Quality Assurance 344:

ECSA Graduate Attributes Project

By: Hlonela Mayosi

22943366

OCTOBER 22 2022

Table of Contents

Introduction

For any business, the is ongoing thought and effort put into how they can perform better and continue to satisfy their customers. This brings quality into the discussion. Quality is something that this online business would like to uphold, and where it is not present, would like to implement as soon as possible. Consistent quality is important to ensure future sales, and to ensure the steady growth of any enterprise. For this to occur, however, we have to investigate what is currently occurring in the business.

In this report, we aim to analyse different aspects of quality. We have been tasked with analysing the processes of an online business to find patterns and gain insight to finally offer recommendations for them to improve their processes, output, and profit.

We are given sales data to be cleaned and analysed. Once the valid data has been extracted, graphs and figures will be produced to gain insight into the sales data and different features within the dataset. This is done through statistical analysis. Once we have performed statistical analysis, we perform process capability analysis using the process capability indices. This gives further insight into how the business is performing.

Statistical process control is process is performed to see how much data is within the specified limits; Type 1 and Type 2 errors are investigated; processes are optimised and we also investigate the reliability. All of this, is performed to gain further insight into how the business can further improve its value offering

1.Data Wrangling

Upon receiving raw data, it is imperative that we perform data cleaning and removal as this greatly affects the results extrapolated by the data. In the sales data we have been given, we need to ensure that negative values (for price for example) and missing values, i.e invalid data, must be extracted and isolated. This has been performed. The valid data has been isolated into its own variable, and invalid data (all data that is not valid) has been isolated into its own variable. Before further explanation, it is beneficial for one to become acquainted with the applicable features:

- **X**: Sales ID, originally in chronological order
- **ID**: A unique value assigned to a client to identify a sale
- **Age**: A feature indicating the age of each client
- **Class**: This is a categorical feature indicating what type of product was purchased. Options are between clothing, food, gifts, household, luxury, sweets and technology
- **Price**: Feature indicating the total amount the customer paid on that sale
- **Year**: Indicates the year the purchase was made
- **Month**: Indicates the month the purchase was made
- **Day**: Indicates the day the purchase was made
- **Delivery Time**: A feature indicating how many hours it took for a delivery to reach a client upon purchase
- **Why Bought**: A feature describing how the client was made aware of the website. Options include Recommended and Website

Again, adequate data cleaning is crucial because without it, we will not be able to draw conclusions or insights from trends and patterns we discover. For context, a screenshot of the data and features is attached:

Filter

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

Figure 1: Snapshot of Dataset and features

2. Descriptive Statistics

It is important that the sales dataset is analysed and that graphs are drawn to extrapolate conclusions. Descriptive statistics allow us to evaluate the data and gives us greater insights into the data and the classes.

2.1Categorical data

2.1.1 Revenue per Class

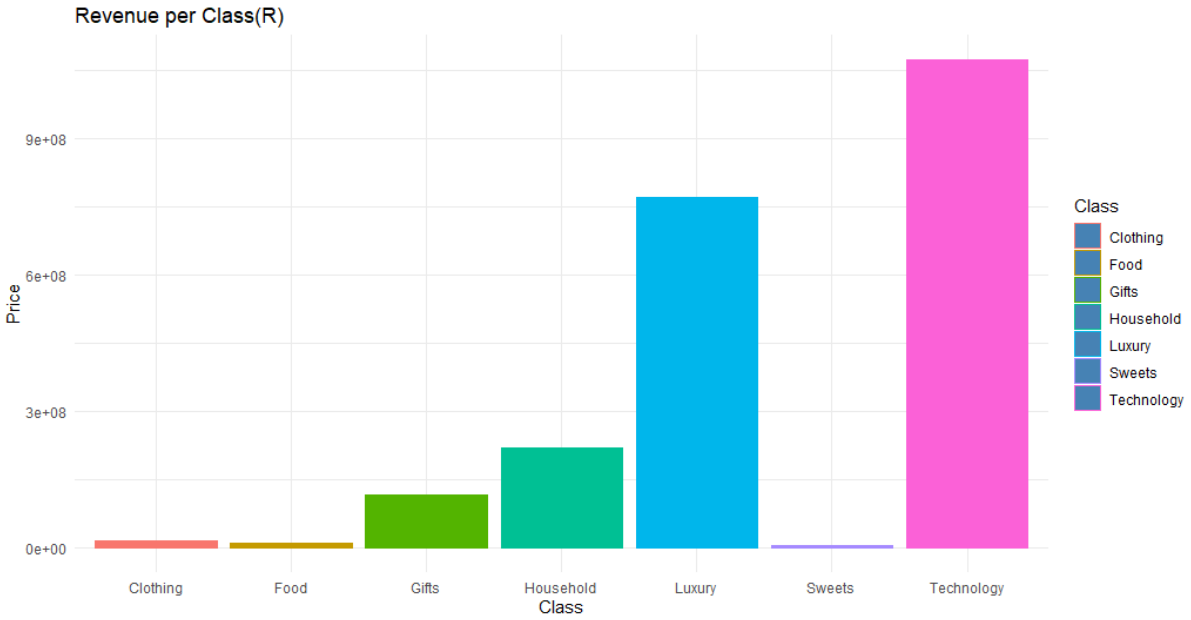


Figure 2: Graph showing Revenue per class

A practical question that business owners would like to answer is Which items bring in the most revenue? This is question is answered in the above graph. It is clear In the above graph that the top two items that brought forth the most revenue was Technology followed by Luxury, with respective revenues of R1072529552 and 769789795 for each item/ class.

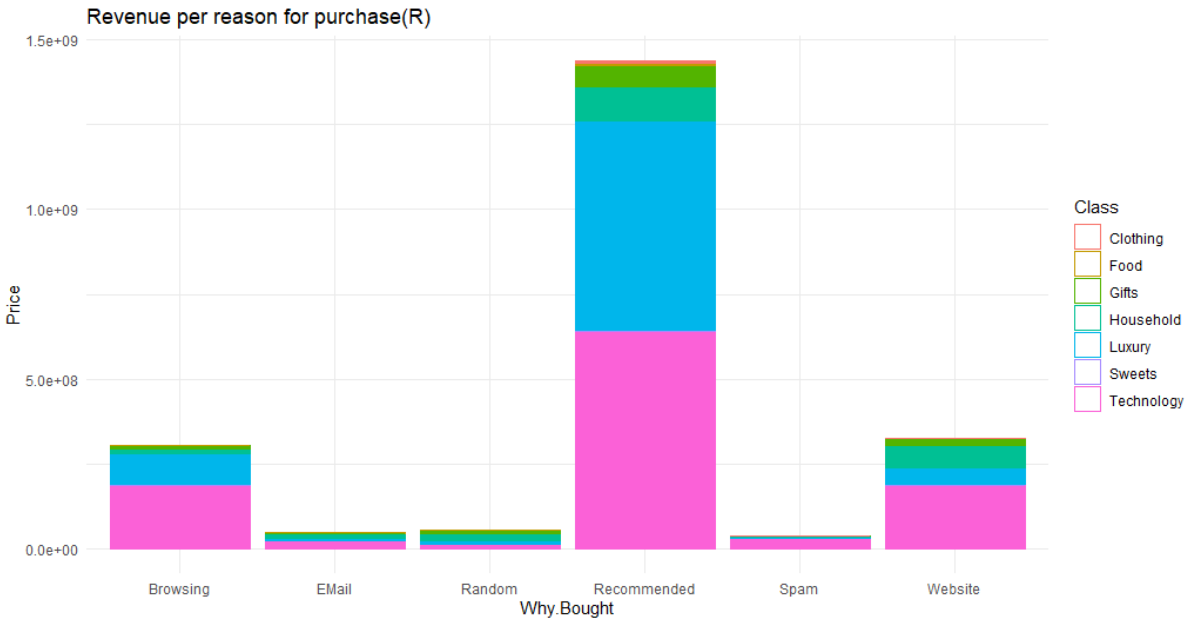
A natural question to follow, is the items sold. This allows us to gauge the class or items with the largest revenue power. This is because some items are more expensive than others, thus a high revenue would not indicate a high demand for the product. Another important aspect is to find out the revenue per "reason" to be bought. The revenues for each class are shown below:

