ECSA Graduate Attributes Project

Quality Assurance 344

Aidan Bristow

23692006

18 October 2022

Contents

List of tables

List of figures

Introduction

This report analyses the sales data of an online business. This report's objectives are to ascertain how well the company is currently operating and to provide thoughtful recommendations regarding the outcomes. Firstly, for the analysis to be accurate, all invalid data must be removed through data wrangling. Then to better comprehend the features, the data is examined for all statistical metrics. These statistical measures are then discussed. To provide advice to the business regarding the current technology process, the process capabilities of the technology class are calculated. Statistical process control is then used to find samples of the data that were out of control. These samples are then investigated. Technologies delivery times are then evaluated with the goal to reduce lost sales by determining the technology class's optimal delivery time. Type 1 and 2 errors for the system are then discussed. A MANOVA table is created to display a link between the data. In the final section of the report the reliability of other business functions is calculated.

1.Part1: Data Wrangling

The complete dataset consists of 180 000 instances and 10 features. The data was separated into two categories being valid and invalid data. This was done by examining the data for cases that were not valid. These cases are removed from the dataset before it is used. All the invalid data was found in the price feature. This included instances of missing values and negative prices. The dataset contains 17 instances of missing values and 5 instances with negative prices. All 22 of the instances of invalid data that were removed are shown in the table below.

Index	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	16320	44142	82	Household	-512	2023	10	2	96	EMail
2	19540	65689	96	Sweets	-512	2028	4	7	6	Random
3	19998	68743	45	Household	-512	2024	7	16	91	Recommended
4	144443	37737	81	Food	-512	2022	12	10	5	Recommended
5	155554	36599	29	Luxury	-512	2026	4	14	7	Recommended
6	12345	18973	93	Gifts	NA	2026	6	11	31	Website
7	16321	81959	43	Technology	NA	2029	9	6	44	Recommended
8	19541	71169	42	Technology	NA	2025	1	19	41	Recommended
9	19999	67228	89	Gifts	NA	2026	2	4	30	Recommended
10	23456	88622	71	Food	NA	2027	4	18	5	Random
11	34567	18748	48	Clothing	NA	2021	4	9	16	Recommended
12	45678	89095	65	Sweets	NA	2029	11	6	4	Recommended
13	54321	62209	34	Clothing	NA	2021	3	24	19	Recommended
14	56789	63849	51	Gifts	NA	2024	5	3	21	Website
15	65432	51904	31	Gifts	NA	2027	7	24	29	Recommended
16	76543	79732	71	Food	NA	2028	9	24	5	Recommended
17	87654	40983	33	Food	NA	2024	8	27	4	Recommended
18	98765	64288	25	Clothing	NA	2021	1	24	17	Browsing
19	144444	70761	70	Food	NA	2027	9	28	5	Recommended
20	155555	33583	56	Gifts	NA	2022	12	9	20	Recommended
21	166666	60188	37	Technology	NA	2024	10	9	43	Website
22	177777	68698	30	Food	NA	2023	8	14	5	Recommended

Table 1: Table of invalid data removed from dataset

2.Part2: Descriptive Statistics

2.1 Features of the data set

Each feature and their associated class are shown. There are six continuous features and five categorical.

Feature	Class
Index	Continuous
X	Continuous
ID	Continuous
AGE	Continuous
Class	Categorical
Price	Continuous
Year	Categorical
Month	Categorical
Day	Categorical
Delivery.time	Continuous
Why.bought	Categorical

Table 2: Table of features class

2.2 Continuous Features

Feature	Count	Percentage Missing	Cardinality	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Standard Deviation
Age	179978	0	91	18	38	53	54.56552	70	108	20.38880835
Price	179978	0	78832	35.65	482.31	2259.63	12294.1	15270.97	116619	20889.15025
Delivery_Time	179978	0	148	0.5	3	10	14.50031	18.5	75	13.95578266

Table 3: Descriptive table of continuous variables

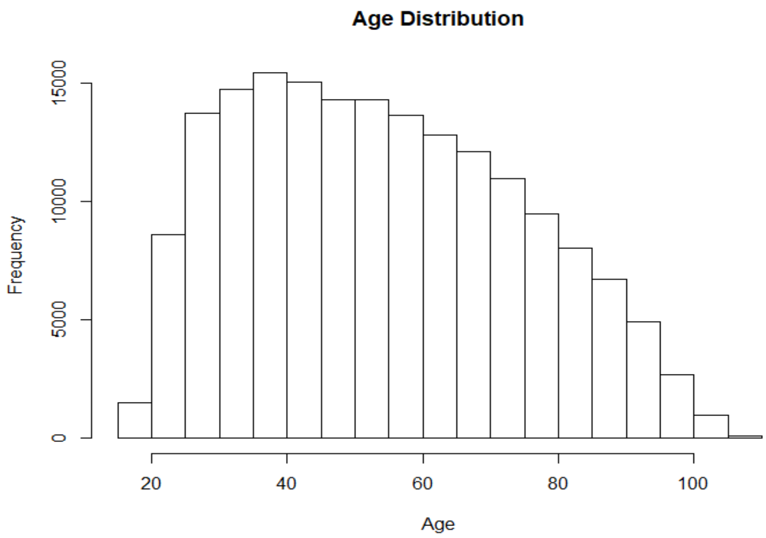


Figure 1: Distribution of age

The distribution of age is unimodal (right-tailed). The distribution is skewed to the right which is simply a tendency towards very low (right skew) values. This makes sense as there are fewer people buying items after the age of 80. Age has a cardinality of 91, shown in Table 3, with most customers ranging around the age of 40.

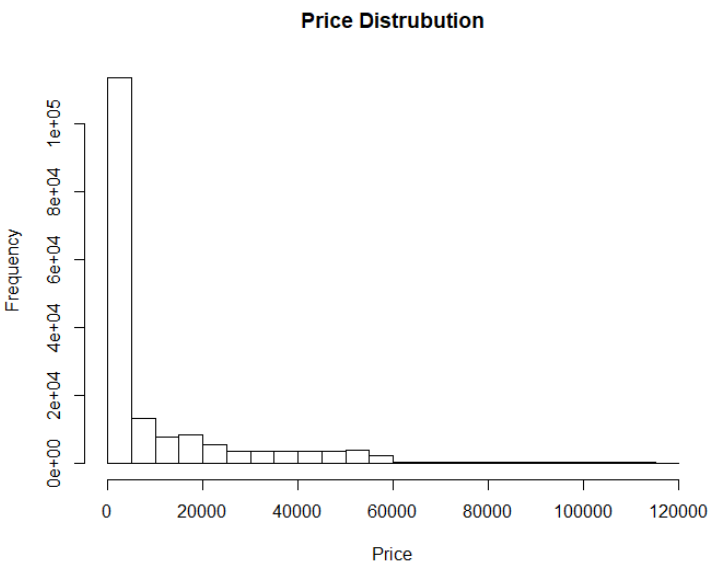


Figure 2: Price distribution

Price shows an exponential distribution. This shows that the likelihood of occurrence for low values is very high, but sharply diminishes as values increase. The range of the minimum price compared to the maximum, shown in Table 3, is large with the range being 116 583.35.

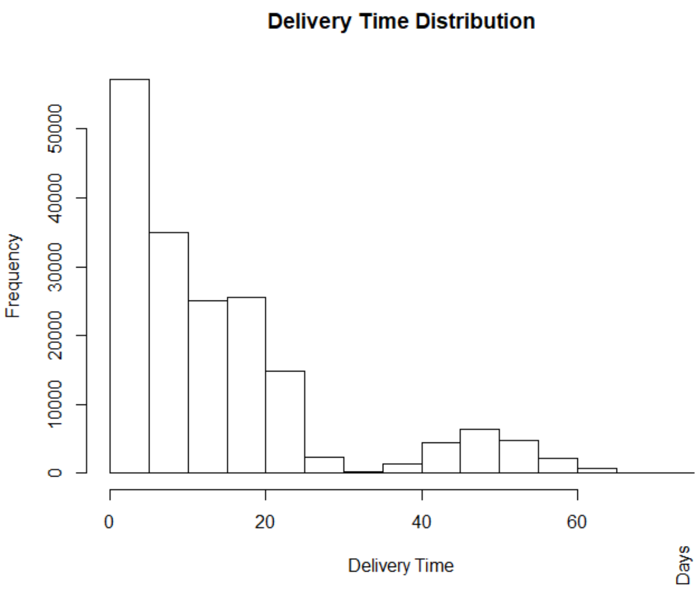


Figure 3: Delivery time distribution

Delivery time shows a multimodal distribution. With two very commonly occurring ranges of values that are clearly separated. Delivery time is most commonly between the range of 0 to 20 with 10 being the median. It is however found that delivery time can be longer, lying between 40 to 60. With no delivery times between these intervals. These longer delivery times result mostly from customers purchasing household items. Shown in Figure 4 below.

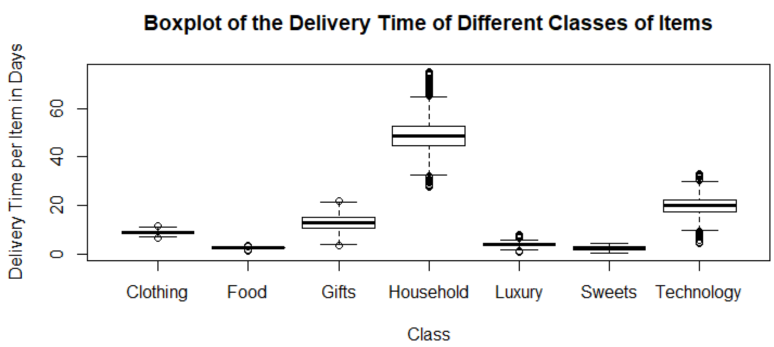


Figure 4: Delivery time vs class

2.3 Categorical features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq	2nd Mode %
Class	179978	0	7	Gifts	39149	21.75	Technology	36347	20.2
Year	179978	0	9	2021	33443	18.58	2029	22475	12.49
Month	179978	0	12	12	15225	8.46	10	15221	8.46
Day	179978	0	30	17	6126	3.4	25	6122	3.4
Why.Bought	179978	0	6	Recommended	106985	59.44	Website	29447	16.36

