UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

Quality Assurance Project Report

Written for:

Engineering Counsel of South Africa (ECSA)

Written by:
Jani Louw
22850708

Date due:

21 Oct 2022

saam vorentoe · masiye phambili · forward together

Department of Mechanical and Mechatronic Engineering
Departement Meganiese en Megatroniese Ingenieurswese
Privaat Sak X1, Private Bag X1, Matieland, 7602
Tel: +27 21 808 4204 | www.eng.sun.ac.za

ENGINEERING
EZOBUNJINELI
INGENIEURSWESE

Abstract

The following report is written for the Engineering Council of South Africa. The report will analyse the data from a dataset provided from a statistical stand point. The report will provide a detailed explanation of the data wrangling process and describe the data provided graphically. The process capabilities of a specific feature will also be analysed and described. Xbar and sbar charts will be provided to further analyse the delivery times of the first 30 and then all the entries of the various product classes. A DOE and MANOVA test will also be done to analyse the connection between specific features within the data. The report will then conclude with extra probability questions from the textbook and a brief summary of the report.

Table of contents

List of figures

List of tables

Introduction

The following report is written for the Engineering Council of South Africa. The report contains the analysis of a dataset, called sales.Table2022, from a statistical point of view. The report begins with the details of why and how the data wrangling process took place as well as a descriptive statistical evaluation of the data. The process capabilities of specific features in the data are also analysed. The report then continues to describe the evaluation of the data using x bar and sbar charts for the first 30 and then all the samples provided in the dataset. Additionally a DOE and MANOVA test is done to further evaluate the connection between the delivery time and price of each product and the class of the various product classes. The report concludes with probability questions from the prescribed textbook and a summary of the findings made throughout the report.

1 Part 1: Data Wrangling

Data wrangling refers to the processes that are designed to transform raw data into formats that are more readily used. For this project, the data wrangling/ cleaning process followed the following steps:

1. Discovery: The data had to be thoroughly looked over and conceptualised to identify data issues and the data's uses. The function, "summary", was mainly used in R to identify and gain knowledge on possible data issues.

2. Structuring: The data was originally mis formatted for its intended application, as the Year, Month and Day columns were not ordered. To rectify this, the data was ordered form least recent to most recent.

3. Cleaning: The of data contained missing values and NA (Non-Applicable data), all of which had to be removed for computational purposes.

Summary:

The original data set (salesTable2022) contained 180000 rows and 10 columns of data. After the completion of the data wrangling process and the addition of a new numbering column, a valid data set was available for all the necessary computations. The valid data set consists of 179978 rows and 11 columns of data. The invalid data set consists of 22 rows and 11 columns of data.

2 Part 2: Descriptive statistics

2.1 Price of product classes sold by company

It is very important for a business to get the pricing of a product correct as it has the greatest impact on profit. The more valuable an item is, the greater the cost price will be, leading to a higher the selling price. A seen in figure 1 below, the most expensive items sold by the company are the luxury items. The product class also has the largest price range in comparison to the other types of products the company sells.

The items with the lowest selling price and smallest price range are sweets. Food and clothing are the second and third cheapest items sold and have low price ranges. Gifts and household products are the third and forth cheapest products but the jump in price range from one to the other, is quite large. Technology is the second most expensive product sold.

The trend detected from the figure 1 is that the more expensive an item class is the larger its price range. This is a good approach for the business to use, as they widen their market by making cheaper options available for the customers who cannot afford to purchase the most expensive product from a specific class.

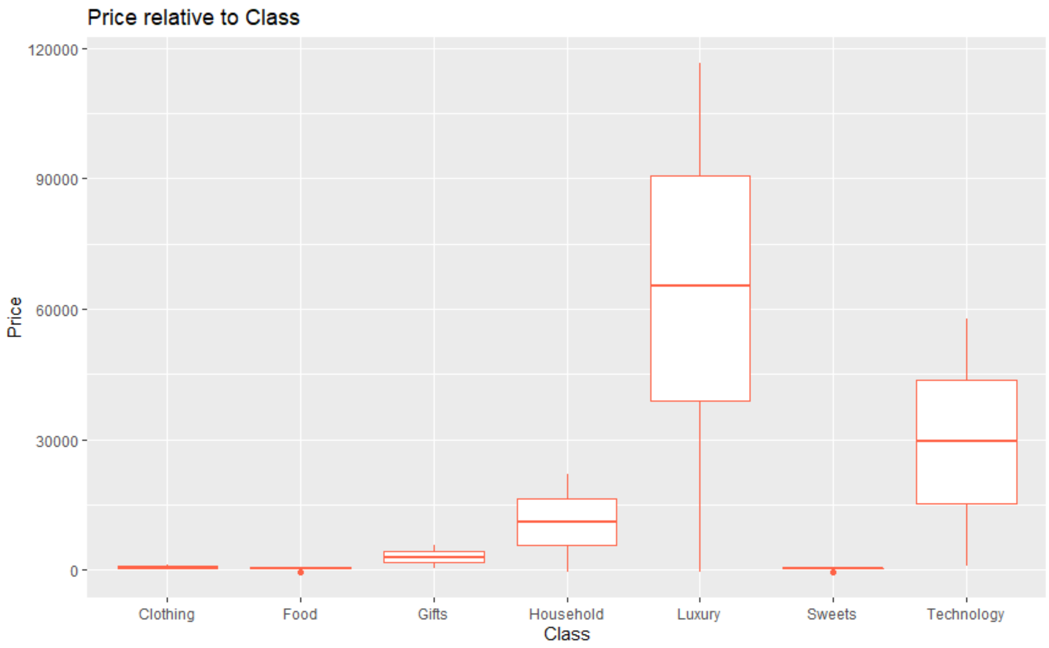

Figure 1: Price relative to Class

2.2 Delivery time of products from product classes

In any business, the delivery time should always be as short as possible to ensure customer satisfaction.

There are many factors that could affect delivery time such as oversea suppliers, inefficient logistics planning, unreliable suppliers, and unreliable delivery systems. From figure 2 below, it is evident that household products have a much longer delivery times as well as a much larger range of time in which the customer may receive their order. This should be a concern to the business since household products are normally products that are used on a daily or weekly basis, which means customers will likely need these items to arrive as soon as possible. Nearer and more reliable suppliers should be investigated.

Products such as luxury items, technology, and gifts are usually products that a customer would not need immediately and therefore can afford to have a longer delivery time because they are not considered necessities.

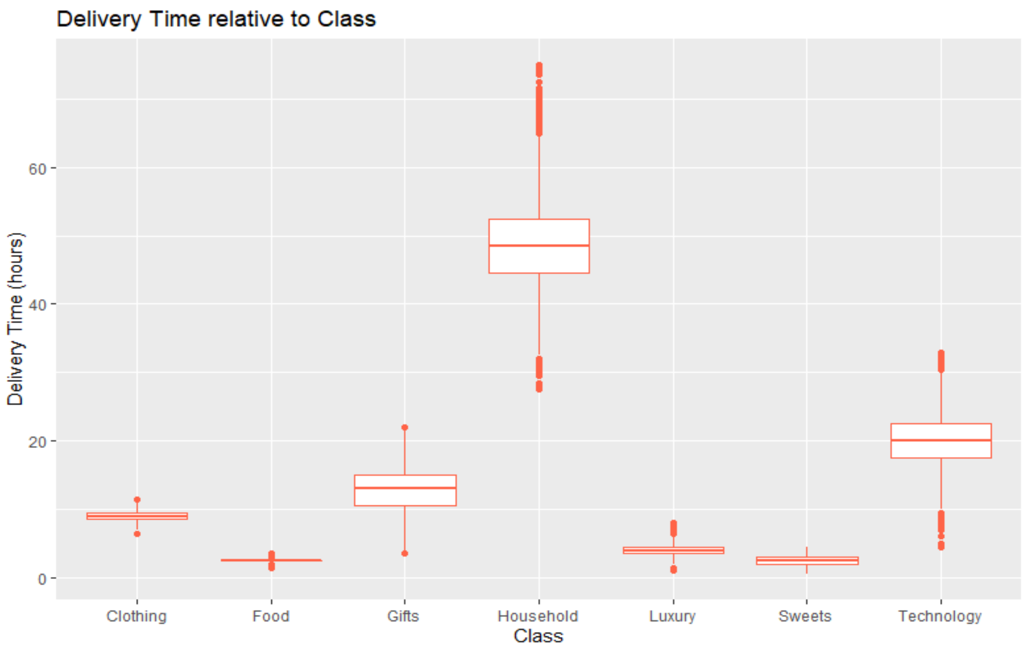

Figure 2: Delivery Time relative to Class

2.3 Yearly sales

From figure 3 below, one can see that the year 2021 had extremely high sales which declined rapidly in 2022. Since 2022 the yearly sales slowly began to increase till 2029 but never again reaching the high from 2021.