Class	**Clothing**	**Food**	**Gifts**	**Household**	**Luxury**	**Sweets**	**Technology**	
Revenue(R)	16911790	10024915	115953129.71	220901078	769789795	6556973	1072529552	

Table 1: Revenue received per class

2.1.2 Revenue per reason

Figure 3: Graph showing reasons for different sales



This graph shows us the revenue associated with "reasons" for which a customer purchased a product. This graph also shows us which classes or items. From this, we can see that most of our items were purchased due to Recommendations. This means that the business must focus on the quality the products and deliveries. This is because if there is a defect, there will be less recommendations and thus a big dip in sales(since they bring most sales)- especially technology and luxury.

Reason for purchase	**Browsing**	**Email**	**Random**	**Recommended**	**Spam**	**Website**	
Revenue(R)	306383885	48126832.8001	56266859.59	1437978192.83	39390669	324520796	

Table 2: Revenue accumulated per reason for sale

2.1.3 Frequency of items per class

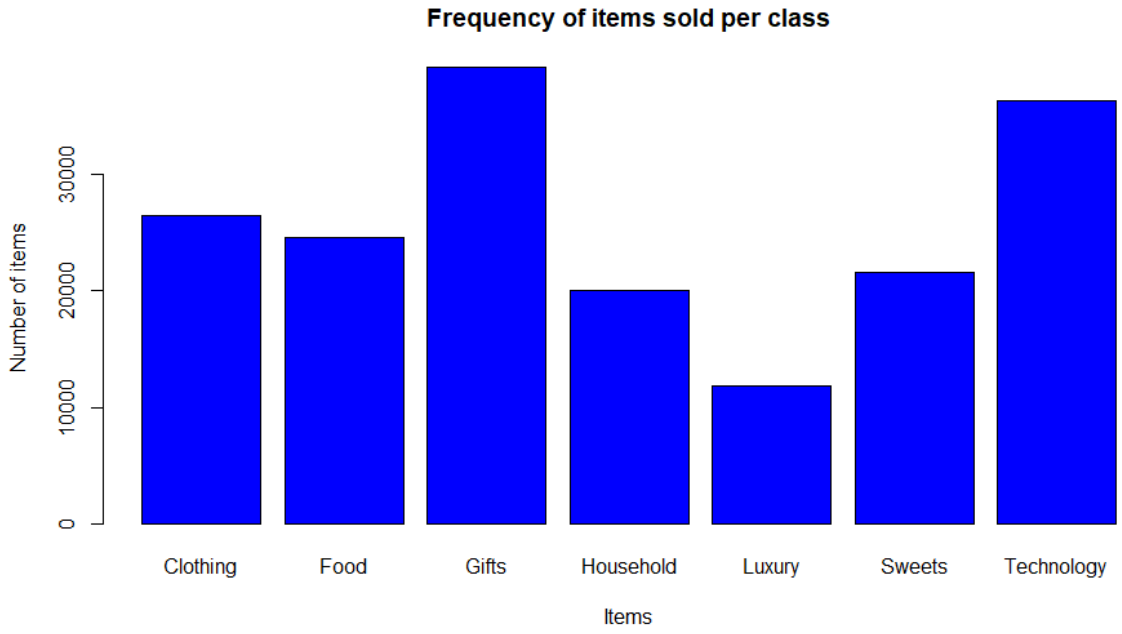


Figure 4: Frequency of items sold per class

This is an important figure, as it goes hand in hand with the total revenue figures. This figure allows us insight into the revenue contribution of the classes. Gifts and Technology are the leading items that have been purchased.

Technology is thus a key player in the business, since it brings forth the most revenue and has also has the second-most items purchased. Gifts may be the most popular items to purchase, but this does not translate into the revenue they bring in. Measures may thus be implemented to take more advantage of this popularity.

Class	Clothing	Food	Gifts	Household	Luxury	Sweets	Technology
Number of items bought	26403	24582	39149	20065	11868	21564	36347