Table 4: Descriptive table of categorical variables

2.3.1 Analysis of class

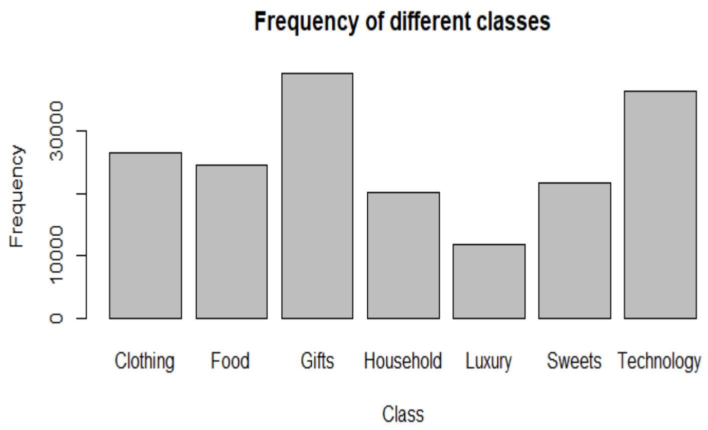


Figure 6: Frequency of different classes

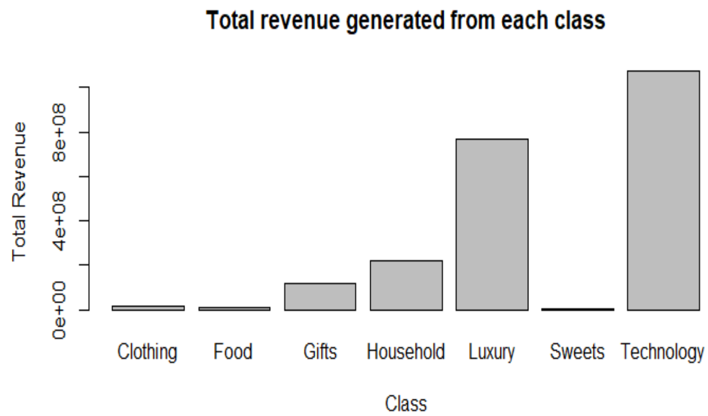


Figure 5: Revenue generated from each class

It is important to identify which items are most popular as well as sales generated from each class. Figure 5 represents the frequency of each class. This shows us that gifts and technology are the most popular items. When this is compared total revenue produced from each class, represented in figure 6, we can see that even though gifts are the most popular item they only generate the fourth highest revenue after households, luxury, and technology. Technology produces the highest revenue generating an income of R1 072 529 553. The business should consider removing items such as food and sweets which generate significantly less revenue in comparison to the other classes. This will allow the business to focus more on improving the classes that are creating the most value for the business.

2.3.2 Reason for why items were bought

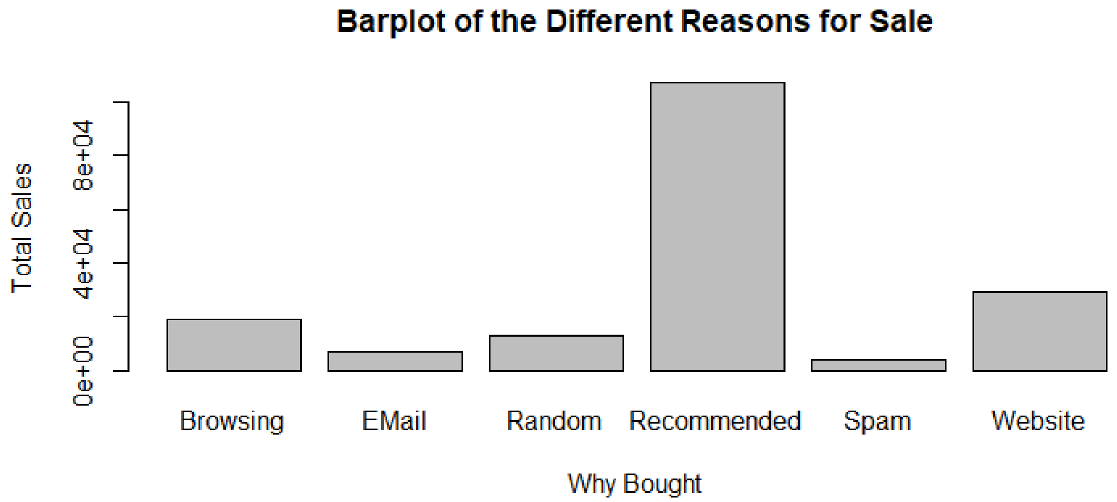
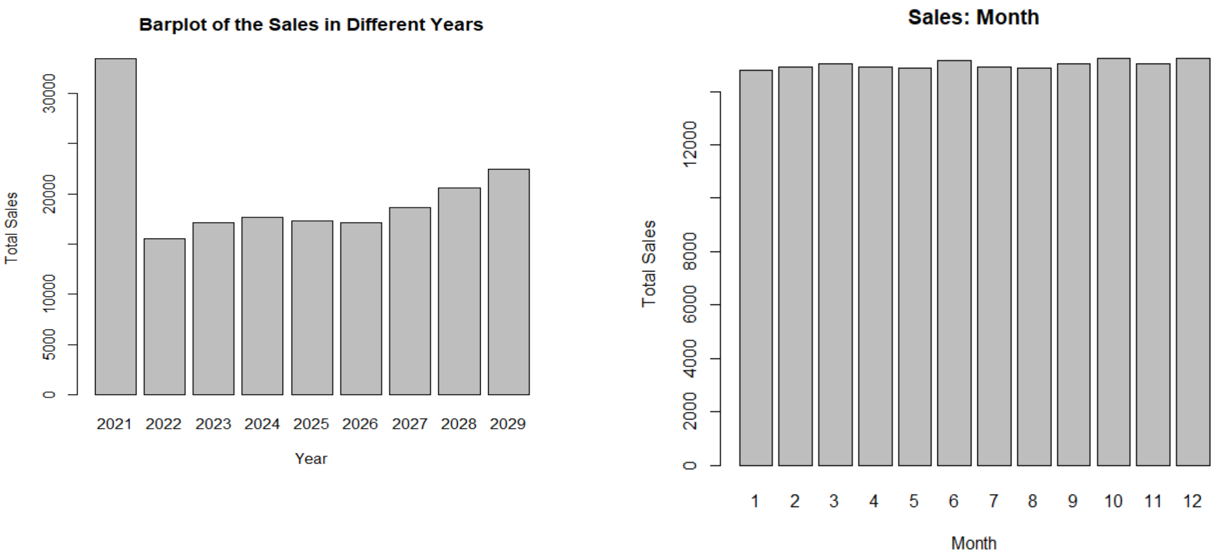
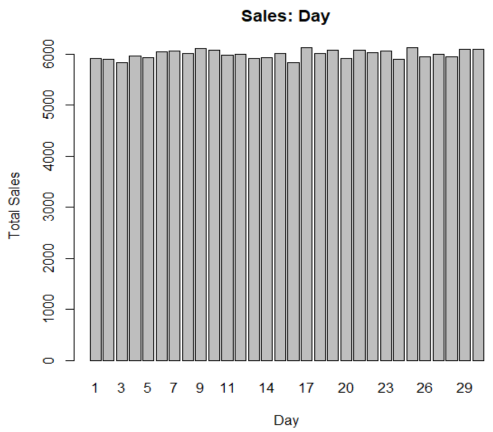


Figure 7: Different reasons for sales

The most frequent reason why someone bought a product is because someone recommended the product to them. The least frequent reason why someone bought a product is because of a spam. This shows that our customers are recommending the store to other people. This is an indication of loyal customers that are satisfied with the quality and service of our products. The business must focus their attention on continuing to meet and exceed customers expectation which will lead to more recommendations. Advertising through the company's website and email distribution could be improved.

2.2.3 Sales in different periods





Total sales in each day and each month are normally distributed with the most sales happening on the 17th day and during December. This is result of people spending more during December holiday. The business must be prepared for this spike in demand to ensure they do not run out of stock. Total sales for each year show that 2021 generated significantly more sales than any other year. This is an area of concern for the business as the sales should be increasing from year to year.

Figure 8: Different period of sales

2.4 Process Capability

The process capability of the company for delivery time of technology items is determined to measure the extent of variation the delivery time experiences relative to the specification limits.

$$USL = 24\ days$$

$$LSL = 0\ days$$

$$Standarddeviation(\sigma) = 3.502$$

$$Mean\ (\mu) = 20.011$$

The Cp ratio shows how well the delivery time of technology (expressed as six standard deviations of the process) fits into the specification range.

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

$$C_p = 14.008$$

Cp > 1 showing that the process can reach the goal it is designed for.

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

$$C_{pu} = 4.657$$

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

$$C_{pl} = 23.359$$

$$C_{pk} = Min(C_{pu}, C_{pl})$$

$$C_{pk} = 4.657$$

The Cpk value is > 1 implying that the process is well within the specification limits. Cpk of 4 is slightly high and indicates that there are loose specification limits.

3.Part3: Statistical Process Control

3.1 Control Limits

The dataset is first sorted chronologically in terms of year, month, and day from oldest to newest. The dataset is then divided into 30 samples of 15 instances each. The X and S charts are then calculated using these samples.

3.1.1 S-Chart

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	0.86656	0.761455	0.656351	0.551247	0.446142	0.341038	0.235934
Household	7.34418	6.45341	5.56264	4.67187	3.7811	2.89033	1.99956
Food	0.437247	0.384213	0.33118	0.278147	0.225113	0.17208	0.119047
Technology	5.18057	4.552222	3.923875	3.295528	2.667181	2.038833	1.410486
Sweets	0.835339	0.734022	0.632704	0.531386	0.430069	0.328751	0.227433
Gifts	2.246333	1.973877	1.701421	1.428965	1.156509	0.884053	0.611597
Luxury	1.511052	1.327777	1.144503	0.961229	0.777955	0.59468	0.411406

Table 5: S-Chart

3.1.2 X-Chart

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	9.404934	9.259956	9.114978	8.97	8.825022	8.680044	8.535066
Household	50.24833	49.01963	47.79092	46.56222	45.33352	44.10482	42.87612
Food	2.709458	2.636305	2.563153	2.49	2.416847	2.343695	2.270542
Technology	22.97462	22.10789	21.24117	20.37444	19.50772	18.641	17.77427
Sweets	2.897042	2.757287	2.617532	2.477778	2.338023	2.198269	2.058514
Gifts	9.488565	9.112747	8.736929	8.361111	7.985293	7.609475	7.233658
Luxury	5.493965	5.241162	4.988359	4.735556	4.482752	4.229949	3.977146