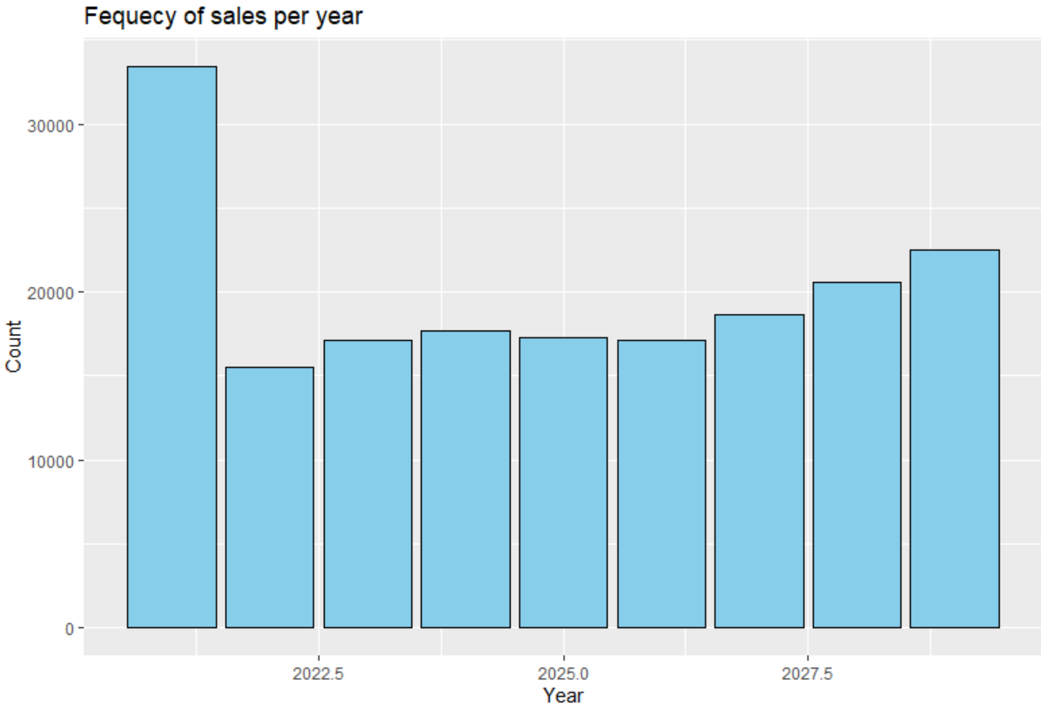

Figure 3: Frequency of sales per year

The above statement can be explained by the figure 4 below. In 2021, clothing and household product sales were very high but both decreased inn 2022. All the other products have had stable sales throughout the years and seem to be only increasing slightly, as they should.

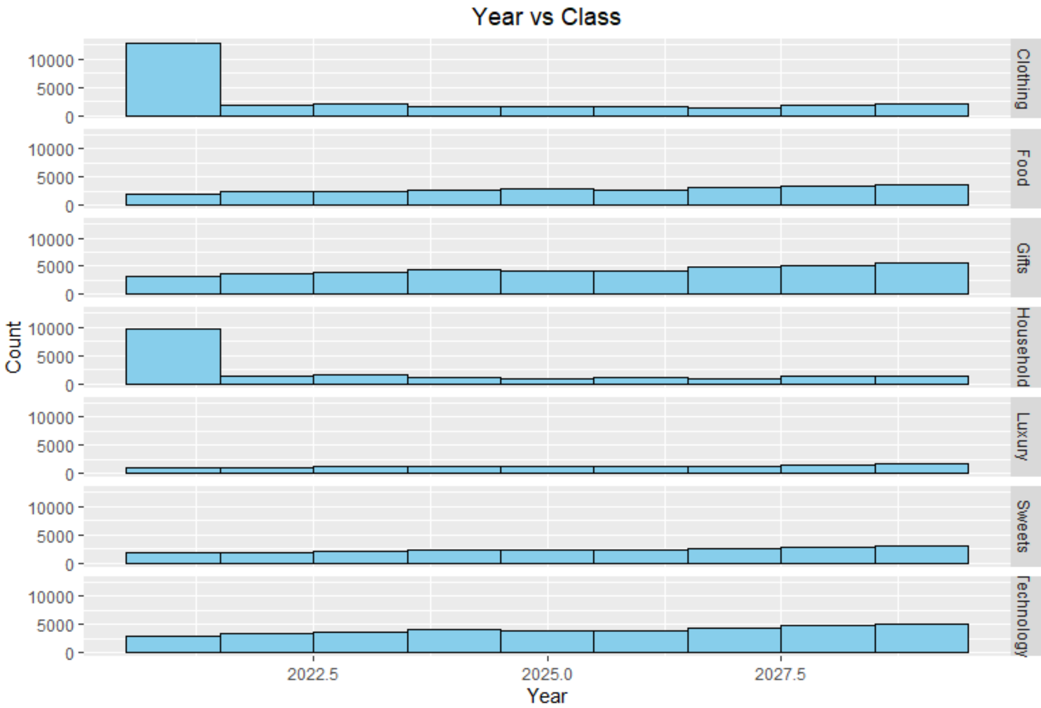

Figure 4: Year vs Class

2.4 Sales of product classes

The business sells various products that can all be categorised into 7 different classes. By analysing the business's sales, the areas where the business should focus on can be determined. From figure 5 below, it is evident that the types of products that are sold most often are gifts and technology.

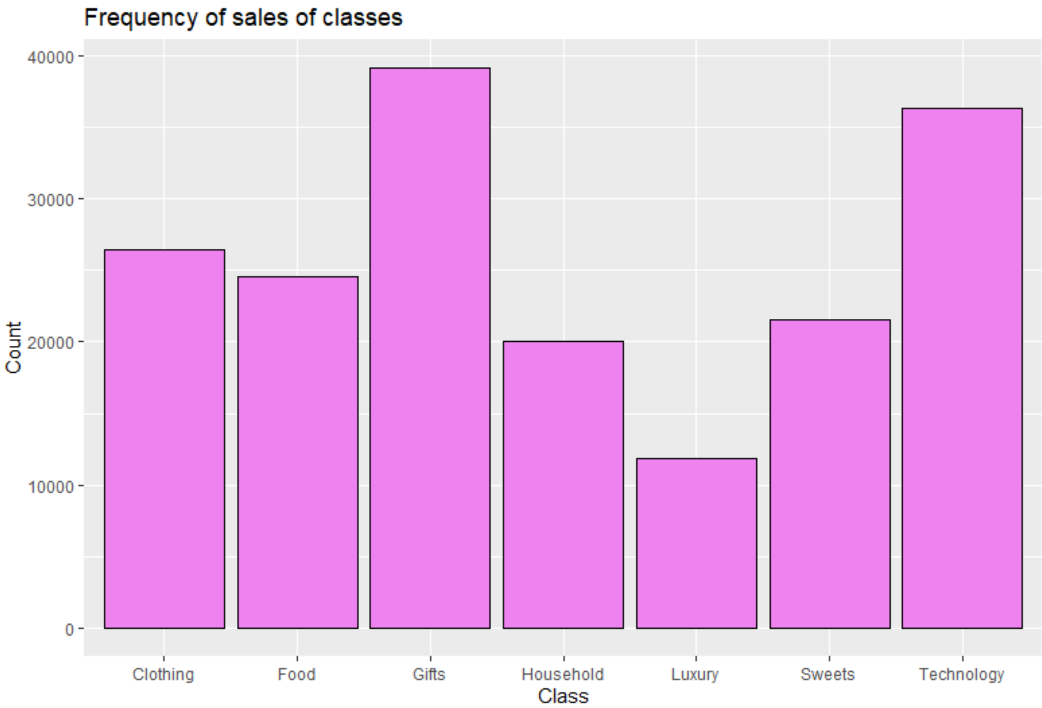

Figure 5: Frequency of sales of classes

2.5 Reasons for customer purchases

There are 6 different recorded reasons as to why a customer would've made a purchase from the business. The various reasons are as follows: browsing, email, random, recommended, spam and website. As seen in figure 6 below, the most common reason across all classes that led to customers making purchases was because the business was recommended to them. This is an indication to the business that keeping their customer service level high must stay a priority.