Table 3: Number of items bought per class

2.2 Continuous data

2.2.1 Customer Age statistics

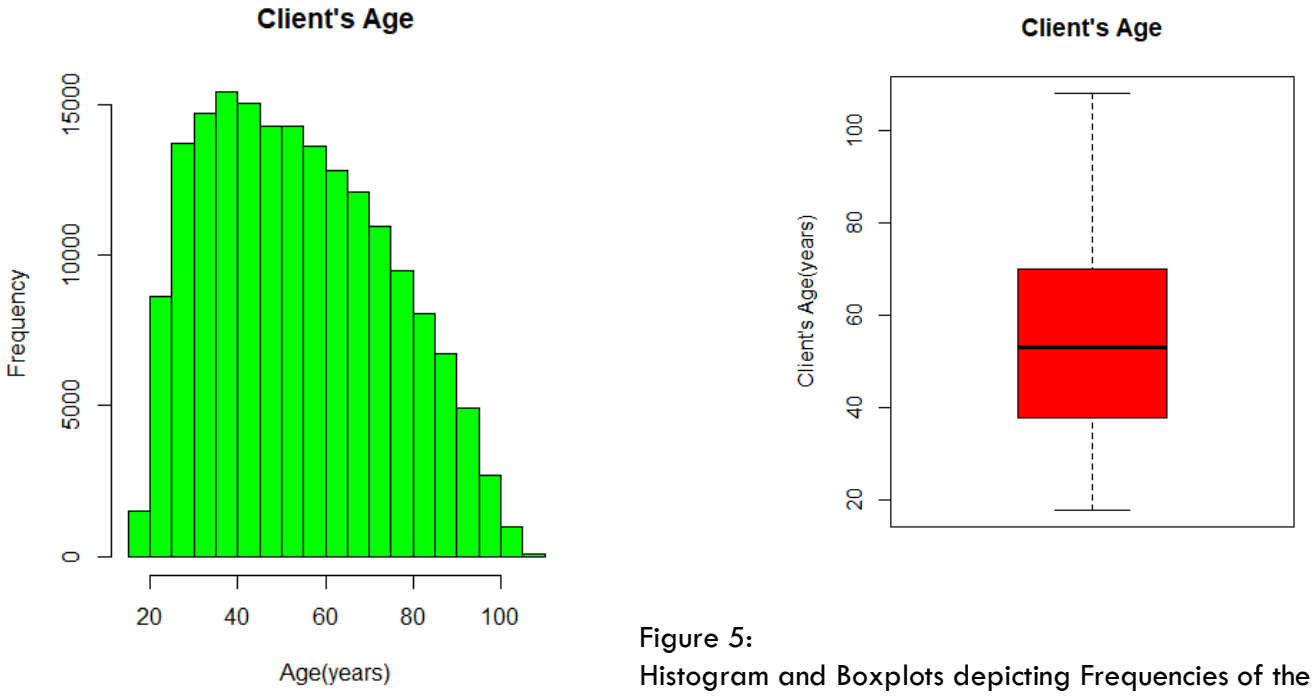


Figure 5:
Histogram and Boxplots depicting Frequencies of the ages of the customers

This Client age histogram tells us that most of the sales occur from customers who are younger, rather than older. We can see this from the plot being skewed to the right, with the peak being on the left-hand side with an age close to 40 years old. From an age close to 60-year-olds, there is a sharp decline in sales. The standard deviation shows that the ages are spread over a large group

Most sales occur between 20 and 45 years old. This is important because it guides the marketing strategy of the company. This could pertain to family specials for the middle-aged, special for newly graduates etc.

	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Std.dev
Age	18	38	53	54.57	70	108	20.38881

Table 4: The Statistic summary for client's age

2.2.2 Purchase Delivery time Statistics

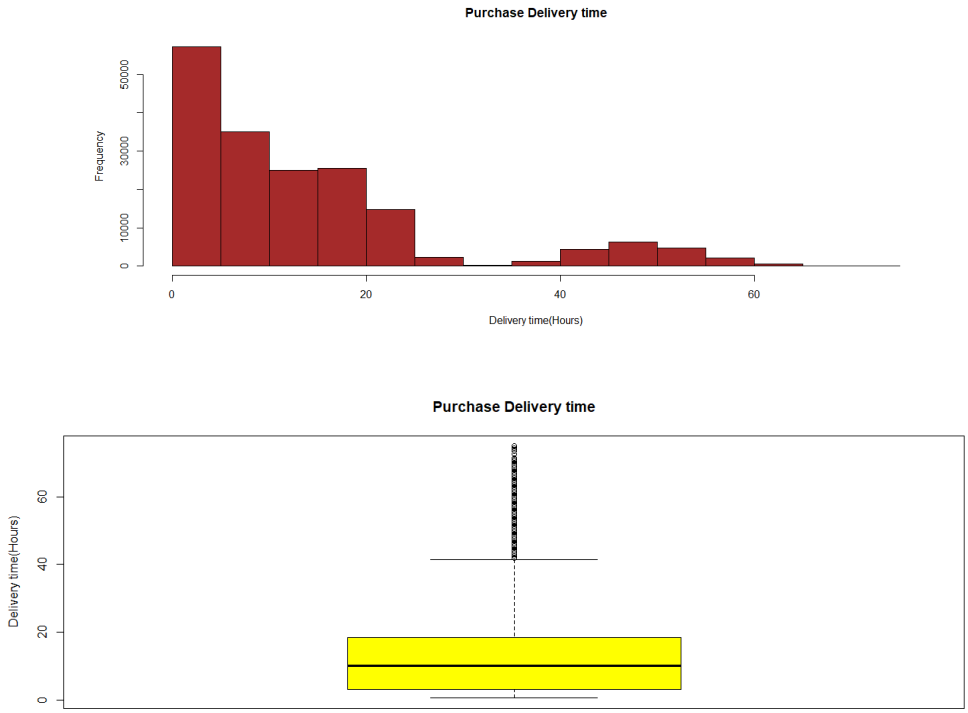


Figure 6: Histogram and Boxplots for Delivery Times

This histogram represents the general representation of the delivery times of the orders purchased. This is a right-skewed distribution with a median closer to the lower quartile. Since there are outliers, these compromise the mean value of the plot. There is a wide range of delivery times, translating to a large variety of types of products sold.

Most of the sales are between 0 and 20 hours. This is encouraging as most of the orders are produced in a shorter timespan than a longer timespan

Table 5: The Statistic summary for Delivery Times

	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Std.dev
Time(hours)	0.5	3	10	14.5	18.5	75	13.95578

