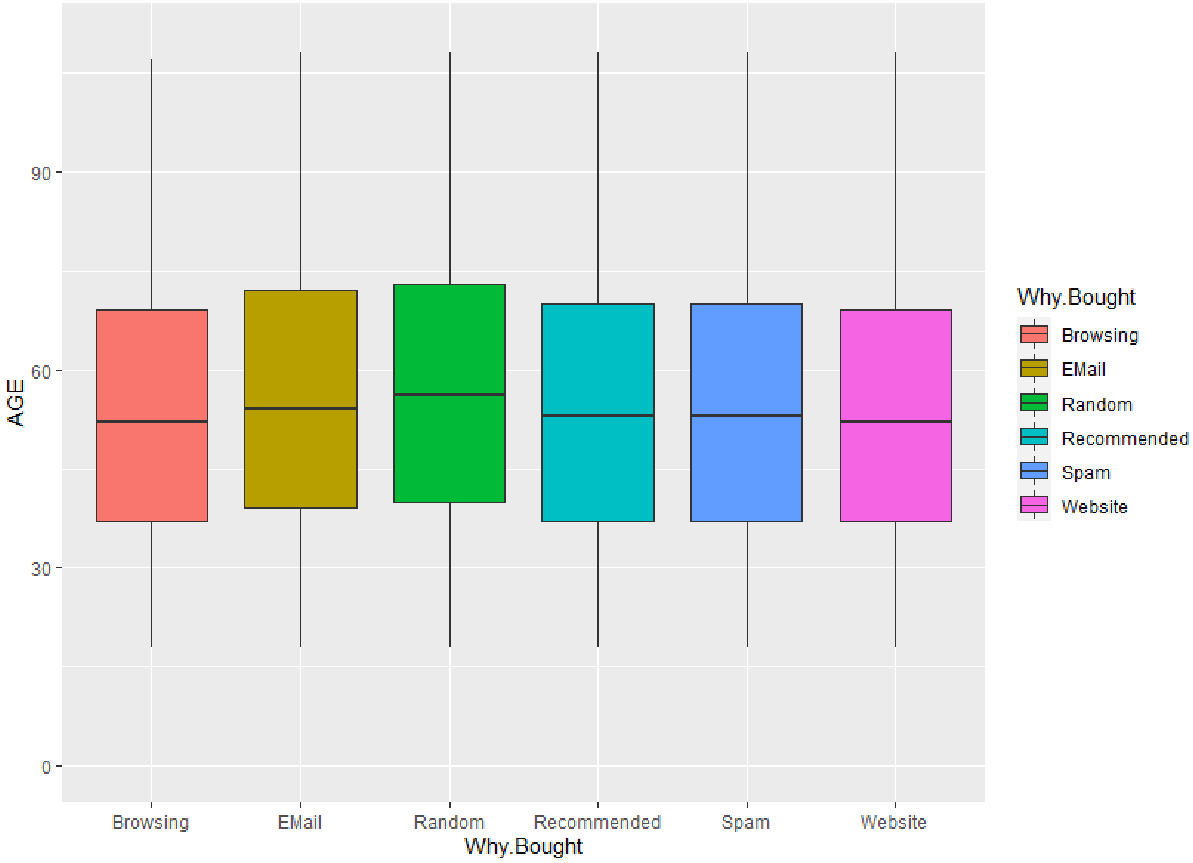


Figure 17 Box plot of Age and reasons Bought

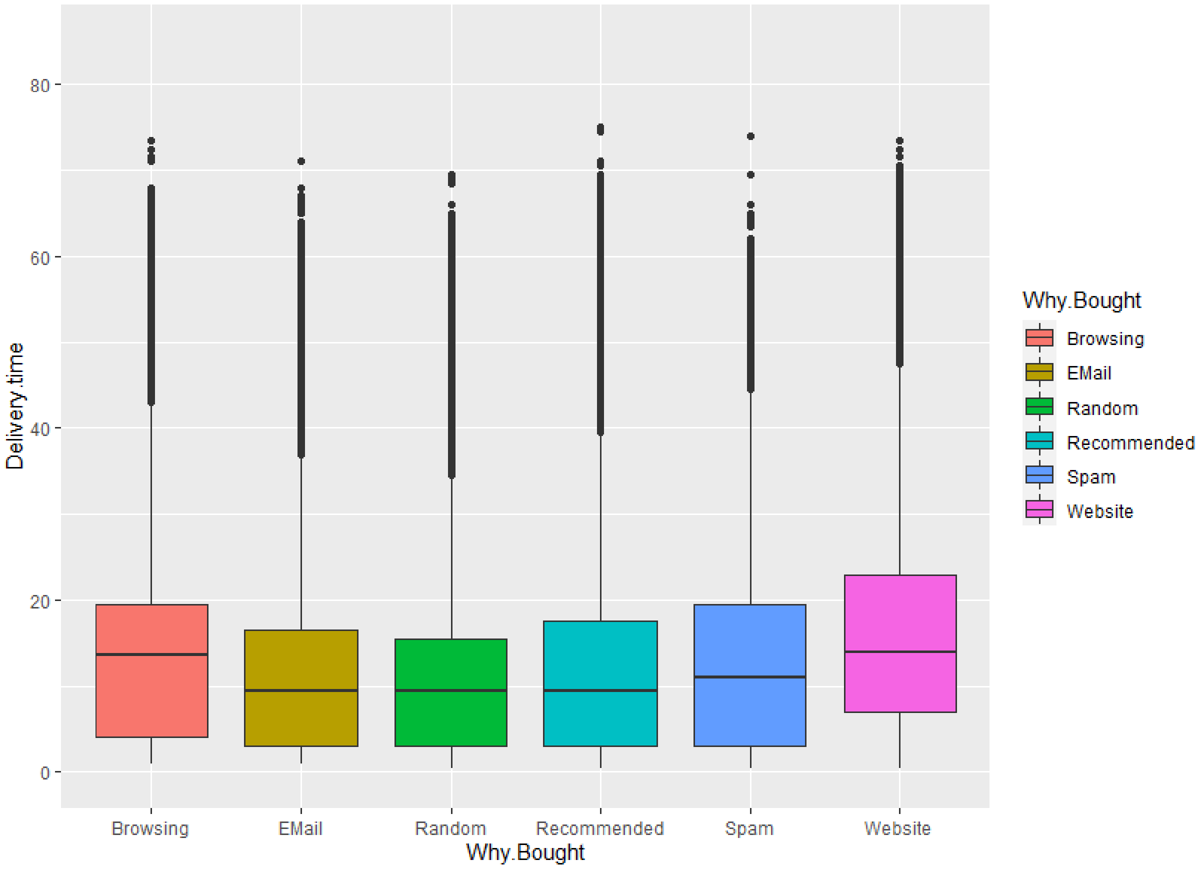
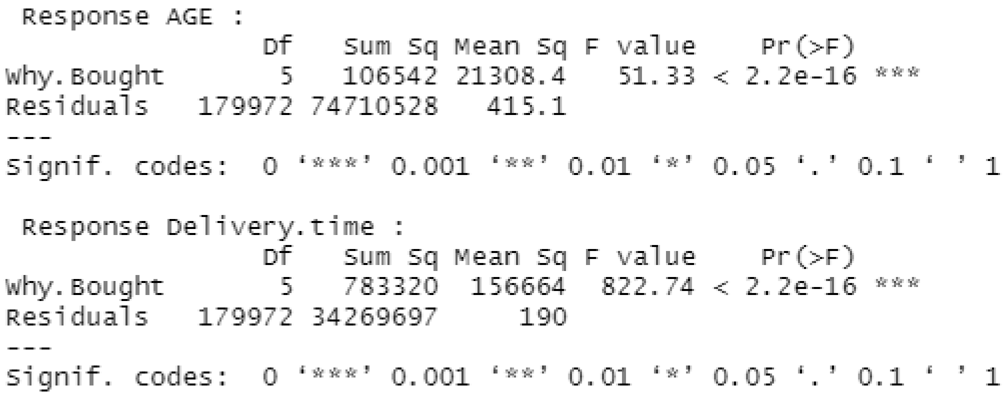


Figure 18 Box plot of Delivery time and reasons Bought

Running the Anova function for reasons the product was Bought in relation to the Age, Delivery time and Price, resulted in the output below.