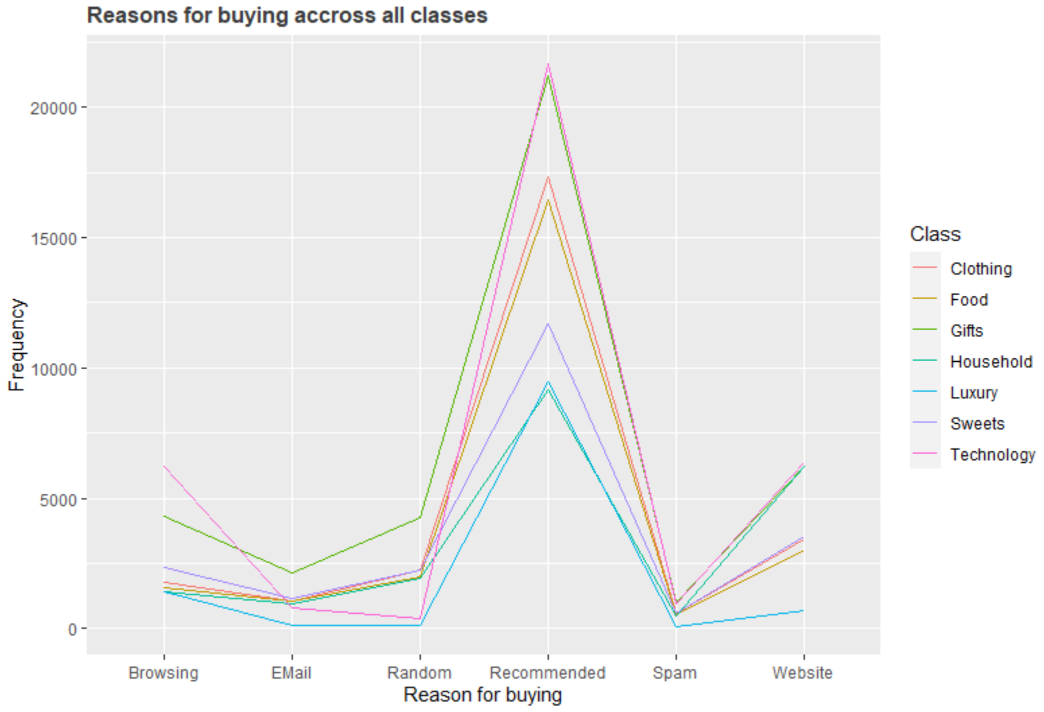

Figure 6: Reasons for buying across all classes

2.6 Ages of customers

It is important for any business to establish target markets for their various products so that they can cater to their customers' need accordingly and therefore increase their sales.

As seen in figure 7 below, clothes are bought most frequently by individuals younger than 30. This can indicate to the business that their marketing should focus on that specific age group.

Food and gift purchases are spread evenly through all ages. Household products are mainly purchased by individuals younger than 60 and has the smallest age range (figure 8). The business should therefore run promotions that will attract customers younger than 60.

Figure 7: Age vs Class

Luxury items are purchased mostly by individuals younger that are around 40. Since they are the most expensive products to buy, the business should look into expanding this market by strategizing to attract a wider age range of customers.

Sweets are hardly purchased by individuals around 40 years of age but has the largest range of customers according to figure 8. Technology purchases decline rapidly for individuals over the age of 60.

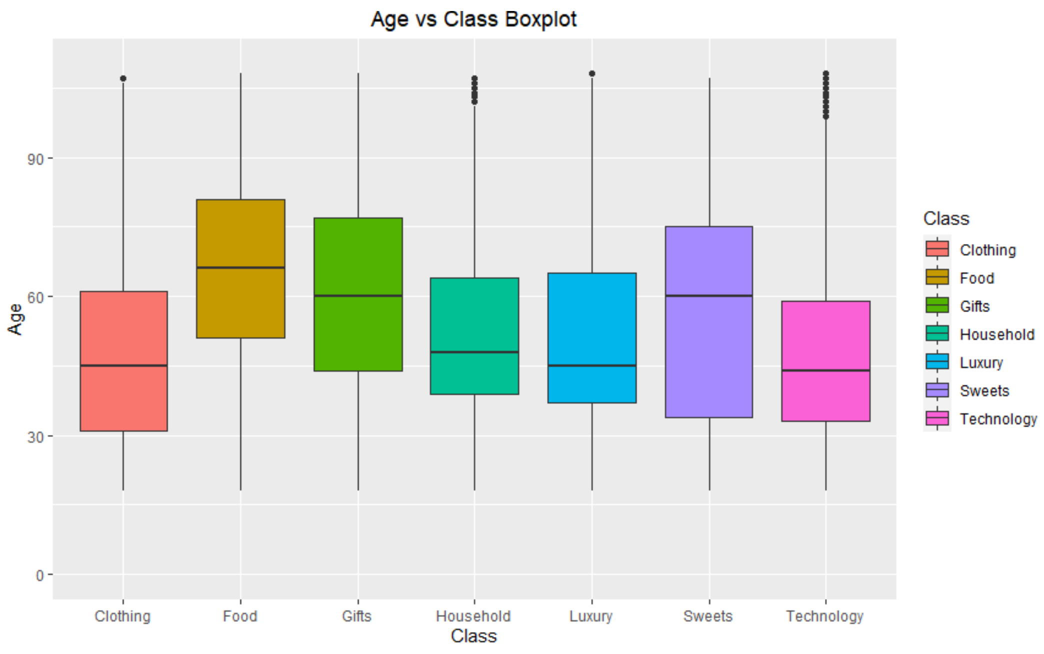

Figure 8: Age vs Class Boxplot

2.7 Process Capability

Process capability indices are used to measure the extent of variation within a process with respect to its specified limits. The indices are also used to assist in the comparison of different processes with respect to reaching given expectations.

The indices required were calculated using the delivery time data for the technology class. The Upper Specification Limit and Lower Specification Limit provided were 24 and 0 respectively. The reason for the Lower Specification Limit being zero is because it is impossible for time to be negative.

The indices were calculated using the next formulas:

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

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

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

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

To perform these calculations the mean (μ) and standard deviation (σ)were calculated as being 20.01092 and 3.501921 respectively.

The results were as follows:

The Cp value is high (more than 1,0) which is an indication that the process is a more capable one and that the specification spread is greater than the process spread. (Hessing, 2021). This mean that the delivery time for technology class items is currently meeting specifications.

The Cpk value is an indication of the location of the data spread and provides more detail on specification achievement than the Cp value. (Hessing, 2021). Since the Cpk value, provided in the table is low (far below 1,0), it means that there are some technology products that are being delivered passed the Upper Specification Limit. This is a bad indication for the business and should be investigated improve customer satisfaction.