2.2.2 Age per class

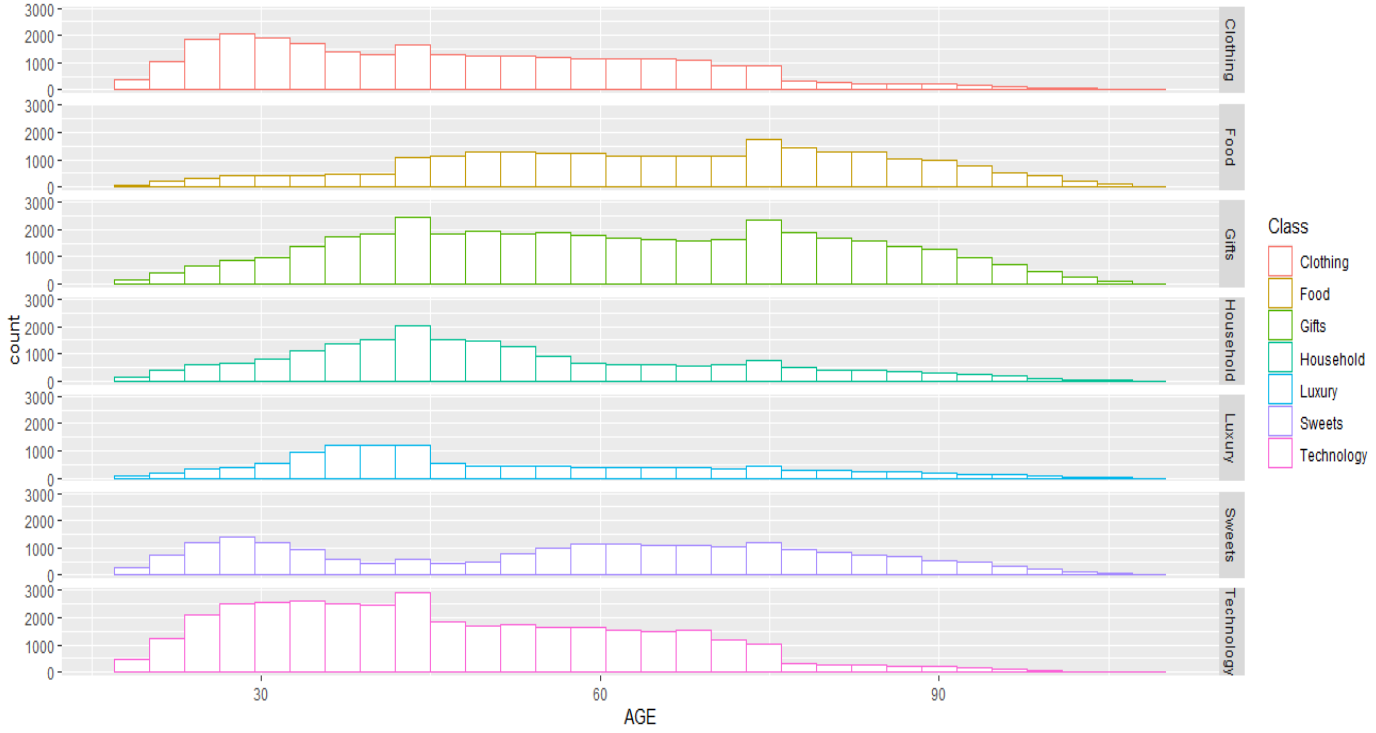


Figure 7: Histograms showing the age distributions per class

This plot is an important one, because it shows the break down of the age of customers per class or item they purchased. Present are both right-skewed and left-skewed distributions. Right-skewed distributions include Clothing, Households, Luxury and Technology. The left-skewed distributions are the Food class. From these figures we can conclude that younger clients purchase the luxury goods, clothing, and sweets. Gifts, however, are more balanced. Older clients however, purchase from Food mostly.

2.2.3 Purchase Delivery time per class

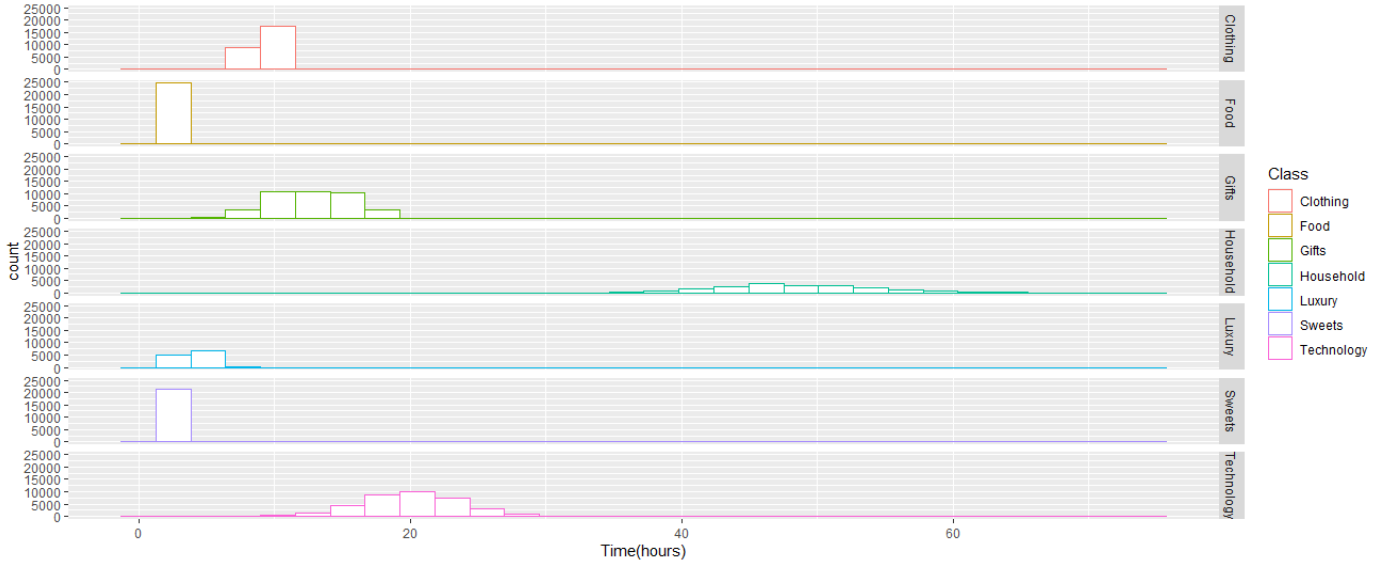


Figure 8: Histograms showing the Delivery Times distributions per class

The above graph gives us the opportunity to gauge the times each class takes to reach customers. This is useful because from here the business can decide where they would like to improve and shorten delivery times for customer satisfaction. From the above graph, we can observe that the items that take the longer to be delivered are Household and Technology. Other goods take a shorter duration such as Food and Sweets. These are perishable and thus must be delivered timely.

2.3 Process Capability Indices for Delivery Time.

Process Capability is a measure of the process variability. It Is thus the study of investigating the ability of a process to meet given specifications. For the company, the process capability assessment is performed with respect the to process delivery times of the technology class items

Process Capability indices are calculated and used to perform the analysis. These being: Capability Potential (Cp), Capability Potential based on Upper specification limit(Cpu), Capability Potential based on Lower specification limit(Cpl) and the actual capability performance(Cpk).

The Upper specification limit(USL) is given to be 24 hours and the Lower specification limit(LSL) is given to be 0 hours. The LSL is logical because the clients desire the delivery to be done as quickly as possible and also we cannot have negative hours. The mean value of Technology class Process times(μ) and the standard deviation (σ) are calculated to be 20.01095 hours and 3.501993 hours.

Thus:

USL=48 hours; **LSL**=0 hours; **μ**=20.01095; **σ**= 3.501993

2.3.1 Capacity potential (Cp)

$$Cp = \frac{USL - LSL}{6\sigma} = 1.142207$$

2.3.2 Capability Potential based on Upper specification limit (Cpu)

$$Cpu = \frac{USL - \mu}{3\sigma} = 0.3796933$$

2.3.3 Capability Potential based on Upper specification limit (Cpl)

$$Cpl = \frac{\mu - LSL}{3\sigma} = 1.90472$$

2.3.4 Actual capability performance(**Cpk**).

$$Cpk = \min(Cpl, Cpu) = 0.3796933$$

Since the Cpk value is less than 1, the process analysed is considered to be poor and incapable of meeting specifications or requirements. Thus the company should seriously consider improve this by reducing the variation and implementing other plans to improve the process delivery times

3. Statistical process control (SPC)

3.1 S and X Charts for each class

Statistical process control is an important process that makes use of procedures to monitor the behaviour of processes within a system. Process control is represented by standard deviation(S) and mean(X-Bar) control charts. These charts were executed on thirty samples of ordered data contained in every class category (technology, clothing, gifts, food, sweets, household, luxury). The data is ordered according to year, month and day. After this, the rest of the remaining data is sampled to analyse the correlation between the delivery times of first thirty samples and the rest of the samples within each respective class.

These charts produce a centre line, upper control limit and lower control limit. These allow for the data points that lie outside the upper and lower control limits to easily be identified. These are referred to as out-of-control data points. X-bar charts are used to gauge stability of delivery times through the years, and the S chart is used to gauge the standard deviations of these stabilities or instabilities.