Table 6: X-Chart

3.2 Graphs for first 30 samples

3.2.1 Clothing

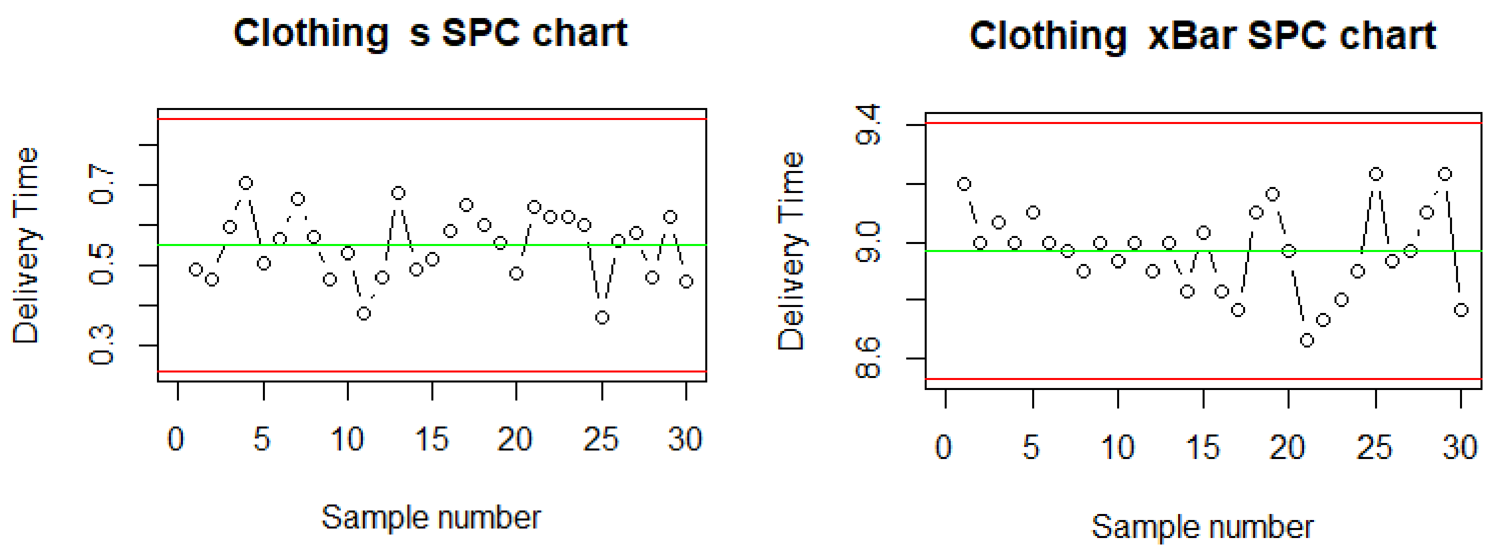


Figure 9: S&X chart for clothing

3.2.2 Household

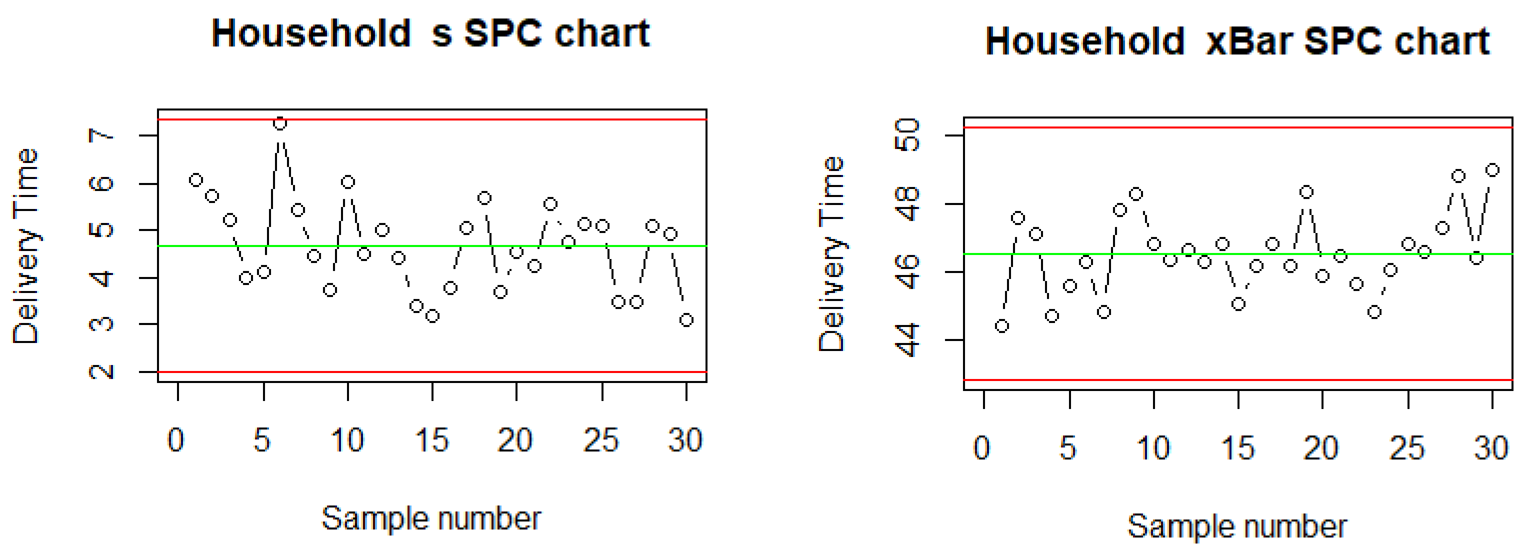


Figure 10: S&X chart for household

3.2.3 Food

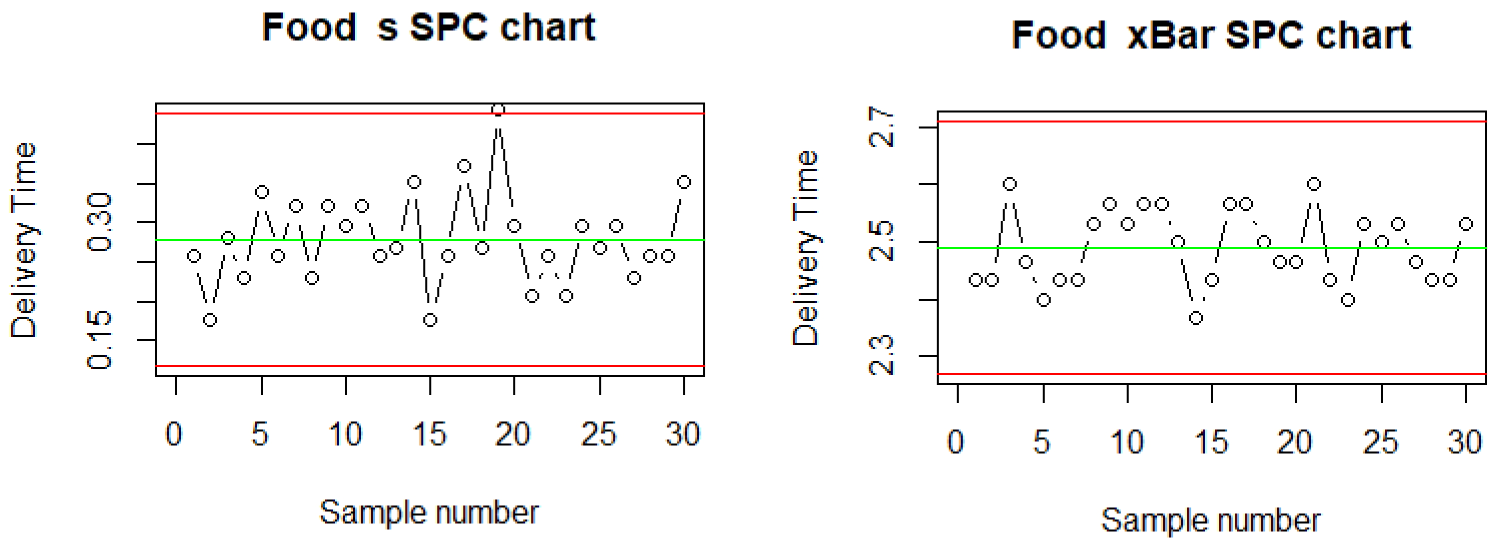


Figure 11: S&X chart for food

3.2.4Technology

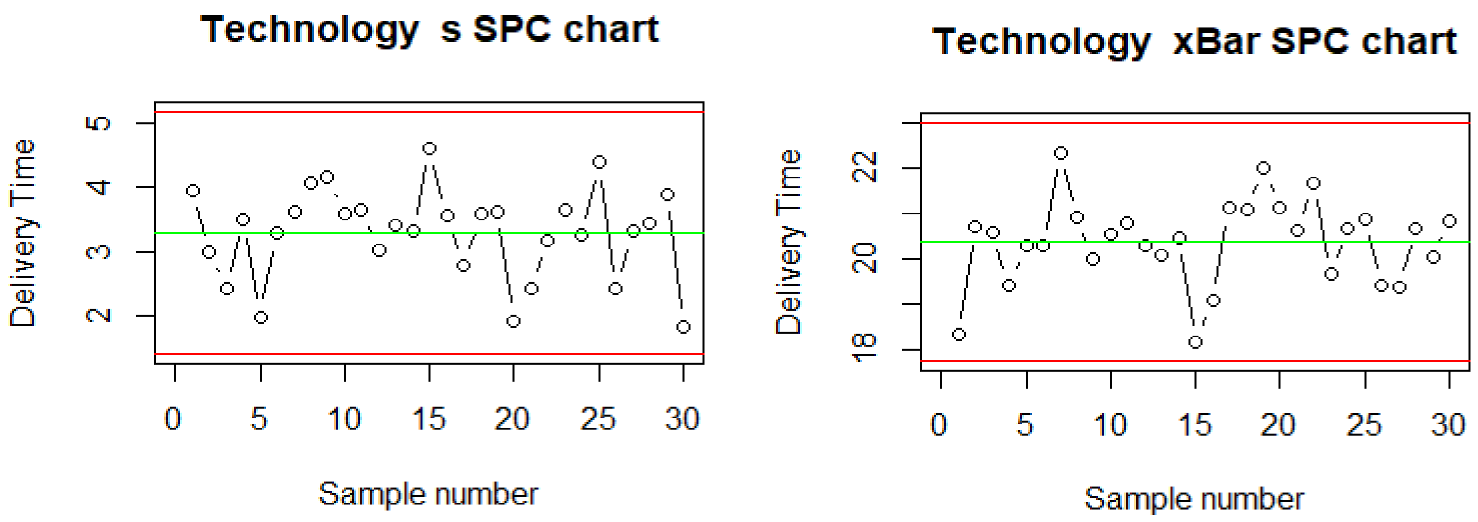


Figure 12: S&X chart for technology

3.2.5 Sweets

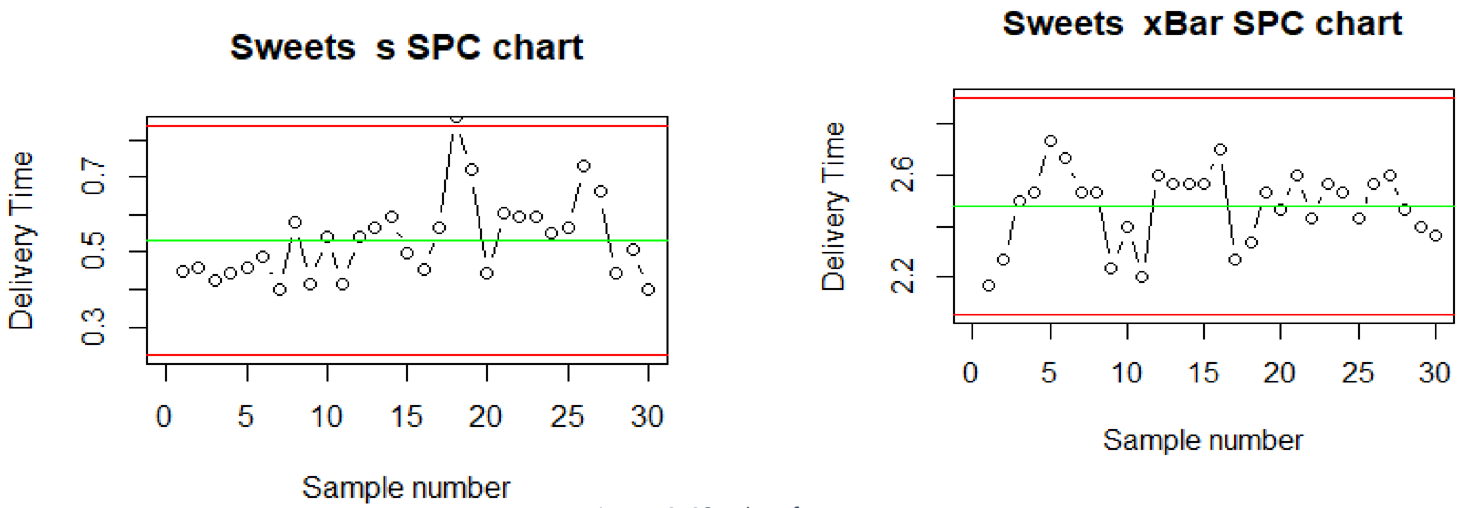
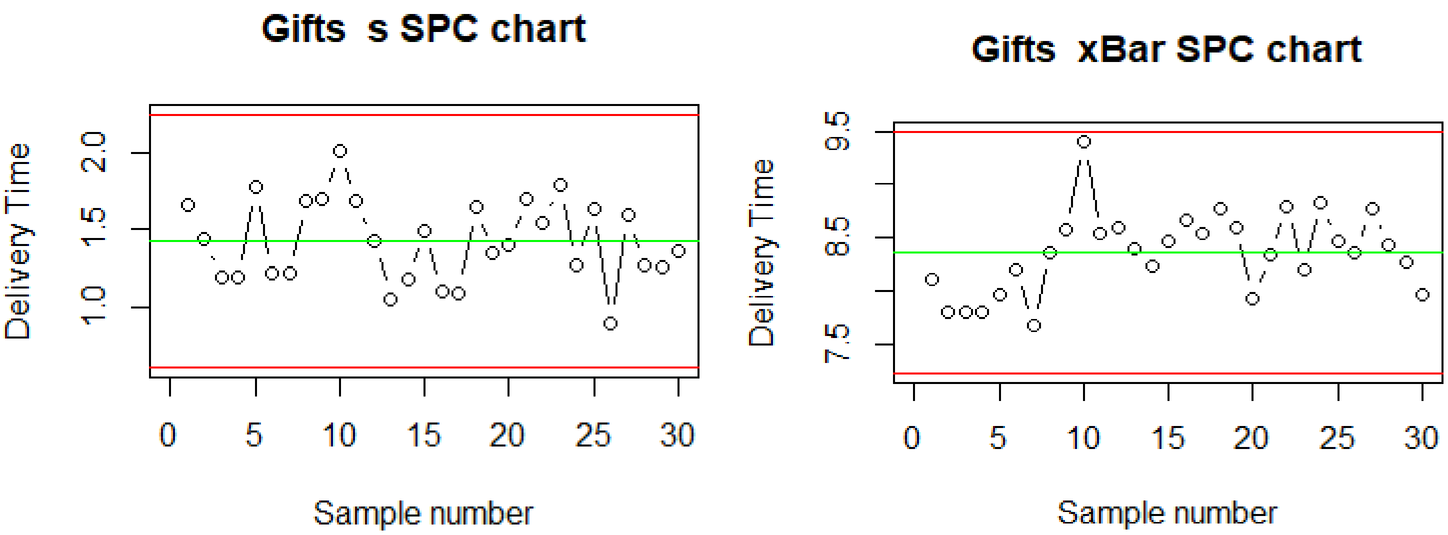


Figure 13: S&X chart for sweets

3.2.6 Gifts



3.2.7 Luxury

Figure 14: S&X chart for gifts

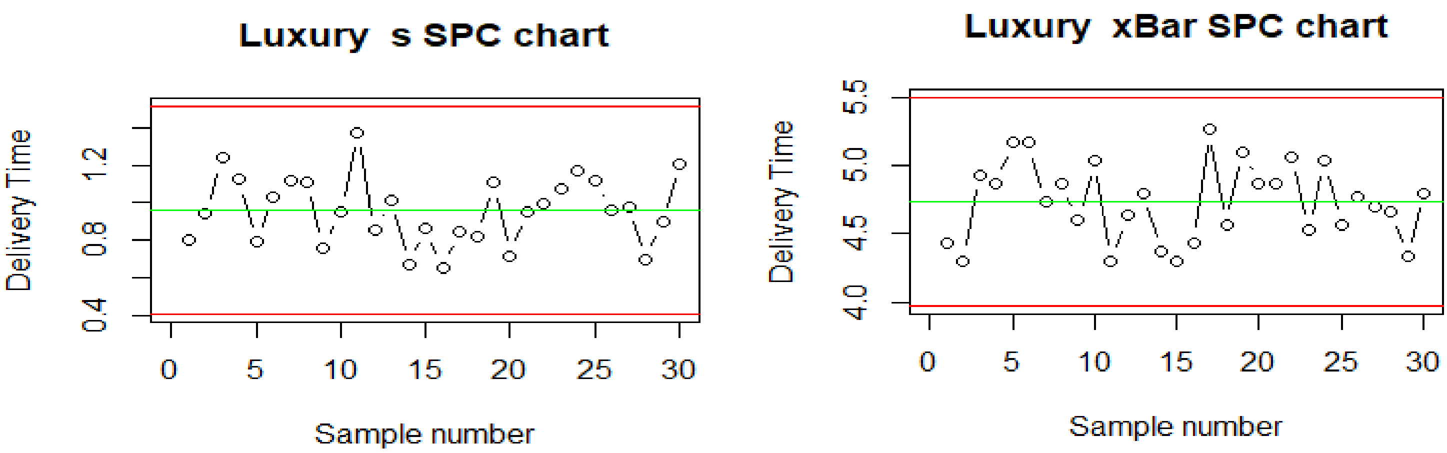


Figure 15: S&X chart for luxury

3.3 Evaluation of S&X charts for first 30 samples

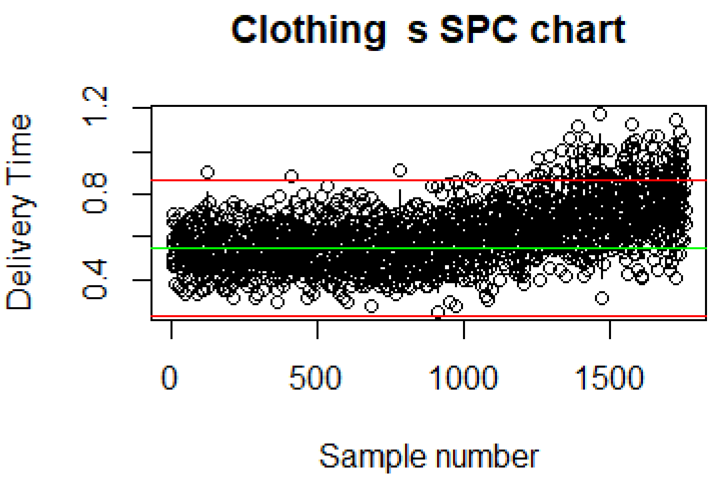
S-Charts should be constructed before X-Charts as they are used to determine whether the corresponding X-Chart is reliable. If a S-Chart has values above the UCL or below the LCL then the chart is out of control and the corresponding X-chart is unreliable. After S-chart has been evaluated X-Charts are then used to determine whether a process is in statistical control.