Cpu and Cpl indicate the location of the process spread as well. Cpu is related to the distance from the Upper Specification Limit while Cpl is related to the distance from the Lower Specification Limit. (Hessing, 2021)

C_p	C_{pk}	C_{pu}	C_{pl}
1.14223	0.3797038	0.3797038	1.904757

Table 1: Process capability indexes for technology classes

3 Part 3: Statistical process control

3.1 Analysing the first 30 samples

3.1.1 Table values of x charts table for the first 30 samples

The following table was generated to calculate the various ulc, u2sigma, u2sigma, centre line, l1sigma, l2sigma and lcl values for each class, for the xbar charts using data from the 1st 30 samples.

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	9.404934	9.259956	9.114978	8.970000	8.825022	8.680044	8.535066
Clothing	50.248328	49.019626	47.790924	46.562222	45.333520	44.104818	42.876117
Household	2.709458	2.636305	2.563153	2.490000	2.416847	2.343695	2.270542
Luxury	22.974616	22.107892	21.241168	20.374444	19.507721	18.640997	17.774273
Food	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
Sweets	5.493965	5.241162	4.988359	4.735556	4.482752	4.229949	3.977146

Table 2: Table of xbar charts for the first 30 samples

3.1.2 Table values of S charts for the first 30 samples

The following table was generated to calculate the various ulc, u2sigma, u2sigma, centre line, l1sigma, l2sigma and lcl values for each class, for the sbar charts using data from the 1st 30 samples.

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	0.8665596	0.7614552	0.6563509	0.5512465	0.4461422	0.3410379	0.2359335
Clothing	7.3441801	6.4534101	5.5626402	4.6718703	3.7811003	2.8903304	1.9995605
Household	0.4372466	0.3842133	0.3311800	0.2781467	0.2251134	0.1720801	0.1190468
Luxury	5.1805697	4.5522224	3.9238751	3.2955278	2.6671805	2.0388332	1.4104859
Food	0.8353391	0.7340215	0.6327039	0.5313862	0.4300686	0.3287509	0.2274333
Gifts	2.2463333	1.9738773	1.7014213	1.4289652	1.1565092	0.8840532	0.6115971
Sweets	1.5110518	1.3277775	1.1445032	0.9612289	0.7779546	0.5946803	0.4114060

Table 3: Table of s charts for the first 30 samples

3.1.3 S charts for first 30 samples evaluation

The sbar charts for each class was generated using the data from the 1st 30 samples.

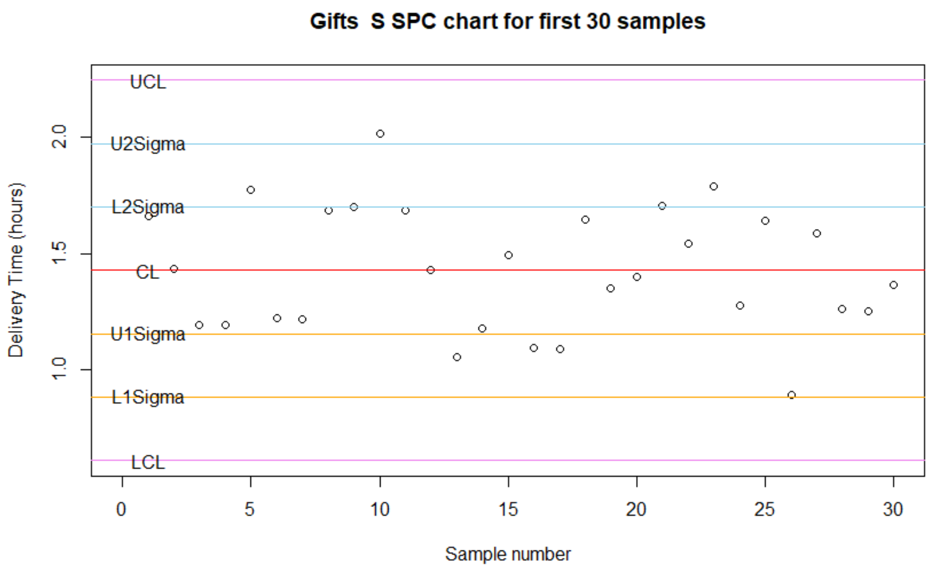

Figure 9: Gifts s SPC chart for the first 30 samples

The chart in figure 9 above is an example of a generally good trend in delivery time as all the values are within the upper and lower control limits.

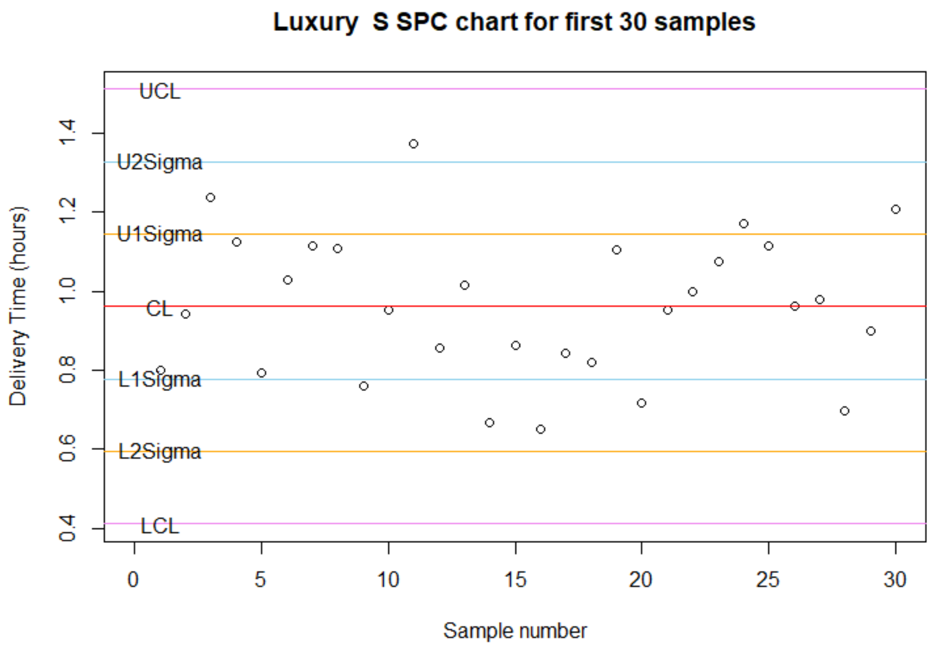

Figure 10: Luxury s SPC chart for the first 30 samples

Although some of the charts of the first 30 samples, such as the one for luxury products in figure 10, do not show any need for concern, the charts are deceiving as one will discover once all the samples are taken into account.

3.1.4 X charts for first 30 samples evaluation

The sbar charts for each class was generated using the data from the 1st 30 samples.

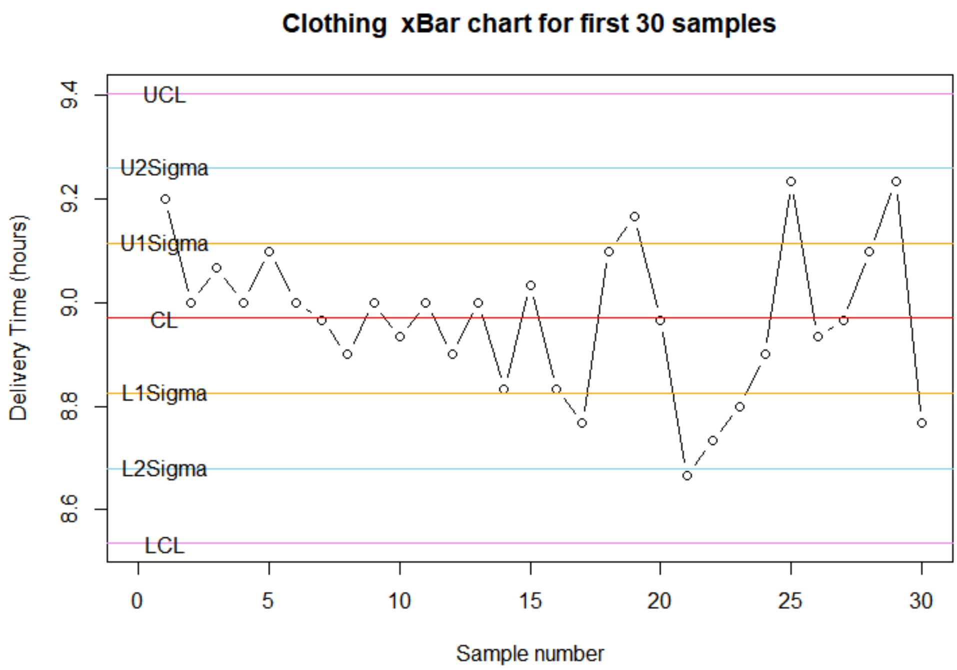

Figure 11: Clothing xbar chart for the first 30 samples

The figure above shows the xbar chart for clothing products, which seems to be losing control as more sample data is considered. This may or may not be a concern for the business but will be easier to evaluate once all the samples are taken into account.