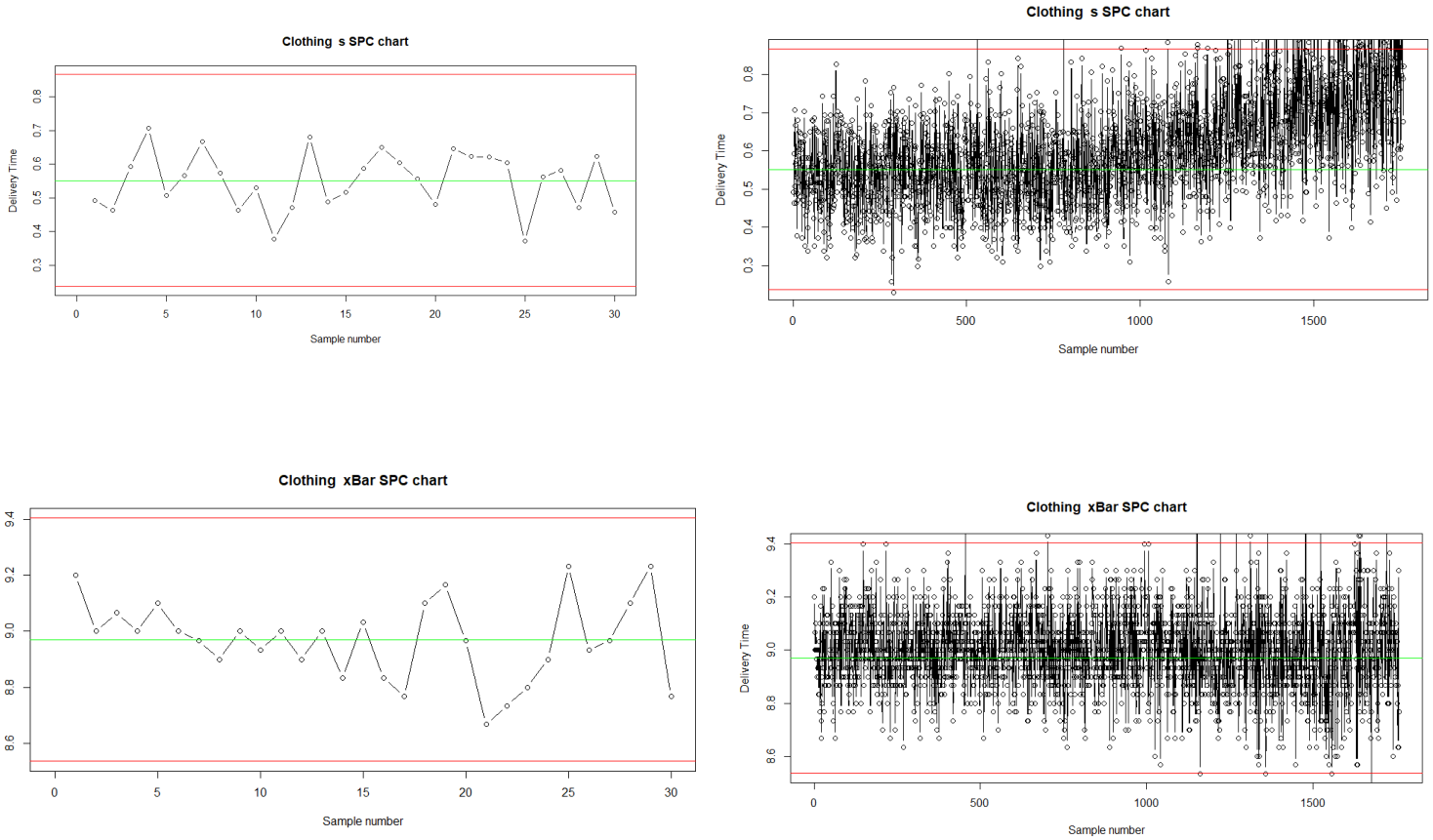


Figure 9: Xbar and s charts for Clothing: For 30 samples and all remaining samples

In these charts, we see that based on the samples taken, the clothing delivery times are within control. However, we see that when we take a larger sample, that the standard deviation begins to be out of control. This translates to a lack of stability and thus means that the mean values documented will become inconsistent and inconclusive(due to high standard deviation). Most of the X-bar values are in control and the few out of control samples may be due to increased demand or lack of capacity

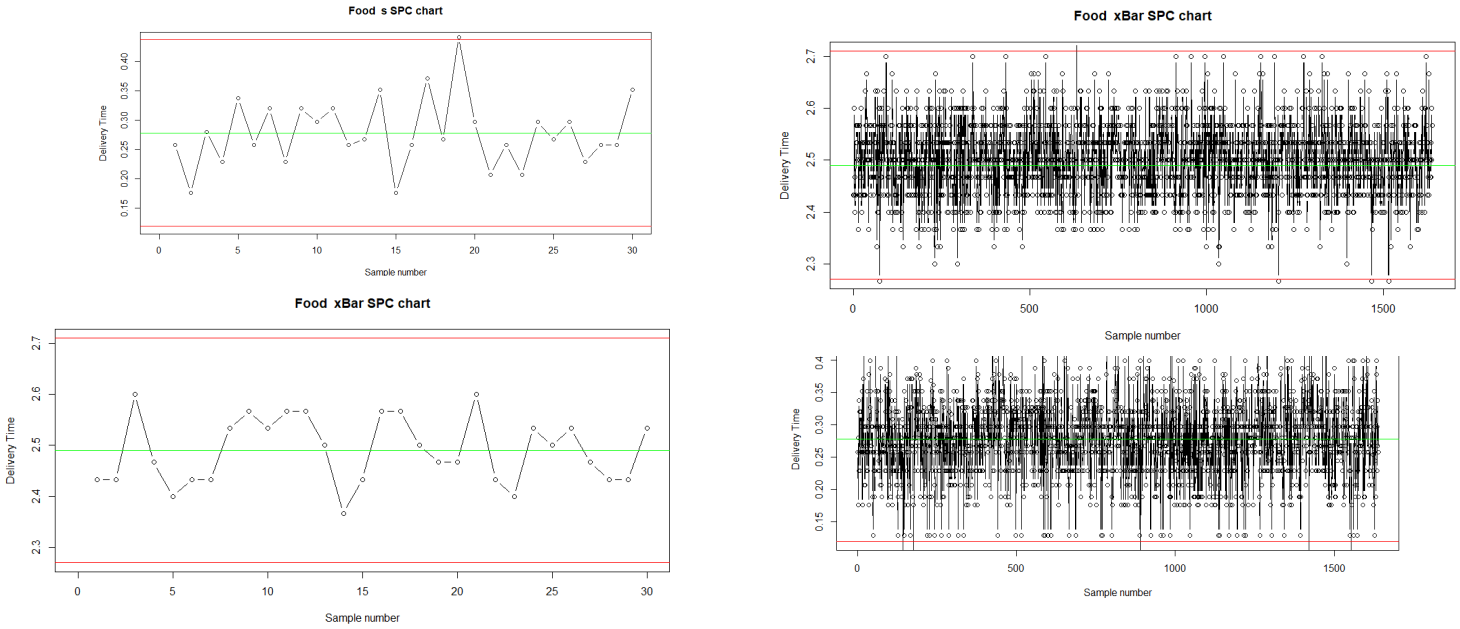
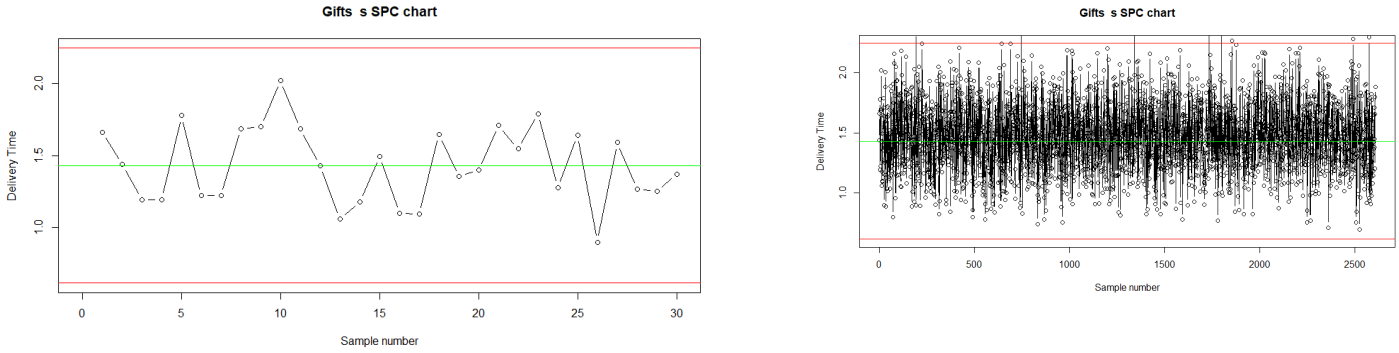
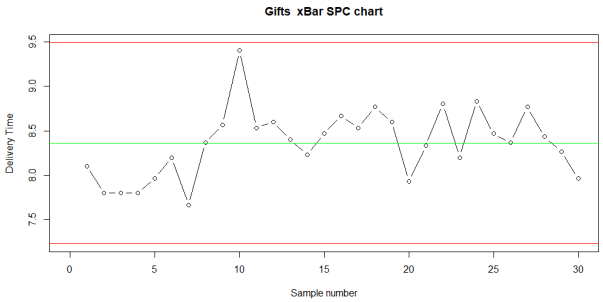