The S-charts for clothing, household, technology, gifts, and luxury shown in figure 9,10,12,14 and 15, shows all these classes are in control and there is no cause of variation in the process of these classes'. The X- bar chart can therefore be evaluated due to the satisfactory S bar chart. The X-chart for these classes shows that the processes are in statistical control.

The S-chart for food and sweets shown in figure 11 and 13, shows that for these classes there are values above the UCL this means that there is a special cause present in the process. Management needs to take action to find the special cause and permanently remove it from the process.

3.4 Graphs for all the samples

3.4.1 Clothing



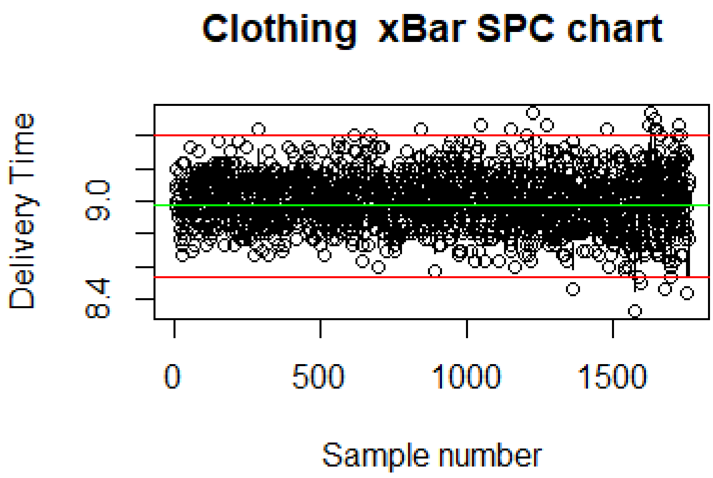


Figure 16: Complete S&X chart for clothing

Up to sample 1250 clothing is in statistical control with only two samples that fall outside the standard deviation upper control limit, these can be seen as outliers. After sample 1250 clothing is out of statistical control with many samples falling above the upper control limit indicating that the standard deviation of the delivery time is too high. With the S-chart not in statistical control the X-chart for clothing cannot be evaluated for these samples.