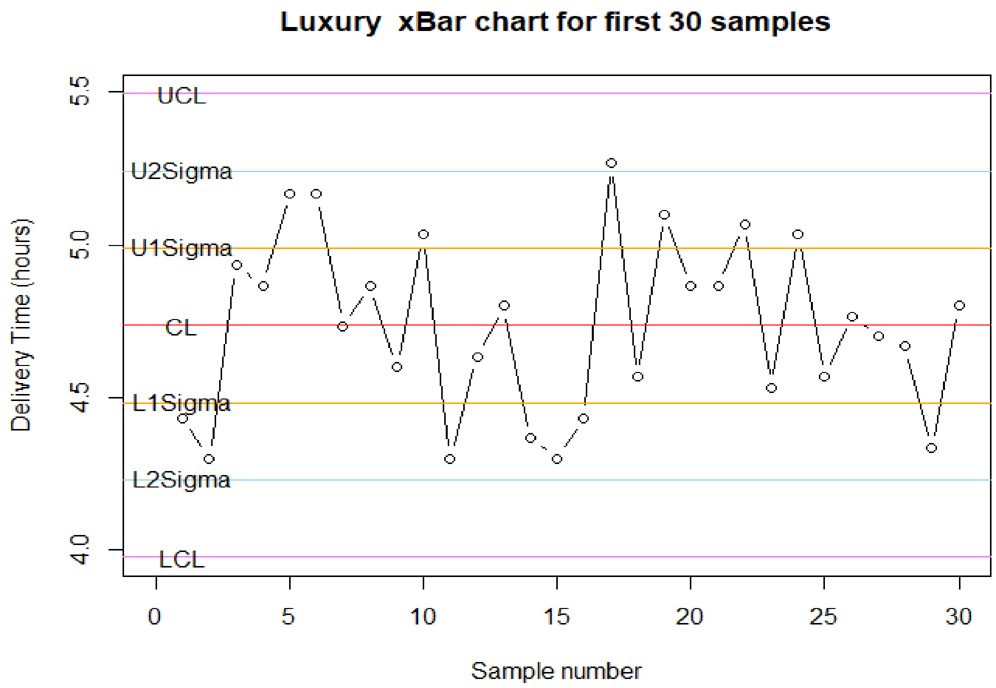

Figure 12: Luxury xbar chart for the first 30 samples

The xbar for luxury products seems to be normal as all the values are well within the upper and lower control limits.

3.2 Analysing all samples

3.2.1 S charts for all samples evaluation

The sbar charts for each class was generated using the data from the 1st 30 samples.

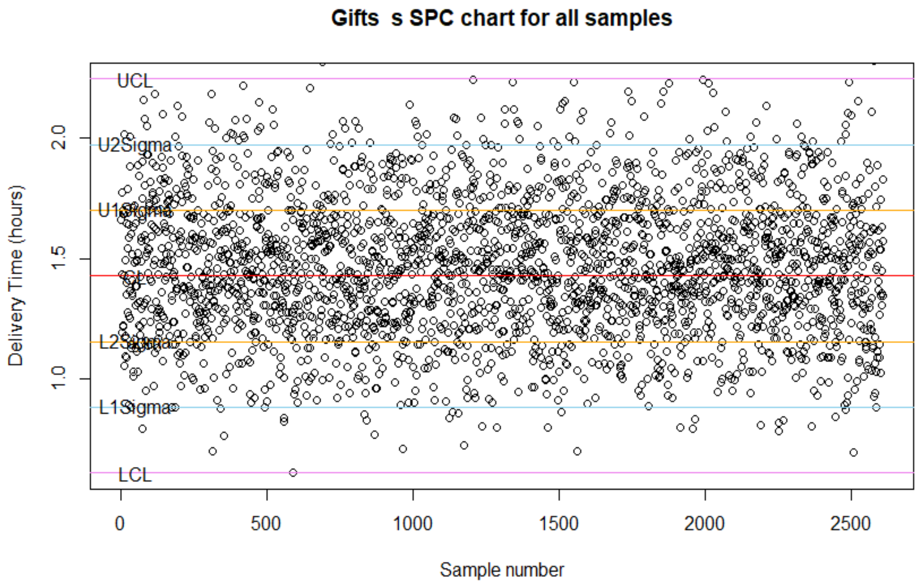

Figure 13: Gifts s SPC chart for all samples

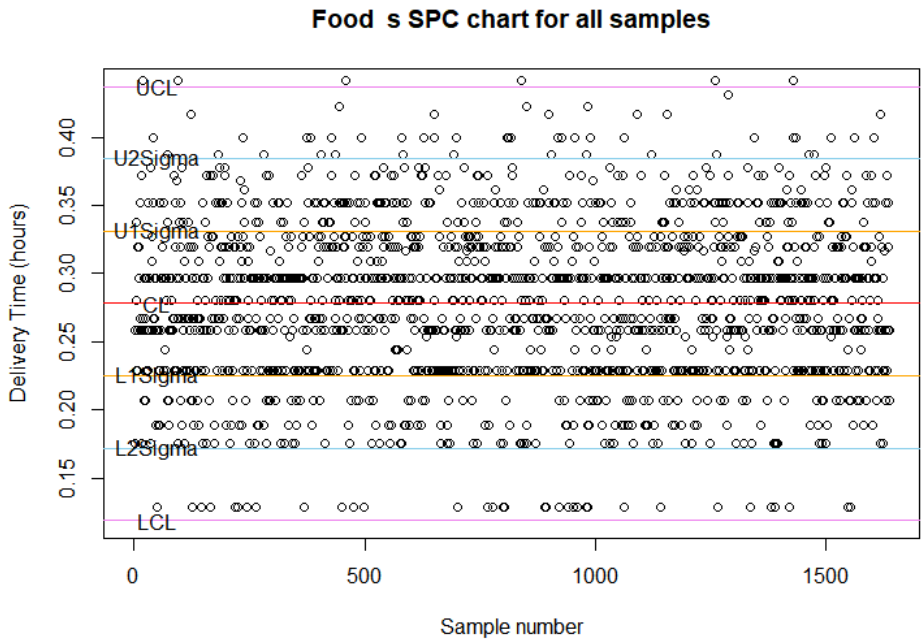

Figure 14: Food s SPC chart for all samples

Both charts in figures 13 and 14 are examples of good delivery time sbar charts as most of the values are within the upper and lower control limits. There are occasional outliers which may be a result of special causes such as loadshedding, breakdowns or late supplier deliveries.

3.2.2 X charts for all samples

The sbar charts for each class was generated using the data from the 1st 30 samples.