Figure 10: Xbar and s charts for Food: For 30 samples and all remaining samples

The food delivery times are under control for the thirty samples and mostly under control for the larger samples as well- besides a few anomalies. This is encouraging and should be kept this way





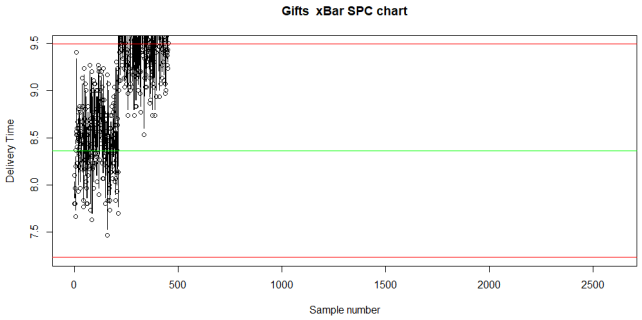
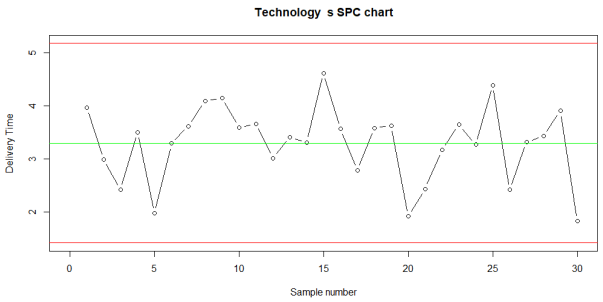
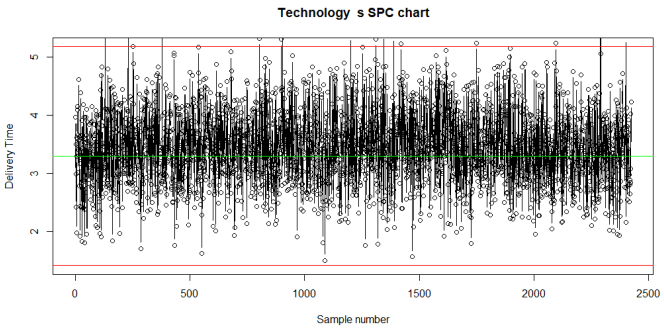
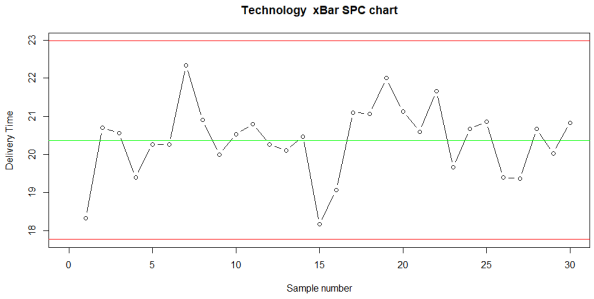


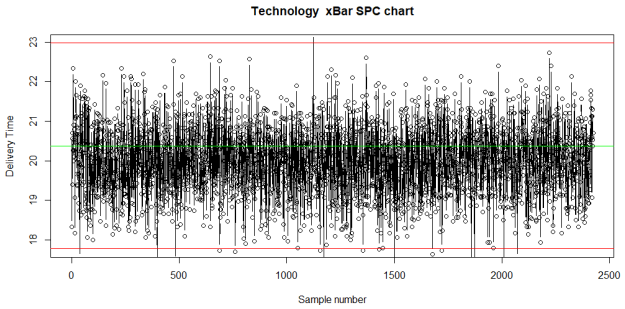
Figure 11: Xbar and s charts for Gifts: For 30 samples and all remaining samples

For gifts, the first thirty samples are under control and are within the upper and lower limit. When we take further samples for this class we find that the standard deviations of the means of the samples are relatively under control but the average delivery times of the samples are dangerously out of control. This is a concern for the business and must be attended to immediately. This may be due to excess demand and not enough capacity, or other technical issues









Technology is a class that is under control relatively both in standard deviation and in the mean value of samples. This stability means that our mean values are conclusive, and we have drawn conclusions from them

Figure 12: Xbar and s charts for Technology: For 30 samples and all remaining samples