3.4.2 Household

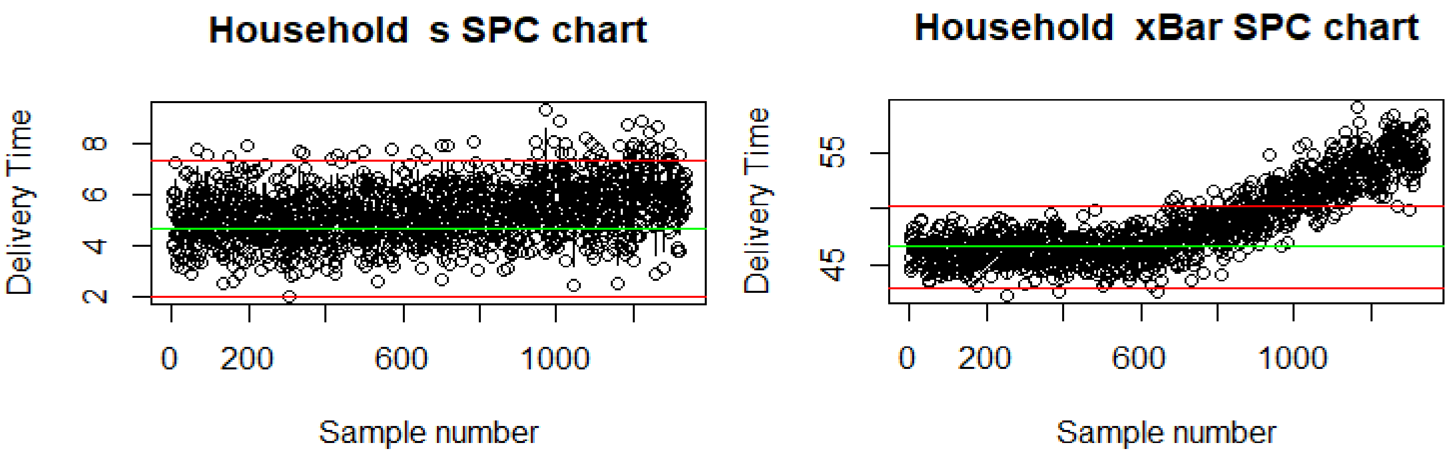


Figure 17: Complete S&X chart for Household

Household delivery time are out of control for all samples after 800. This is a concern for the business, management must investigate the cause for these late delivery times.

3.4.3 Food

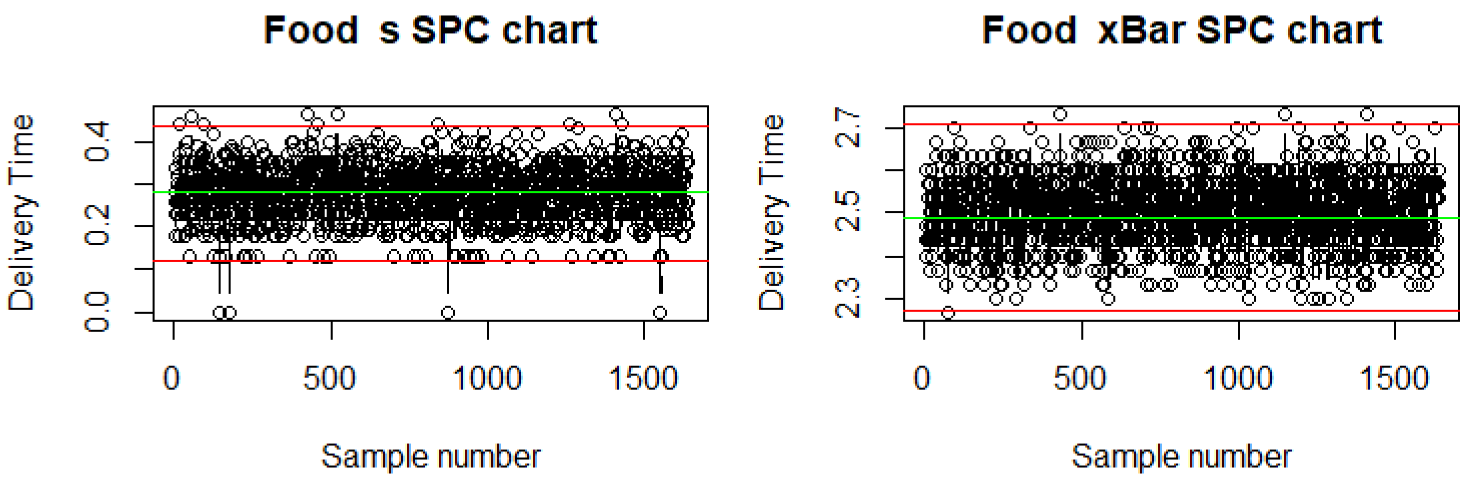


Figure 18: complete S&X chart for food

Food is in statistical control for both the standard deviation and mean of delivery time. There are still the occasional values that fall outside of the limit these can be seen as outliers.

3.4.4 Technology

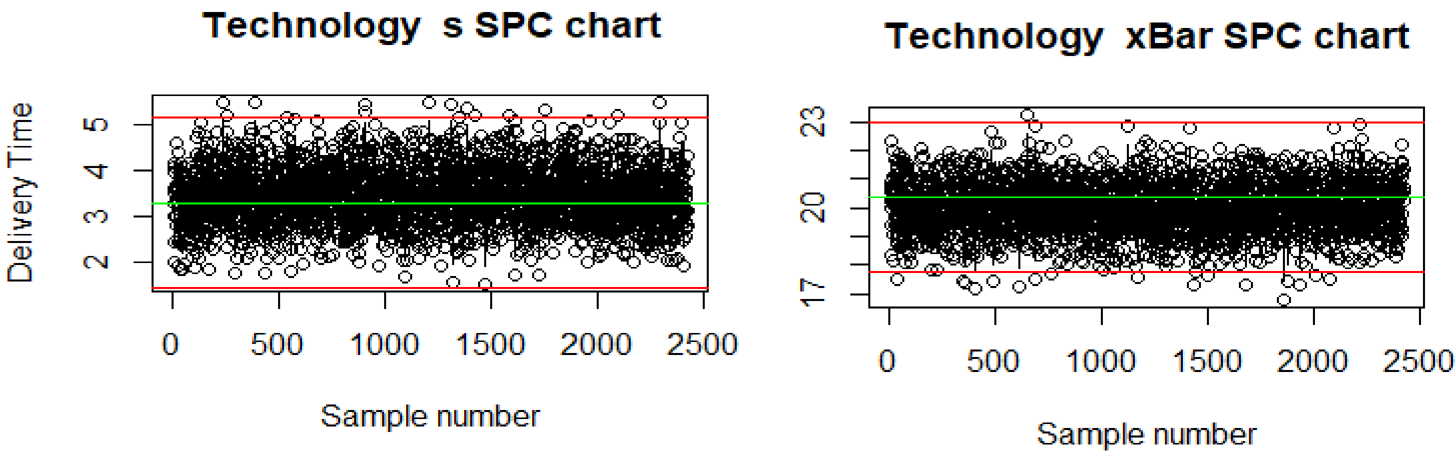


Figure 19: Complete S&X chart for technology

Majority of the samples for technology are in statistical control with the occasional sample falling just outside the limits. The company should work on getting all samples within the limits because technology produces the highest revenue, having consistent delivery times for this class is crucial to keeping loyal customers

3.4.5 Sweets

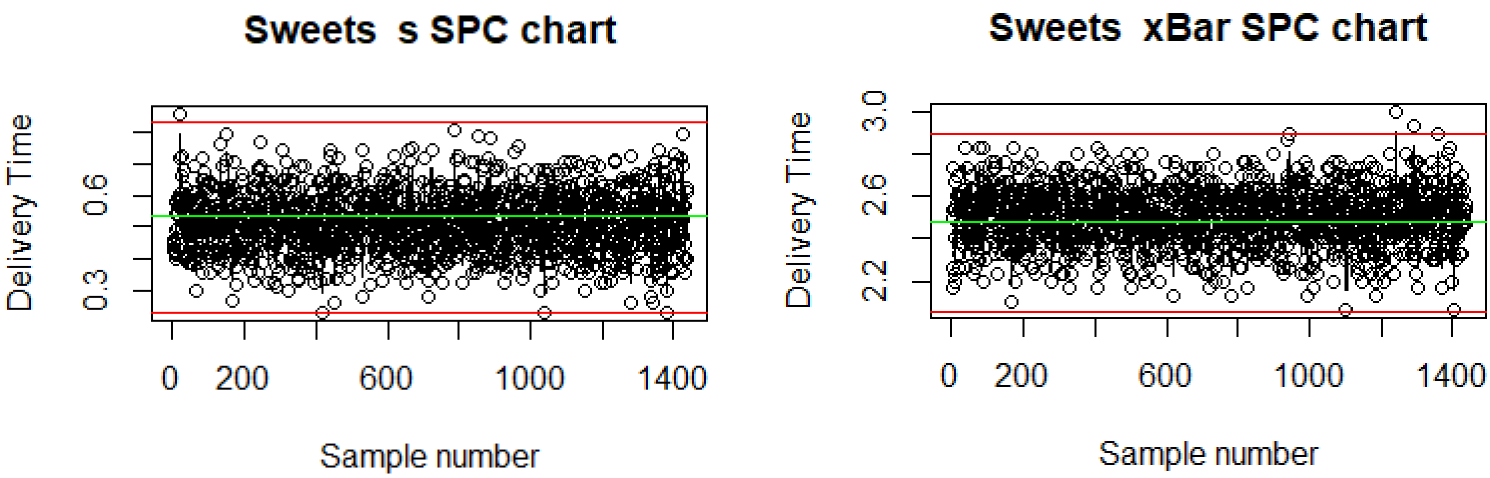


Figure 20: Complete S&X chart for sweets

Sweets delivery times are performing well with one of the samples having a standard deviation out of the control limits and two samples with a mean out of the limits.