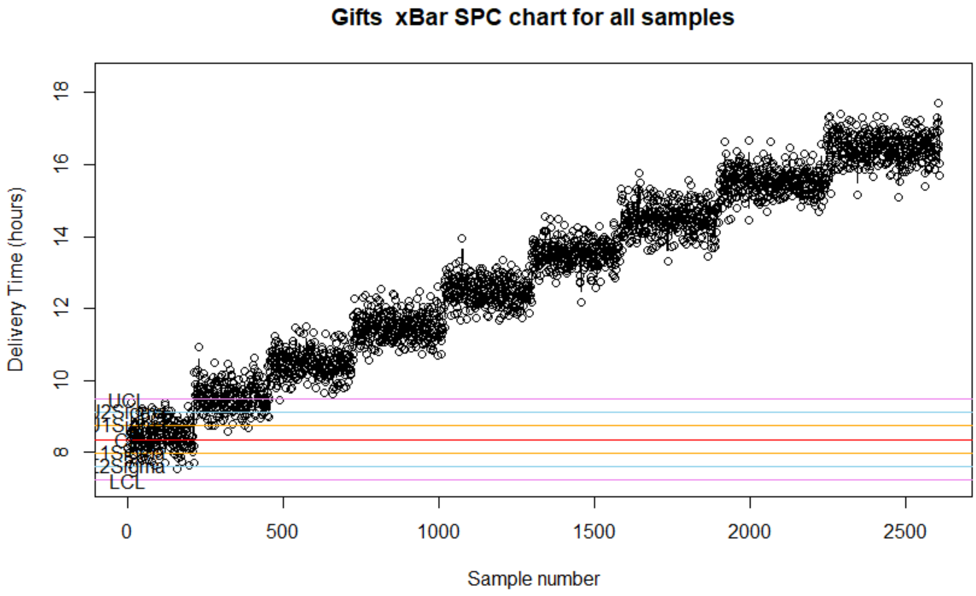

Figure 15: Gifts xbar chart for all samples

Figure 15 above is cause for great concern as multiple samples have recorded excessively out of bound/ late deliveries which should be a concern to the business. The process is very much out of control and way beyond its upper control limit. The mean delivery time is increasing which is something that the business will need to look into.

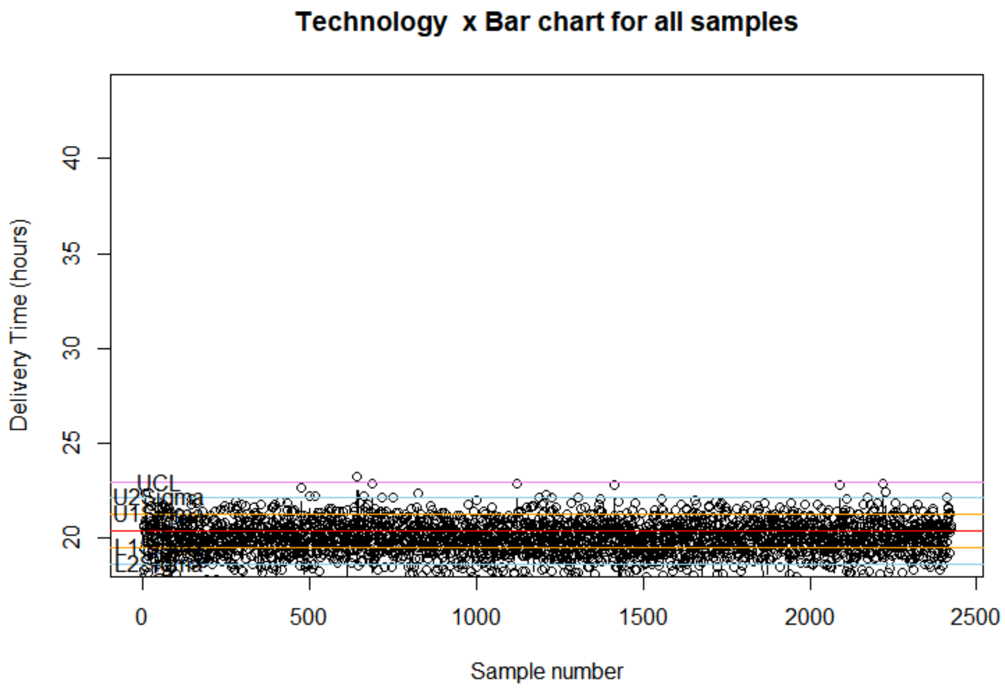

Figure 16: Technology xbar chart for all samples

The figure 16 above is an example of a healthy/ good xbar chart. The process in under control with very few cases of the delivery time being out of the required limits.

4 Part 4:

4.1 Analysins of out of contol process

4.1.1 A: Identifying sample outside of control limits

Class	Total found	1st	2nd	3rd	3rd last	2nd last	Last
Clothing	20	282	837	1048	1695	1723	1756
Household	395	252	387	643	1335	1336	1337
Food	4	75	432	1149	1408	0	0
Technology	19	37	345	353	1933	2009	2071
Sweets	4	942	1243	1294	1358	0	0
Gifts	2287	213	216	218	2607	2608	2609
Luxury	440	142	171	184	789	790	971

Table 4: Table of samples outside of control limits

Extreme outliers:

The following figures 17-19, show the identified extreme cases of out of control xbar charts for delivery time vs class.

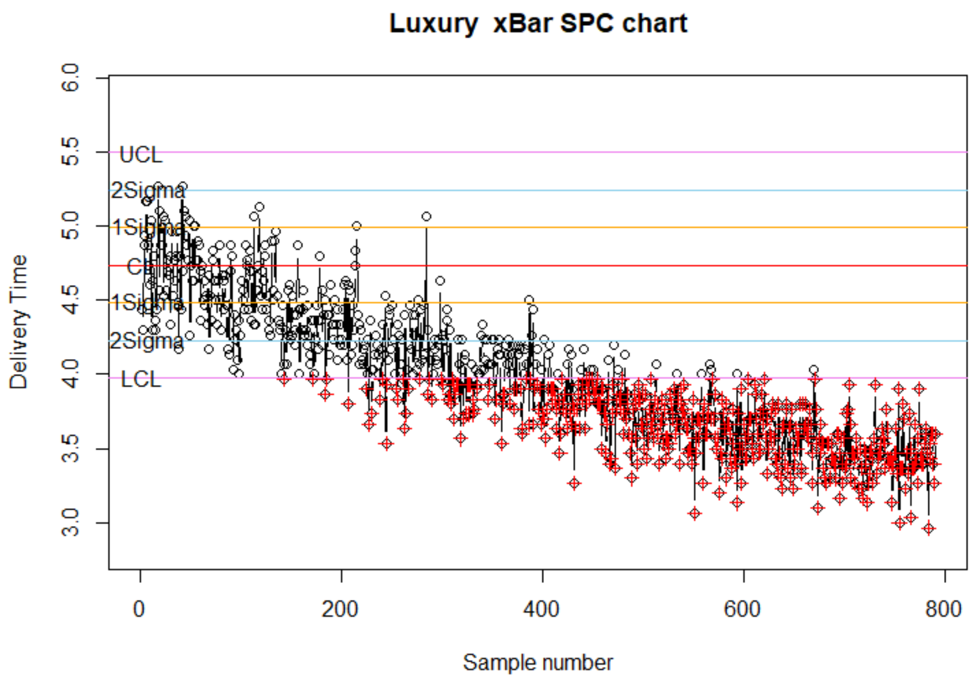

Figure 17: Luxury data outliers

The large number of consecutive outliers in figure 17 above indicate excessive variation within the process or that the mean of the process has changed. In this case, the mean has shifted down, which means the average delivery time for luxury products is decreasing. This is a good indication for the business.