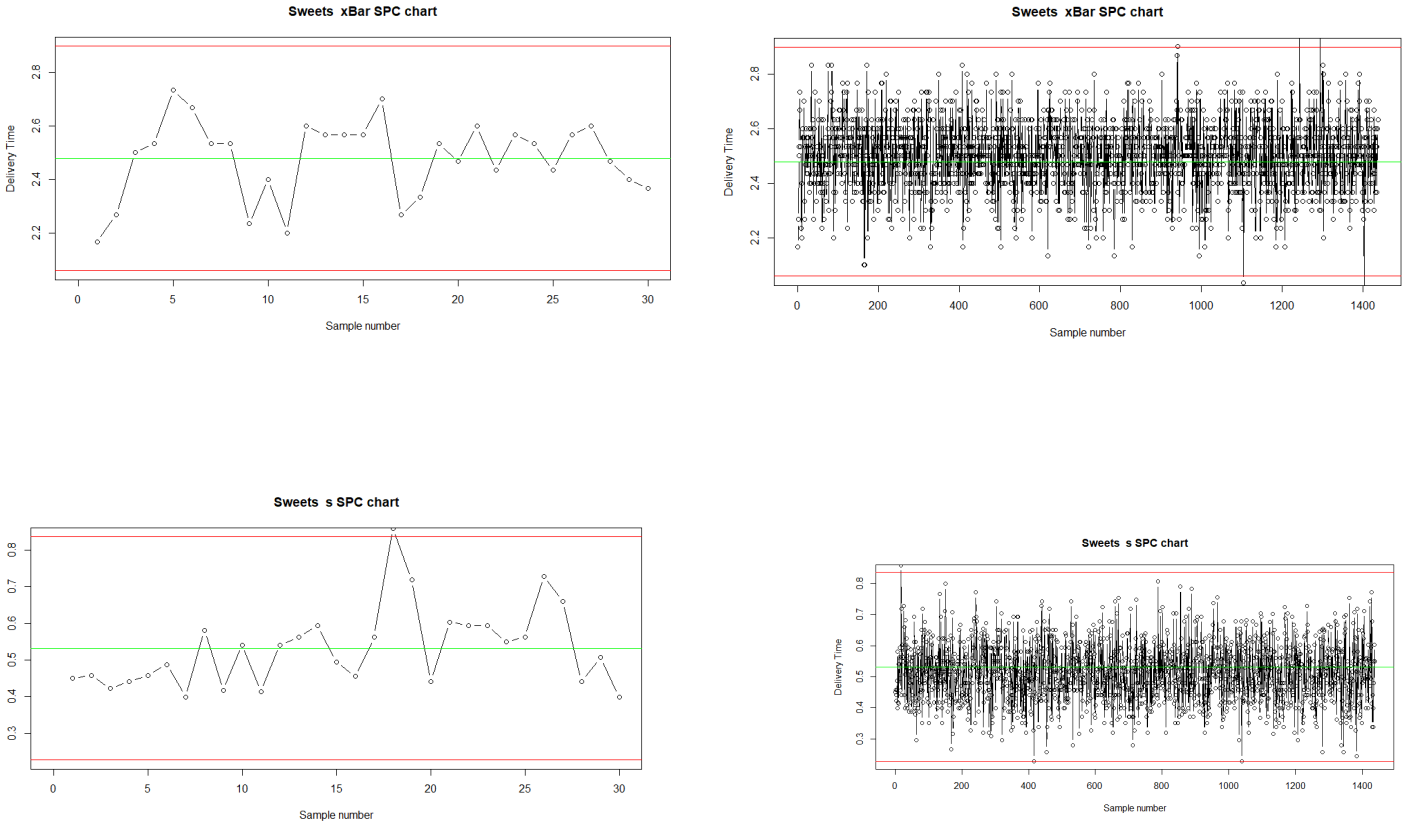
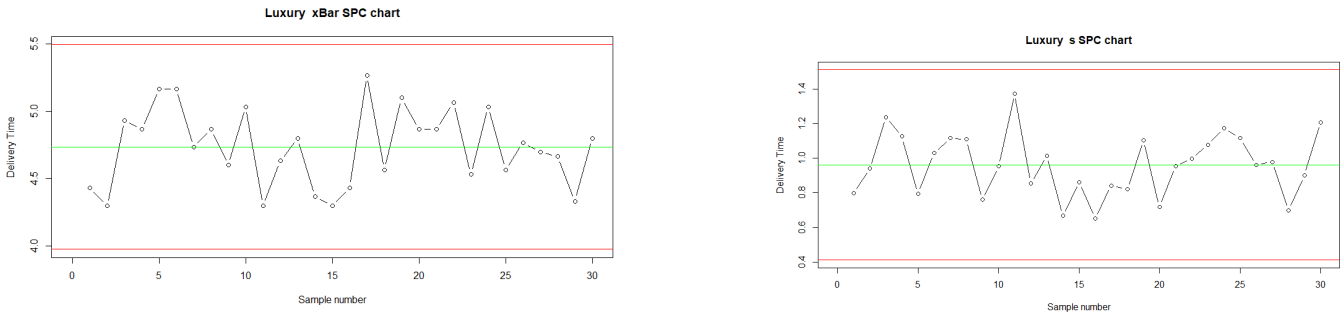
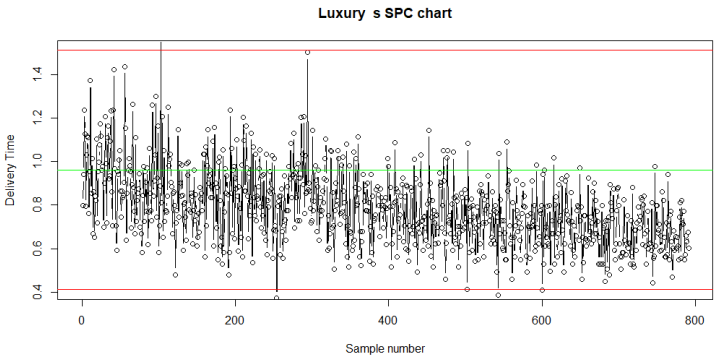


Figure 13: Xbar and s charts for Sweets: For 30 samples and all remaining samples

Sweets is another category that is in control. A vast majority of samples are in control from both the thirty samples and the rest of the samples in the category. The few anomalies may be investigated for further clarification