3.4.6 Gifts

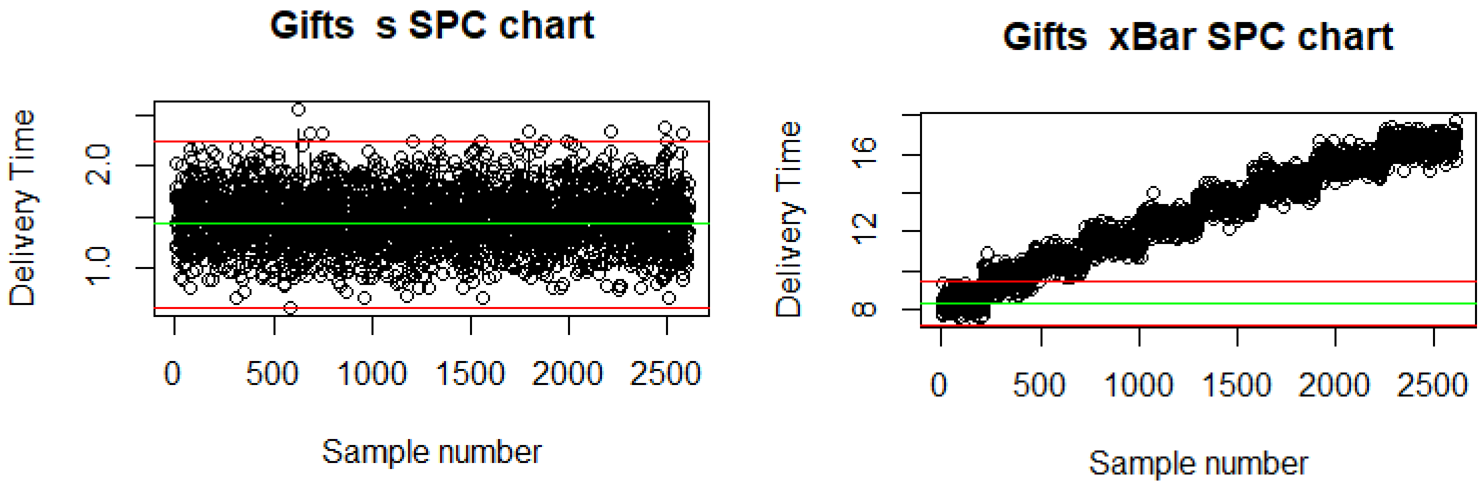


Figure 21: Complete S&X chart for gifts

The X-chart of the delivery times of gifts, shows that this delivery process is not in statistical control. There is a clear upward trend in the delivery times. Most of the samples fall above the upper control limit. This is bad for the business as longer delivery times are associated with poor quality service. The gift class is an important class as it is the most frequently sold item. With delivery times increasing, the business may lose loyal customers. The business main reason for purchases is customers recommending the store, late delivery times will cause customers to lose faith in the company and not recommend their services. The business must implement new strategies in order to decrease the delivery times of gifts.

3.4.7 Luxury

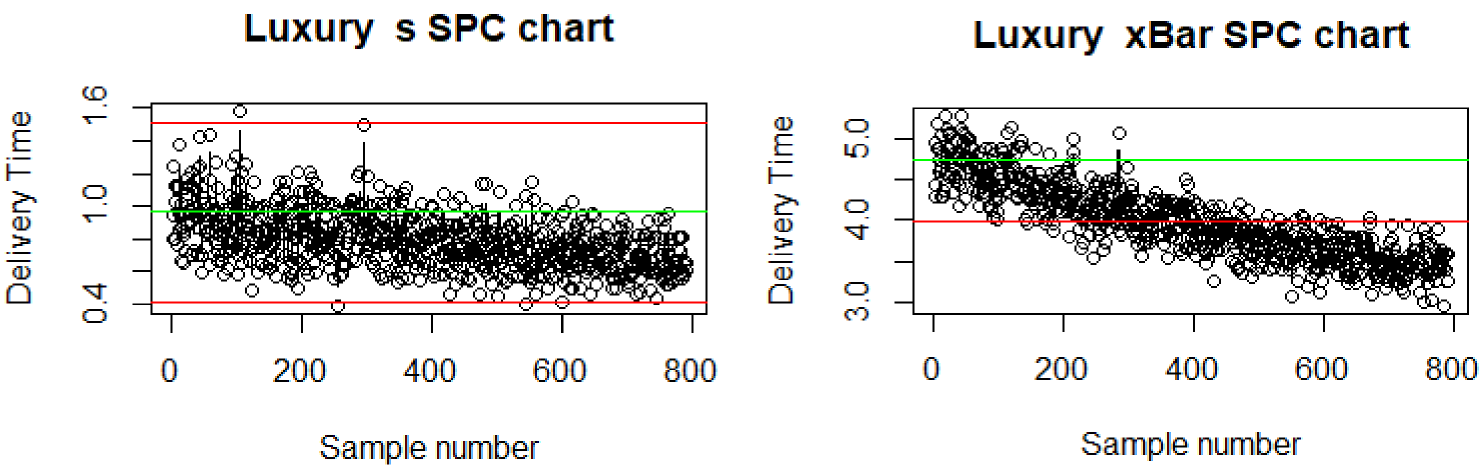


Figure 22:Complete S&X chart for luxury

The delivery times for the luxury class are out of control. Many of the samples fell below the lower control limit. This is not necessarily bad because the graph follows a downward trend, resulting in clients get their luxury items faster. This will lead to an increase in customer satisfaction which will promote customer loyalty and increase recommendations for the store. This should still be investigated to see whether the change was done on purpose. New control limits need to then be calculated.

4.Part 4

Part 4 focuses on optimising the delivery process using delivery times of all the samples.

4.1A: Samples with mean delivery times outside control limits

Class	Samples Outside of limits	1st Out	2nd Out	3rd Out	3rd Last Out	2nd Last Out	Last Out
Clothing	20	282	837	1048	1695	1723	1756
Household	395	252	387	643	1335	1336	1337
Food	4	75	432	1149	NA	NA	NA
Technology	19	37	345	353	1933	2009	2071
Sweets	4	942	1243	1294	NA	NA	NA
Gifts	2287	213	216	218	2607	2608	2609
Luxury	440	142	171	184	789	790	791

Table 7: Samples with mean delivery times outside the control limits

Table 6 above, clearly shows that households and luxury have a large number of samples that are out of control. With the gift class having a very large number of samples out of control,2287 samples fall out of the upper control limit. This is clearly seen in Figures 17,21,22 which confirms the tables finding of these classes having many samples outside of the upper and lower control limits.

Luxury and gifts have the highest number of samples outside of the control limits, 440 and 2287 respectively. Looking at figures 5 and 6 above we can see that both luxury and gifts are very important classes in the business with luxury generating the second highest revenue out of all the classes and gifts being the class with the highest frequency of sales. This is an area of concern for the business, management should investigate these two items to find where the problems lie with the delivery of these items. The rest of the classes show few samples that fall out of the limits.

4.1B:

Class	Most consecutive samples in range	Last sample in range
Clothing	4	223
Household	3	45
Food	4	223
Technology	6	1776
Sweets	4	1243
Gifts	7	2477
Luxury	4	63

Table 8:Consecutive samples in given range

The most consecutive samples between the -0.3 and 0.4 sigma control limits indicate the stability of the delivery process. The more consecutive samples that fall within an interval, the longer the process is stable. When there are many consecutive samples that lie between the -0.3 and 0.4 sigma control limits, then the company can be certain that there is little variation in the process. This means that the company can confidently promise delivery times they know they can deliver. Meeting promised delivery times will help grow customer loyalty to the business and increase recommendations. Table 8 shows that for all the classes there are few consecutive samples before a sample fall outside of the control limits. The company needs to work on decreasing the variability of delivery times for all classes.

4.2. likelihood of making a Type I Error

A Type I error is the probability of rejecting a true null hypothesis. In this case if we reject our null hypothesis, we are saying process is unstable, where in fact the process is stable.

Ho – Stable process (the process is in control and centred on the centreline calculated using the first 30 samples.)

H1 – Process is not stable (process in not in control and has moves from the centreline or has increased or decreased in variation.)

4.2.1 Probability of type 1 error for 4.1 A

Probability of type 1 error for 4.1 A = pnorm(-3)*2

= 0.002699796

= 0.2699796%

The probability of a type 1 error is 0.2699796%

Therefore, there is a 0.27% chance of rejecting a true null hypothesis. This means that there is a 0.27% chance that a process will be classified as not stable where it is stable. This will result in unnecessary time and money being spent on fixing a process that did not need to be fixed.

4.2.2 Probability of type 1 error for 4.1 B

Probability of type 1 error for 4.1 B = pnorm(1.2, lower.tail = FALSE) + pnorm(-0.5)

= 0.4236072

= 42.36072%

The probability of a type 1 error is 42.36072%.

Therefore, there is a 42.36072% of rejecting a true null hypothesis.

4.3 Optimal delivery time for delivery process

If it takes longer than 26 hours for a technology class item to get delivered, it is considered as a late delivery and the company incurs a cost of R329 per item per hour delivered late. it costs the business R2.5/item/hour to reduce the average time by one hour. The optimal number of hours that the delivery time should be reduced by in order to reduce cost for late delivery needs to be calculated.

The company currently incurs a cost of R758 674 for late deliveries of technology. By running multiple different scenarios to find the optimal number of hours to reduce the delivery time for lowest cost resulted in reducing the delivery time by 7 hours. This leads to no late deliveries for technology. For the company to reduce these delivery times they must pay R2.5 for each delivery time that must be reduced. There are 1356 delivery times that are late this means that the company will have a total cost of R23 730. If the company introduces the improved delivery times, they will save R734 944.

4.4 Type ll Error

Type ll error is the probability of failing to reject a false null hypothesis. Meaning we fail to investigate a process, but unknown to us the process is unstable.

The likelihood of making a type II (Consumer's) Error for the technology class given that the delivery process average moves to 23 hours. The type II error would occur when the Ha is true, but we fail to identify this, due to the sample Xbar value being between LCL and UCL. The type two error is calculated as follows:

(pnorm(UCL_for_Technology,mean=23,sd=(UCL_for_Technology-LCL_for_Technology)/6)-pnorm(LCL_for_Technology,mean=23,sd=(UCL_for_Technology-LCL_for_Technology)/6))*100 = 48.83177

The probability of a Type ll error is 48.83%. This means that there is a 48.83% chance that we may assume that the technology product was delivered on time given it was late. This probability is relatively high, and the company should take action to ensure that they don't fail to investigate processes that are out of limits. If the company does not address this, they will lose customers from late deliveries they are unaware of.

5. Part 5. MANOVA

MANOVA is a well-known technique that can be used to determine the impact of different features on one another. The influence of the dependent variables - delivery time, year, and price on the independent variable - class, is investigated. The constructed MANOVA has the following hypotheses.

H0: Delivery time, year and price made no significant change to the buying pattern for each class.

H1: At least one feature has an influence on the buying pattern.

After the MANOVA was run in R studio a p-value of 2.2e-16 was produced. Using an alpha value of 0.05, $p<0.05$. This indicates that there are significant differences among groups on a linear combination of the dependent variables. This means that the null hypothesis should be rejected. From this we can conclude that at least one of the features has an impact on class.