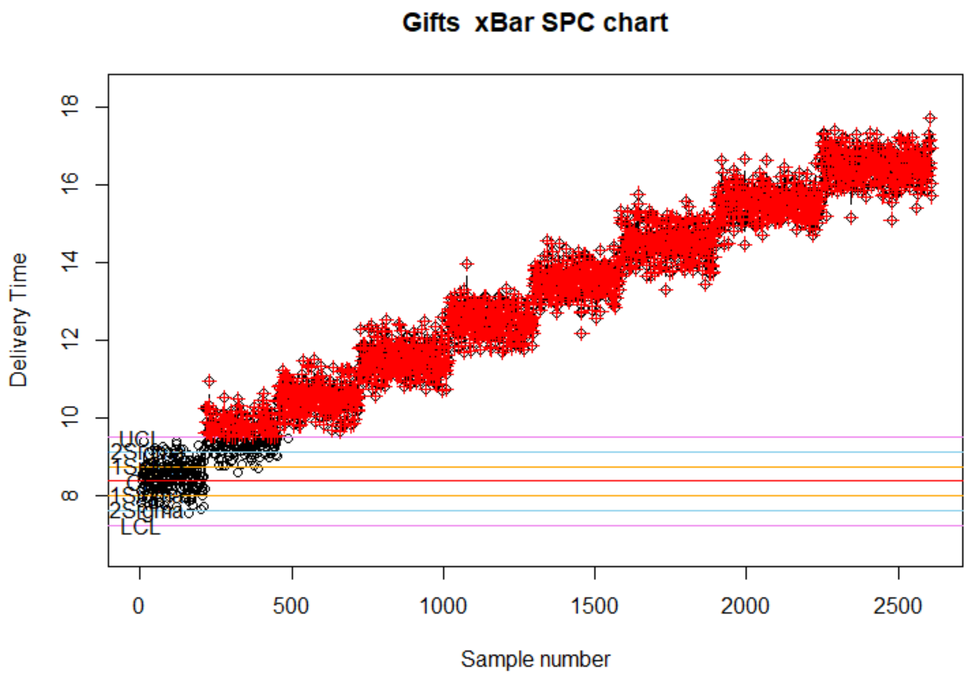

Figure 18: Gifts data outliers

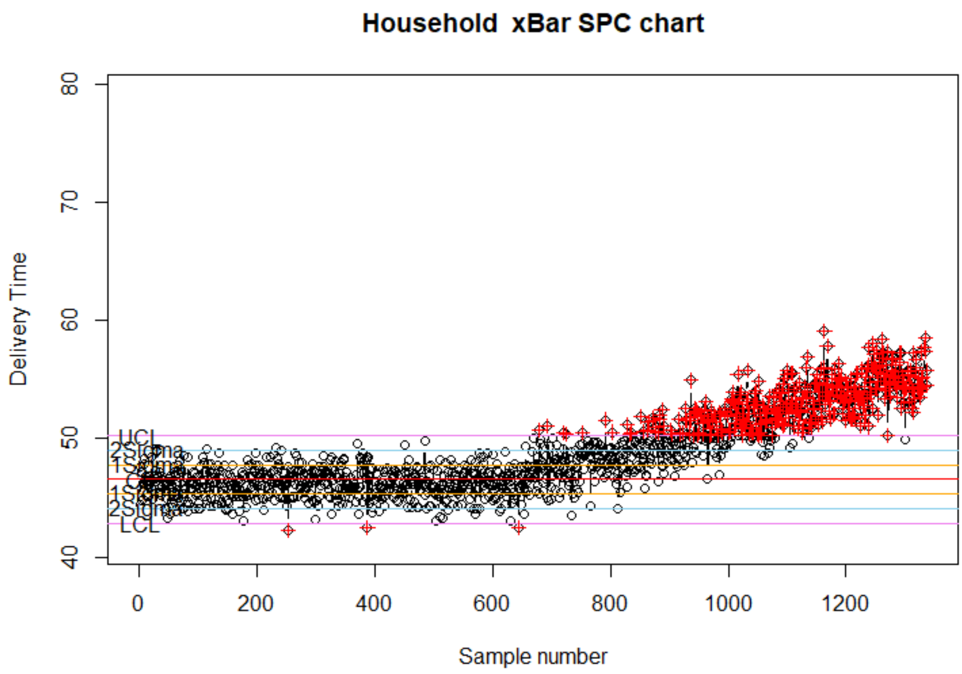

Figure 19: Household data outliers

Figures 18 and 19 indicate that mean delivery times for both household and gifts have increased dramatically. This is a bad indication, as it shows that both the processes have lost control and must be investigated by management.

4.1.2 B: Analysis of sbar charts and standard deviations for each product class

The following table is used to depict the number of consecutive samples of sample standard deviations between sigma control limit values of -0.3 and 0.4.

Classes	Index	Total/ Ending sample number
Clothing	1186	14
Household	1202	7
Food	1598	6
Technology	2384	21
Sweets	1371	10
Gifts	2476	20
Luxury	271	5

Table 5: Table of consecutive samples within -0.3 and 0.4 control limits

From analysing table 5 above, it is clear that luxury products have by far the largest number of standard deviations within the sigma control limits. Upon studying the table, one can identify that the larger the index number the larger the larger the ending sample number is.

4.2 Type 1 error for A and B

The following calculations were done to identify the probability of making a **manufacturer's** error.

	Process is under control	Process is not under control
SPC indicates process is under control	Do nothing	Type 2 error
SPC indicates process is not under control	Type 1 error	Fix process

Table 6: Error types identification

If the SPC was to indicate that a process was not under control (not within $\pm$ 3 sigma, while it actually was, a type 1 error would be made. The % of this occurring is calculate below for A and B:

$\underline{Type\ 1\ calculations\ for\ A}$

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

$= 0.002699796$

$= 0.2699796\%\ chance\ of\ making\ a\ type\ 1\ manufacturer's\ error$

$\underline{Type\ 1\ calculations\ for\ B}$

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

$= 0.2733332$

$= 27.33332\%\ chance\ of\ making\ a\ type\ 1\ manufacturer's\ error$

4.3

4.4 Type 11 error for A and B

The following graph was generated to identify the probability of making a **consumers'** error for A.

As previously stated in table 6, a type two error may occur when the SPC indicates that a process is under control, while it is actually no (this would mean that is actually not within the $\pm$ 3 sigma control limits).

As indicated in the graph, the percentage of making a type 2 error is 48.83%.

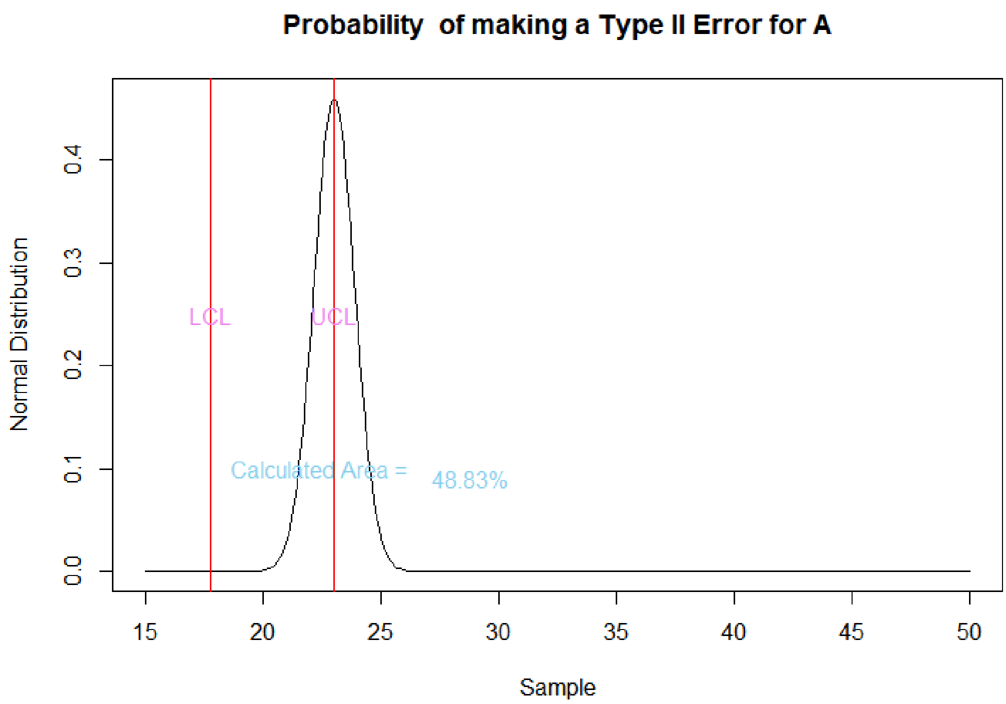

Figure 20: Probability of making a consumers' error

5 Part 5: Doe and MANOVA

5.1 MANOVA test

5.1.1 What is a MANOVA test?

A MANOVA test is used to determine whether multiple levels of independent variables have an effect on dependant variables, be it on their own or in combination with one another. (Frost, 2021).