```
 Response AGE :
                Df   Sum Sq Mean Sq F value    Pr(>F)    
Why.Bought       5   106542 21308.4   51.33 < 2.2e-16 ***
Residuals   179972 74710528   415.1                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Delivery.time :
                Df   Sum Sq Mean Sq F value    Pr(>F)    
Why.Bought       5   783320  156664  822.74 < 2.2e-16 ***
Residuals   179972 34269697     190                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In the output above the greater the F value the greater the evidence that there is a difference between the group means. While having a p value smaller that the alfa of 0.05 would result in the rejection of our Ho and thus conclude that there is a statistically significant difference between the groups compared.

Referring to the output we see that the resulting probability for a Ho is equal to 2.2e-16 with a F value of 51.33, this is for AGE in relation with Why Bought. Figure 17 reflects this Ho value since the graphs are similar but not identical and thus, we reject the Ho and conclude that there is a statistically significant difference.

Referring to the output we see that the resulting probability for a Ho is equal to 2.2e-16 with a large F value of 822.74, this is for Delivery time in relation with Why Bought. Figure 18 reflects this Ho value since the graphs are similar but not identical and thus, we reject the Ho and conclude that there is a statistically significant difference.

Part 6: Reliability of Service and products

6.1

Problem 6 (p.363)

Specifications = 0.06 ±0.04 cm.

Scrapping cost = $45 per part.

$L(x) = k*(x – T)^2$

$k = 45/ (0.04)^2$

k = 28 125

Thus, the Taguchi loss is represented by the following equation:

$L(x) = 28125*(x – 0.06)^2$

Problem 7 (p.363)

Reduced scrapping cost = $35 per part

a)

$L(x) = k*(x – T)^2$

$k = 35/ (0.04)^2$

k = 21875

Thus, the Taguchi loss is represented by the following equation:

$L(x) = 21875*(x – T)^2$

b) Process déviation (T) = 0.027

The Taguchi Loss thus equal to $L= 21875*(0.027)^2$ = $15.95

Thus, for each part with a process deviation, there would be an additional cost of $15.95

6.2

Problem 27 (p.365)

a)

In the case that all the backup machines are out of order, the system reliability would be calculated as follows:

Reliability = 0.85 * 0.92 * 0.9 = 0.7038

The system reliability without backup machines would equal 0.7038*100 = 70.38 %

b)

If the backup machines are in working order the machines would be in parallel, thus increasing the calculated reliability as follows:

Reliability = $(1 – (1 – 0.85)^2) * (1 – (1 – 0.92)^2) *(1 – (1 – 0.9)^2) = 0.9615$

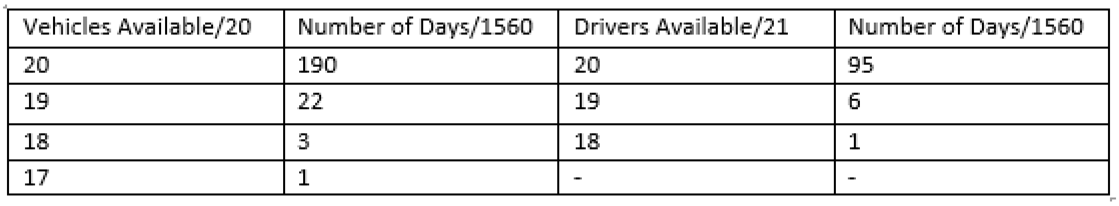
Thus, the reliability of the system if all the machines are in working order would amount to 0.9615*100 = 96.15 % this is substantially more than the 70.38 % system reliability due to the backup machines being out of order. Thus, the improvement in reliability is due to identical machines in parallel 96.15-70.38 = 25.77%

6.3

Binominal Probability’s

For the delivery process deliver reliable service there should be 19 of the 20 delivery vehicles available.