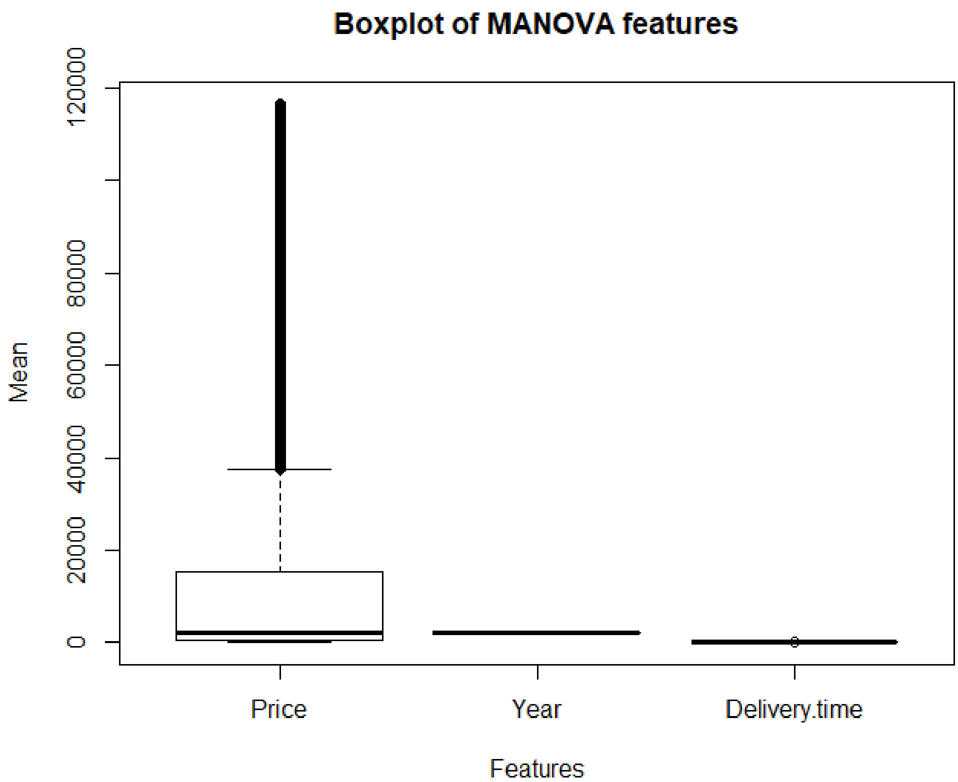


Figure 23: Distribution of MANOVA classes

Figure 16 above shows the distribution of the different features used for in the MANOVA. The different features mean, and distributions are clearly not similar and thus the null hypothesis must be rejected. Agreeing with the hypothesis test.

6. Part 6

6.1 Taguchi loss function

6.1.1 Problem 6

Food deliveries must be kept cool during transit. The company has a subsidiary, Lafrideradora, who makes components for their units. The blueprint specification for the thickness of a refrigerator part at Lafrigeradora, Inc. is 0.06+-0.04 centimetres (cm). It costs $45 to scrap a part that is outside the specifications. The Taguchi loss function for this problem will be determined.

$$L = k(y - m)^2$$

$$L = 45$$

$$k = \frac{45}{0.04^2}$$

$$k = 28125$$

$$L = 28125(y - 0.06)^2$$

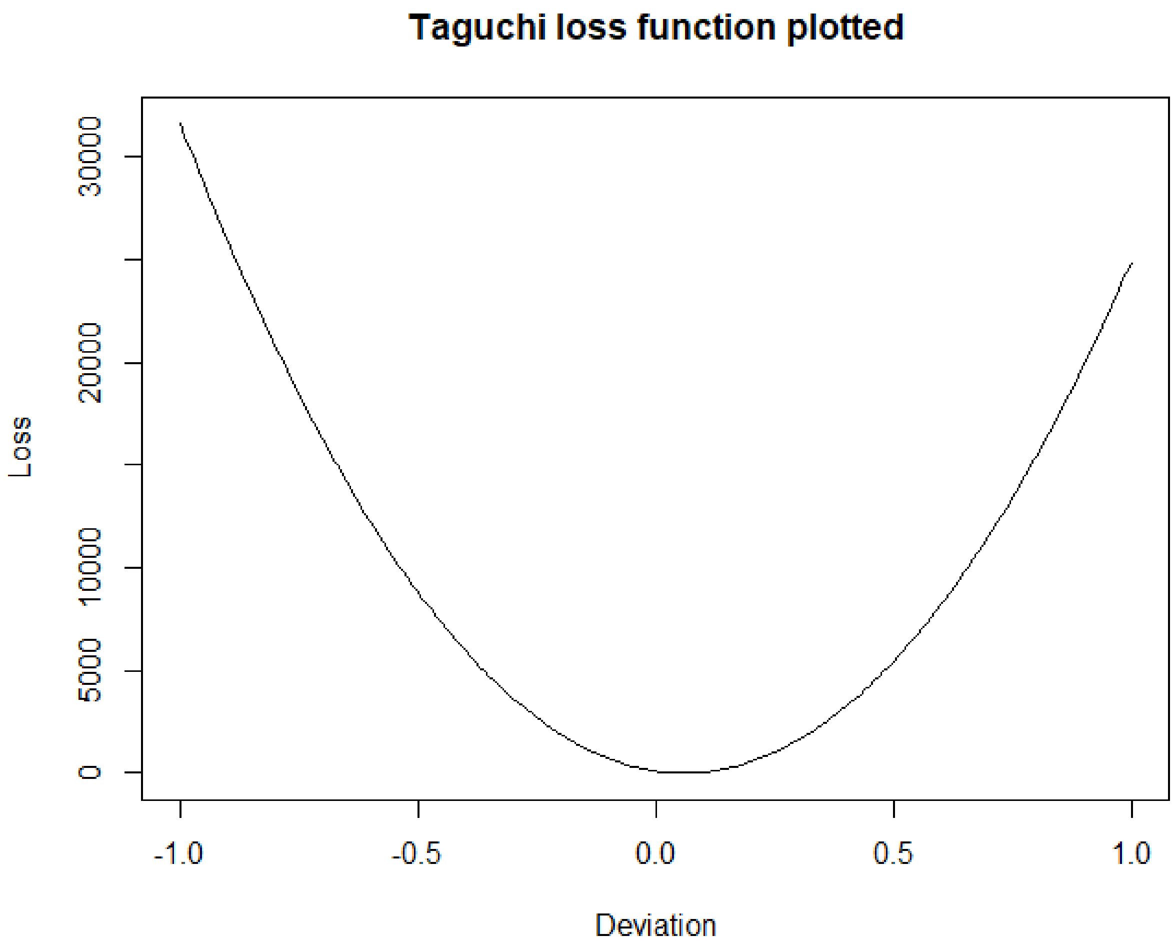


Figure 24:Taguchi loss function problem 6

Figure 24 graphs deviation to loss. As deviation increases your loss increases hence the more expensive it becomes. The business must aim to keep the deviation as low as possible.

6.1.2 Problem 6 A

A team was formed to study the refrigerator part at Lafrigeradora, Inc. described above. While continuing to work to find the root cause of scrap, they found a way to reduce the scrap to 25$ per part. The Taguchi loss function is calculated.

$$k = \frac{35}{0.04^2}$$

$$k = 21875$$

$$L = 21875(y - 0.06)^2$$

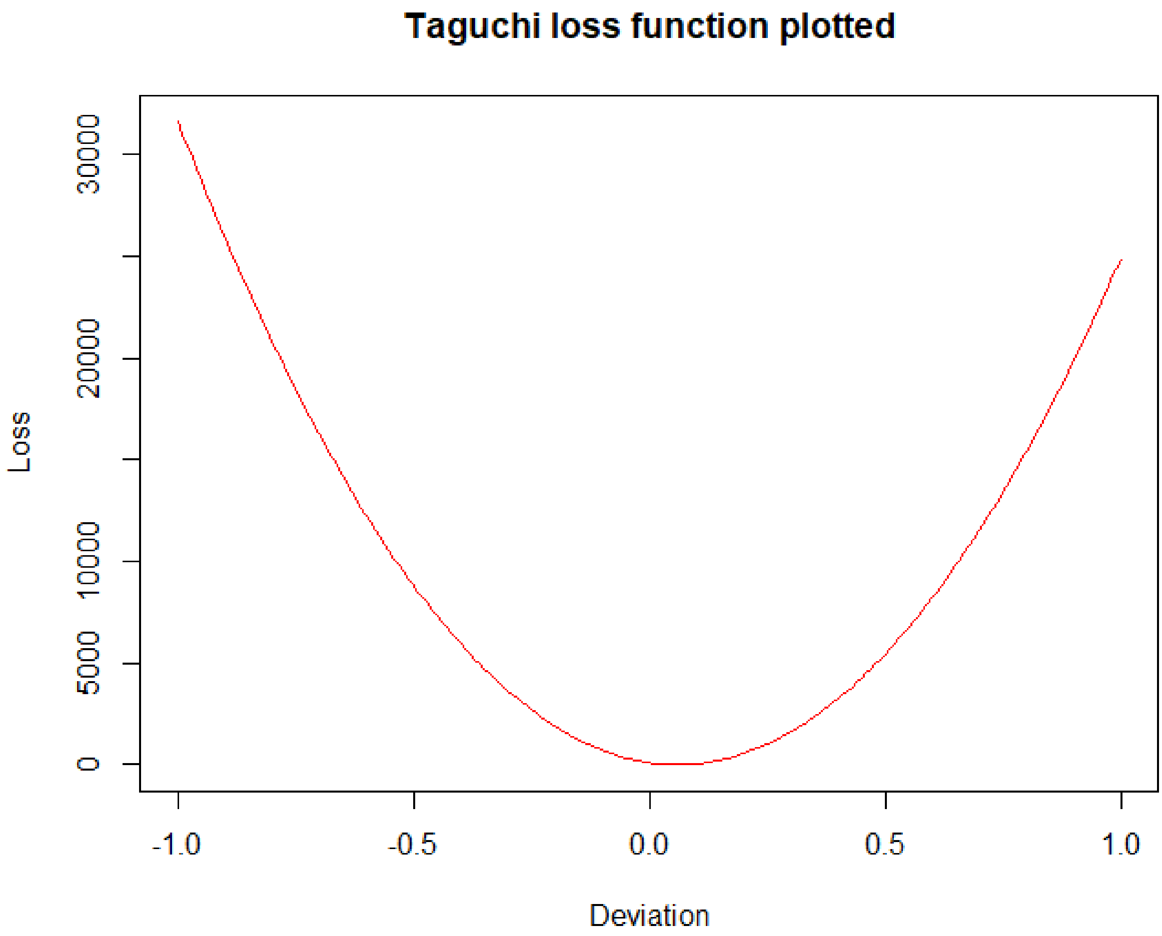


Figure 25: Taguchi loss function problem 6A

Management must continue with efforts to reduce deviation which will result in a decrease of loss.