For the given set of data set a MANOVA can created to determine the influence that the specific class of a product has on the different product prices and delivery times. The class shall be the independent variable and the price and delivery times, the dependant variables.

5.1.2 Assumptions to be made

For MANOVA to be used effectively, certain conditions must be met as the results would be biased otherwise. MANOVA assumes the following:

1. Homogeneity of variances across the range of predictors
2. Linearity between all pairs of dependent variables, covariates, and variable co-variates in each cell.
3. Normal distribution of groups of dependant variables.

5.1.3 Set up Hypothesis to be made

$H_0\ Price\colon The\ class\ of\ the\ product\ does\ not\ have\ a\ significant\ influence\ on\ the\ price\ of\ the\ product$

$H_1\ \ Price\colon The\ class\ of\ the\ product\ does\ have\ a\ significant\ influence\ on\ the\ price\ of\ the\ product$

$H_0\ Delivery\ time\colon The\ class\ of\ the\ product\ does\ not\ have\ a\ significant\ influence\ on\ the\ delivery$

$time\ of\ the\ product$

$H_1\ Delivery\ time\colon The\ class\ of\ the\ product\ does\ have\ a\ significant\ influence\ on\ the\ delivery$

$time\ of\ the\ product$

5.2 MANOVA tables for Delivery time and Price and results

5.2.1 Tables

MANOVA results for delivery time					
Value	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Class	6	33456906	5.58e+06	629515	0
Residual	179954	1594005	8.86e+00	NA	NA

Table 7: MANOVA table for Delivery time

MANOVA results for Price					
Value	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Class	6	5.72e+13	9.53e+12	80224	0
Residual	179954	2.14e+13	1.19e+08	NA	NA

Table 8: MANOVA table for Price

5.2.2 Results

As seen in both tables above, the p values are zero for both tests. These p values may have been rounded to zero as they were so small. Both of these p values indicate the there is indeed a significant influence that the class of a product has on the delivery time and price of a product, therefore both H_1 hypotheses are true.

The results of both tests can be further proven to be true by figures 1&2 from part 1, where the box plots of both the class vs delivery times and class vs price where plotted and discussed.

6 Part 6: Reliability of the services and products

6.1 Taguchi Loss Function problems

(The flowing calculations are based on problems 6 and 7 on page 363)

$\boldsymbol{Problem\ 6}$

$Specifications = 0.040 \pm 0.035\ cm$

$Cost\ to\ scrap = 30\ dollars\ per\ part$

$Calculations$:

$L(x) = k(x - T)^2$

$30 = k(0.035)^2$

$k = 24489.796$

$Taguchi\ loss\ function\ is\ therefore \rightarrow L(x) = 24489.796(x - T)^2$

$\boldsymbol{Problem\ 7}$

$Cost\ to\ scrap = 25\ dollars\ per\ part$

$Calculations$:

$L(x) = k(x - T)^2$

$25 = k(0.035)^2$

$k = 20408.163$

$Taguchi\ loss\ function\ is\ therefore \rightarrow L(x) = 20408.163(x - T)^2$

6.2 System reliability problems

(The following calculations are done based on problem 27 from chapter 7)

$\boldsymbol{Problem\ 27}$

a. The production system is in series since the backup machines are not in operation.

   $Calculations$:

   $Rs = R1 * RS * R3$

   $= 0.85 * 0.92 * 0.90$

   $= 0.7038$

   This means that the machines are 70.38% reliable when the backup machines are not in operation.

b. The production system is in parallel since the backup machines are in operation.

   $Calculations$:

   $Rp = 1 - (1 - R1) * (1 - R2) * (1 - R3)$

   $= 1 - (1 - 0.85) * (1 - 0.92) * (1 - 0.90)$

   $= 0.9988$

   This mean that the machines are 99.88% reliable when the backup machines are in operation.

6.3 Binomial probability problems

6.3.1 Information

The business has 21 drivers available whom each work 8 hours a day.

Drivers out of 21	Days drivers were available out of 1560
20	95
19	6
18	1

Table 9: Driver details

For the delivery process, there are 20 vehicles, of which 19 are required to function at any given time.

Vehicles available out of 21	Days vehicles were available out of 1560
20	190
19	22
18	3
17	1

Table 10: Vehicle details

6.3.2 How many days per year should be expected to have reliable delivery times given there are 21 vehicles?

To compute, the following code in figure was used in r to calculate the probability of all 21 vehicles being reliable:

```
#21 vehicles
#set variables
unreliable_vehc<-18
availble_vehc<-21
prob_unreliable<-(1560-22-190)/1560 #(total days-days 19 vehicles were available- days 20 vehicles were available)/total days

#Probability of being reliable
p_reliable_21v<-1-pbinom(unreliable_vehc,availble_vehc,prob_unreliable,lower.tail = TRUE,log.p = FALSE)
```

Figure 21: 21 vehicle reliability probability code

The following code in figure was used to calculate the probability of all 21 drivers being reliable:

```
#21 drivers
#set variables
new2_unreliable_vehc<-18
new2_availble_vehc<-21
new2_prob_unreliable<-(1560-95-6)/1560 #(total days- days 20 drivers were available- days 19 drivers were available)/total days

#Probability of being reliable
p_reliable_21d<-1-pbinom(new2_unreliable_vehc,new2_availble_vehc,new2_prob_unreliable,lower.tail = TRUE,log.p = FALSE)
```

Figure 22: 21 driver reliability probability code

The results were as follows:

```
###results 1
data.frame(p_reliable_21v,p_reliable_21d)
p_reliable_21v p_reliable_21d
     0.4420258      0.8484614
```

Figure 23: Reliability probability results with 21 vehicles

The days per year that one could expect a reliable delivery will then be:

$$Reliable\ days\ per\ year = p.reliable.21v * p.reliable.21d * 365$$

$$= 136.89$$

$$= 136\ days$$

6.3.3 How many days per year should be expected to have reliable delivery times given there are 22 vehicles?

To compute, the following code in figure was used in r to calculate the probability of all 22 vehicles being reliable:

```
#22 vehicles
#set variables
new_unreliable_vehc<-18
new_availble_vehc<-22
new_prob_unreliable<-(1560-22-190-3)/1560 #(total days- days 19 vehicles were available- days 20 vehicles were available- days 18 vehicles were available)/total days

#Probability of being reliable
p_reliable_22v<-1-pbinom(new_unreliable_vehc,new_availble_vehc,new_prob_unreliable,lower.tail = TRUE,log.p = FALSE)
```

Figure 24: 22 vehicle reliability probability code

The results were as follows:

```
###results 2
data.frame(p_reliable_22v,p_reliable_21d)
p_reliable_22v p_reliable_21d
     0.6399049      0.8484614
```

Figure 25: Reliability probability results with 22 vehicles

The days per year that one could expect a reliable delivery will then be:

$$Reliable\ days\ per\ year = p.reliable.22v * p.reliable.21d * 365$$

$$= 198.17$$

$$= 198\ days$$

7 Conclusion

After using various methods to statistically analyse the data provided, it is recommended that the business investigate the delivery processes of gifts, household and luxury products.

8 References

- Capability index - explained (no date) The Business Professor, LLC. Available at: https://thebusinessprofessor.com/en_US/mgmt-operations/capability-index-definition (Accessed: October 19, 2022).

- Frost, J. (2021) Multivariate anova (MANOVA) benefits and when to use it, Statistics By Jim. Available at: https://statisticsbyjim.com/anova/multivariate-anova-manova-benefits-use/ (Accessed: October 20, 2022).

- Hessing, T. (2021) Process capability (CP & CPK), Six Sigma Study Guide. Available at: https://sixsigmastudyguide.com/process-capability-cp-cpk/ (Accessed: October 19, 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 19, 2022).