Table 7 Values for binominal Probability’s



Vehicles Available/20	Number of Days/1560	Drivers Available/21	Number of Days/1560
20	190	20	95
19	22	19	6
18	3	18	1
17	1	-	-

To determine the number of days per year we could expect reliable delivery times, well need to use the equation below.

Equation 2

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

Since there is no requirement for number of drivers, we only calculating the probability for 21 vehicles.

- P (21) = (1560-190-22-3-1)/1560 = 0.86153 = $\binom{21}{0}\, p^0\ (1-p)^{21-0}$ thus P (0 fail) = 9.2997x10^-19
- P (20) = 190/1560 = 0.12179 = $\binom{21}{1}\, p^1\ (1-p)^{21-1}$ thus P (1 fail) = 0.19045
- P (19) = 22/1560 = 0.01410 = $\binom{21}{2}\, p^2\ (1-p)^{21-2}$ thus P (2 fail) = 0.03187
- P (18) = 3/1560 = 0.0019231 = $\binom{21}{3}\, p^3\ (1-p)^{21-3}$ thus P (3 fail) = 0.000009137
- P (17) = 1/1560 = 0.00064102 = $\binom{21}{4}\, p^4\ (1-p)^{21-4}$ thus P (4 fail) = 9.99577109x10^-10

Weighted P = 1344(9.2997x10^-19) +190(0.19045) +22(0.03187) +3(0.000009137) +1(9.99577109x10^-10)/1560 =0.02365

- P (21) = $\binom{21}{0}\, p^0\ (1-p)^{21-0}$ thus P (0 fail) = 0.604945
- P (20) = $\binom{21}{1}\, p^1\ (1-p)^{21-1}$ thus P (1 fail) = 0.307723
- P (19) = $\binom{21}{2}\, p^2\ (1-p)^{21-2}$ thus P (2 fail) = 0.074539

Expected number of days for 0 failures = 0.604945 * 1560 = 943

Expected number of days for 1 failure = 0.307723 * 1560 = 480

Expected number of days for 2 failures = 0.074539 * 1560 = 116

Thus, the expected number of days of reliable delivery = 1539/1560 *365 = 360.08 days

Increasing the number of vehicles of 22

Weighted P = 1344(9.2997x10^-19) +190(0.19045) +22(0.03187) +3(0.000009137) +1(9.99577109x10^-10)/1560 =0.02365

- P (22) = $\binom{22}{0} p^0 \; (1-p)^{22-0}$ thus P (0 fail) = 0.590638
- P (21) = $\binom{22}{1} p^1 \; (1-p)^{22-1}$ thus P (1 fail) = 0.314753
- P (20) = $\binom{21}{2} p^2 \; (1-p)^{22-2}$ thus P (2 fail) = 0.080054
- P (19) = $\binom{21}{3} p^3 \; (1-p)^{22-3}$ thus P (3 fail) = 0.0129276

Expected number of days for 0 failures = 0.590638 * 1560 = 921.39

Expected number of days for 1 failure = 0.314753 * 1560 = 491.01

Expected number of days for 2 failures = 0.080054 * 1560 = 124.88

Expected number of days for 3 failures = 0.0129276 * 1560 =20.167

Thus, the expected number of days of reliable delivery = 1557.45/1560 *365 = 364.40 days

Conclusion

Technology, Gifts and Household items are the 3 features that contributes most to the sales and therefore it is extremely important that these features are promoted as much as possible to generate more sales and ultimately increase the profit. The delivery times for Gifts, Luxury and Household items should be analysed and improved to help to improve our customer service level and to make the company more reliable and attractive. Neglecting this could lower the company's recommendations and directly influence their sales, especially since the company's best method of attracting customers are through recommendations. Marketing profiles should be adjusted according to the recommended age group for features, taking in account the age appropriateness of the products.

Reverences

1. Dr TG Driske van Schalkwyk (2022). *Statistics for QA344.* Retrieved from https://learn.sun.ac.za/course/view.php?id=70723