6.2 Problem 6 B

The deviation is reduced to 0.027 which results in a Taguchi loss of:

$$L = 21875 \times 0.027^2$$

$$L = 15.95$$

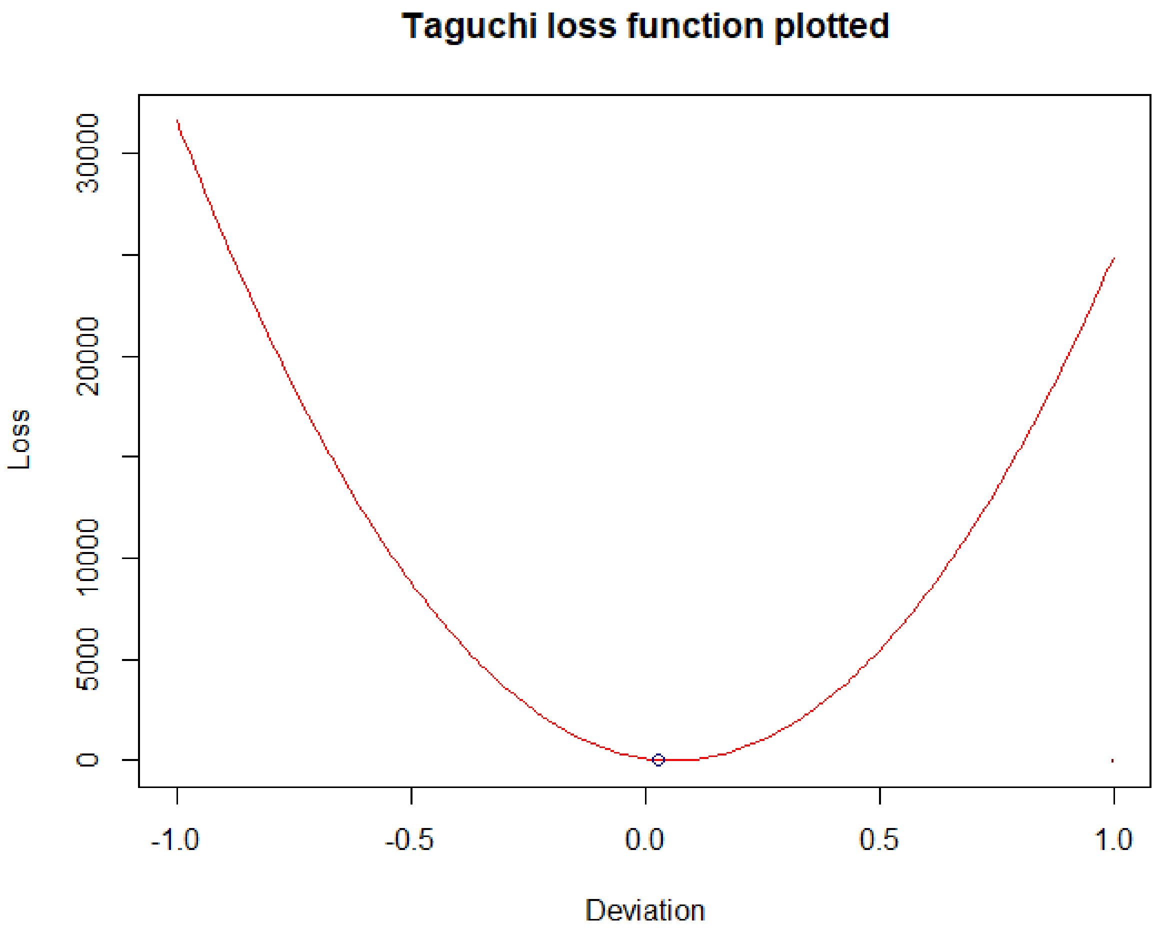


Figure 26: Taguchi loss function problem 6B

The point of deviation corresponding to the loss of 15.95 is shown in figure 18. As shown on the graph this process is doing well. Lafrigeradora should try maintain this deviation and improve it where possible.

6.3 Part 6.2

Magnaplex, manufactures some your technology items. Management thinks their current process, shown in figure 26, is wasteful due to the identical machines used as backup in case of failures.

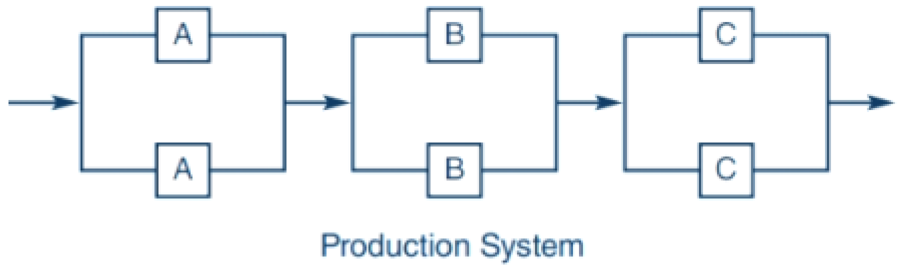


Figure 27:Current production system

The reliability of each machine is shown in figure 27 below.

Machine	Reliability
A	0.85
B	0.92
C	0.90

Figure 28: Reliability of each machine

To evaluate whether management is right a comparison will be made between the reliability of the process with identical machines in parallel compared to only one machine each.

If only one machine each in A, B and C is working, the reliability is simply: RA*RB*RC = 0.85*0.92*0.9 = 0.7038

If machines are working in parallel shown by the current system in figure 26 then if one of the machines breaks down the system can still continue. For components A, their combined reliability is 1 – probability(both fail). Both A's fail with a probability (1-0.85)2 = 0.0225 and then their combined reliability is RAA = 1- 0.0225 = 0.9775 Both B's fail with a probability (1-0.92)2 = 0.0064 and then their combined reliability is RBB = 1- 0.0064 = 0.9936 Both C's fail with a probability (1-0.9)2 = 0.01 and then their combined reliability is RCC = 1- 0.01 = 0.99 The reliability is then simply: RAA*RBB*RCC = 0.9775*0.9936*0.99 = 0.9615. This is a 26% improvement from only having one machine each in the production system.

6.4 Vehicle availability

The company has 20 delivery vehicles available, of which 19 is required to be operating at any time to give reliable service. During the past 1560 days, the number of days that there were only 20 vehicles available was 190 days, only 19 vehicles available was 22 days, only 18 vehicles available was 3 days and 17 vehicles available only once.

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

Figure 29:Binomial equation

1560-190-22-3-1 = 1344 days we had all vehicles available.

The p value for defective vehicles is calculated using the equation shown in Figure 29.

dbinom(0, 21, prob= p, log= FALSE)*1560-1344 = 0 , this results in a p value of :

p = 0.007071661

19 vehicles are required to be operating at any time to give reliable service. Therefore, we look at the probability of having 2 or less cars available in a day.

pbinom(2,21,p) = 0.9996264, there is a 99,96% chance that we will have reliable delivery in a day.

6.4.1 Drivers' availability

There are 21 drivers, who each work an 8-hour shift per day. During the past 1560 days, the number of days that there were only 20 drivers available was 95 days, only 19 drivers available was 6 days and only 18 drivers available, once only.

1560-95-6-1 = 1553 days we had all workers available.

The p value for drivers is calculated using the equation shown in Figure 29.

dbinom(0, 21, prob= p, log= FALSE)*1560-1553 = 0 , this results in a p value of :

p = 0.007071661

19 vehicles are required to be operating at any time to give reliable service. Therefore, we look at the probability of having 2 or less drivers available in a day.

pbinom(2,21,p) = 1 there will always be enough drivers in a day for the 19 vehicles.

6.4.2 Expected reliable delivery times in a year

$0.9996264 \times 1 \times 365 = \ 364.8636$

There will be 364 reliable delivery days in a year. Only one day in the year may have unreliable delivery days this is a good sign for the business. Further improvements can be made to get a 100% reliability.

Part2: Increased number of vehicles

An extra vehicle is purchased by the company resulting in 22 vehicles in total.

The same process is covered previously to calculate the new p value using 22 vehicles.

dbinom(0, 22, prob= p, log= FALSE)*1560-1344 = 0

p = 0.00675129

19 vehicles are still required to be operating at any time to give reliable service. Therefore, we look at the probability of having 2 or less cars available out of 22 vehicles in a day.

pbinom(2,22,p) = 0.9999887, there is a 99,99887% chance that we will have reliable delivery in a day.

6.4.4 New reliable delivery time in a year

$0.9999887 \times 1 \times 365 =$ 364.9959

A new vehicle still does not get 365-day reliability exactly but is very close to 100% reliability.

Conclusion

The data for the online business was cleaned of all invalid instances an analysis was done on the different classes of products in the business.

After analysing the data, it is clear that luxury and technology produce a significantly higher total revenue then any other class. The business should consider removing items such as sweets and food which produce very little revenue in comparison to the other classes. This will help the business focus on the more profitable classes.

The business has created a loyal customer base with most of the reasons behind sales coming from customers that were recommended to the store. To ensure this customer base continuous to grow the business must make sure that they continue to meet and exceed customers' expectations in terms of service. One of the most important aspects of our service to our customers is reliable delivery times.

After analysing the X and S charts of delivery times for different classes it is clear that gifts delivery time is a major concern for the business. Gifts is the most frequently sold item. This means that there are a large number of customers getting their deliveries late. The business must take immediate action to create more reliable delivery times for gifts.

The probability of making a type one error is significantly smaller than making a type two error. The company should therefore create better processes to ensure that their delivery was actually on time rather than assuming their delivery was on time where it was late.

In conclusion, business must focus on luxury and technology products. As well as increasing all efforts to meet customers expectations, especially with reliable delivery times. With immediate attention on creating more reliable delivery times for gifts.

References

Sthda.com. 2022. *MANOVA Test in R: Multivariate Analysis of Variance - Easy Guides - Wiki - STHDA*. [online] Available at: <http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance> [Accessed 17 October 2022].

Cran.r-project.org. 2022. *Index of /web/packages/MANOVA.RM/vignettes*. [online] Available at: <https://cran.r-project.org/web/packages/MANOVA.RM/vignettes/> [Accessed 17 October 2022].

Hernandez, F., 2022. *Data Analysis with R - Exercises*. [online] Fch808.github.io. Available at: <http://fch808.github.io/Data-Analysis-with-R-Exercises.html> [Accessed 17 October 2022].