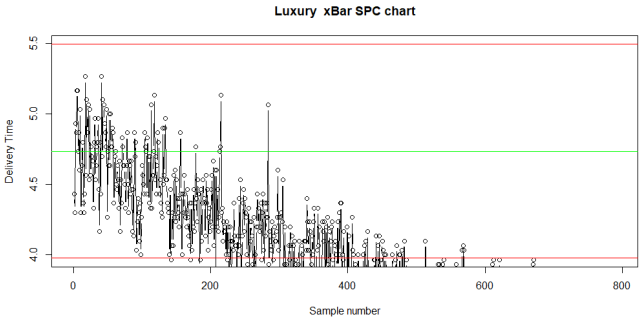
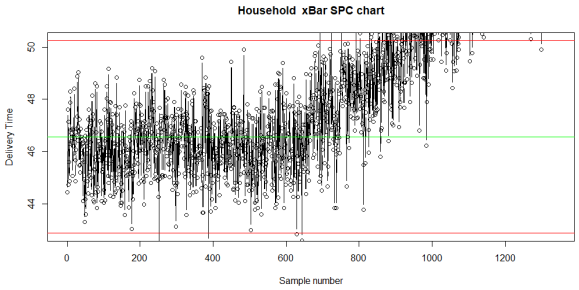
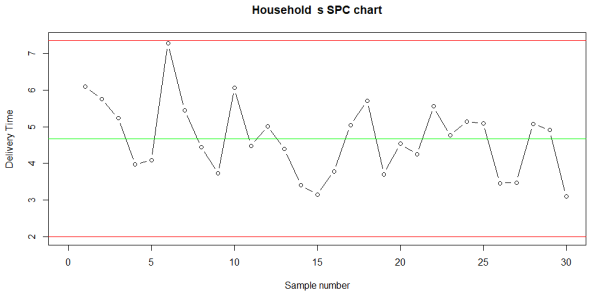
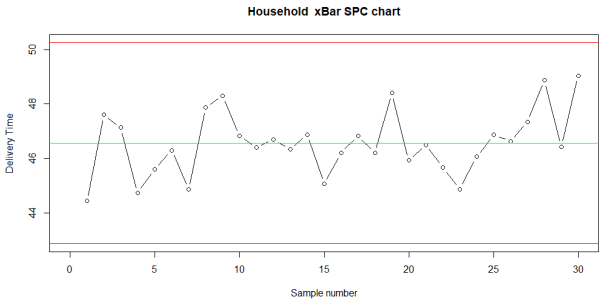


Figure 14: Xbar and s charts for Luxury: For 30 samples and all remaining samples

The concern for the luxury delivery times is the sharply decreasing average delivery times. These are out of control and continue to decrease into being out of control. One perspective may be that the items are being delivered faster, but this may be too fast to allow recovery or replenishment, hence being out of control. This is something that must be investigated and fixed urgently.







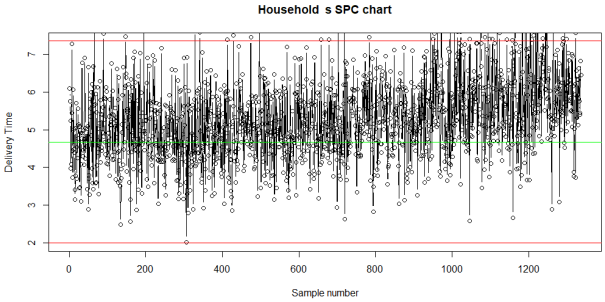


Figure 15: Xbar and s charts for Household: For 30 samples and all remaining samples

The household items is another class with an out of control mean(X-Bar) delivery time that neds to be urgently controlled and stabilised. Although the thirty samples can be encouraging, the rest of the remaining samples show a better representation. The standard deviations between sample 1000 and sample 1200 also show loss of control

Below are the summary tables depicting the centre lines, upper limit, lower limit,U1Sigma,U2Sigma, L1Sigma and L2Sigma for each class:

For the X-Chart

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL	
Technology	22.97462	22.1079	21.24117	20.37444	19.50771	18.64098	17.77427	
Clothing	9.404934	9.259956	9.114978	8.97	8.825022	8.680044	8.535066	
Household	50.24833	49.01962333	47.79092167	46.56222	45.33351	44.10481667	42.87612	
Luxury	5.493965	5.25116	4.98835	4.735556	4.48275	4.22994	3.977146	
Food	2.709458	2.6363	2.56315	2.49	2.4168	2.3436	2.270542	
Gifts	9.488565	9.11274	8.7369	8.361111	7.985293	7.60947	7.233658	
Sweets	2.897042	2.7572	2.617534	2.477778	2.3380	2.1982	2.058514	

Table 6: Control chart values for the X-chart

For the S-Chart

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL	
Technology	5.18057	4.552222	3.923875	3.295528	2.66718	2.038833	1.410486	
Clothing	0.8665596	0.761453 5	0.656350	0.551246 5	0.44614 2	0.34103	0.235933 5	
Household	7.34418	6.45341	5.56264	4.67187	3.7811	2.89033	1.99956	

Luxury	1.511052	1.3277	1.144503	0.9612289	0.77795	0.59468	0.411406	
Food	0.4372466	0.384213	0.33118	0.2781467	0.225113	0.172080	0.1190468	
Gifts	2.246333	1.97387	1.70142	1.428965	1.156509	0.884053	0.6115971	
Sweets	0.8353391	0.73402	0.632703	0.5313862	0.43006	0.3287509	0.2274333	

Table 7: Control chart values for the S-chart

4. Optimising the delivery processes

4.1 Analysis of the sample means and standard deviations out of control

	Clothing	Household	Food	Technology	Sweets	Gifts	Luxury
No. of samples outside control limits	17	400	5	17	5	2290	434
First 3 Samples	455 702 1152	252 387 629	75 633 1203	37 398 483	942 1104 1243	213 216 218	142 171 184
Last 3 Samples	1677 1723 1724	1335 1336 1337	1203 1467 1515	1872 2009 2071	1243 1294 1403	2607 2608 2609	789 790 791
Most consecutive number of samples of sample standard deviations between the -0.3and +0.4 sigma control limit	4	3	7	6	4	5	4

4.2 Type 1 error analysis

A type 1 error, otherwise known as a Manufacturer's error, is a well-known term in quality engineering. A type 1 error occurs when we reject the null hypothesis(H0) when it is true. In other words, it is a false positive conclusion such as being told something is wrong, given it is correct.

In this specific case, it is the rejecting the null hypothesis that the process is under control. When in fact it is under control. We investigate the likelihood of making a type 1 error when identifying x-bar(mean) samples outside the control limits(A) and when identifying the highest number of consecutive standard deviation values with -0.3 and +0.4 of the control limits.

For Rule A: The probability of making a Type 1 Error occurs when we identify samples outside the outer control limits. These are set at LCL=-3 and UCL=+3. The centreline is at 0.

P(Z>3) = 1-P(Z<3)= 1-0.9986501= 0.001349898. Since the distribution is two-sided(symmetrical) we multiply this by two to obtain:

P(Z>3)*2= 0.001349898*2=2.699796e-3 =0.26997%=0.27%

This is a low chance that the company may be willing to take. If this risk proves false however, financial costs will be incurred.

For Rule B: The probability of making a Type 1 Error occurs when we identify samples between -0.3 sigma and +0.4Sigma. Thus:

P(-0.3<Z<0.4)= P(Z<0.4)-P(Z<-0.3)= 0.6554217- 0.3820886

=0.2733332= 27%.

We have 27% of a type one error occurring for standard deviation samples between -0.3 sigma and +0.4Sigma.

4.3 Optimizing the Delivery Process for Maximum Profit.

The business is wanting to find the optimal hours to reduce the delivery times by, in order to maximise the profit incurred. The current mean delivery time that we are tasked to optimize is 20.01hours the business currently has a mean of If the delivery of a technology item is slower than 26 hours, we will lose R326 per item-late-hour in lost sales. Another cost Incurred is one of R2.5 per time/hour to reduce the average time for delivery time.

After iterations, we can determine the minimum loss cost to be R340870, corresponding with a 3 hour reduction(seen on the figure below figure 16). This is indicative to the business that the delivery times for the technology class should be reduced by 3 hours to 17.01095 hours

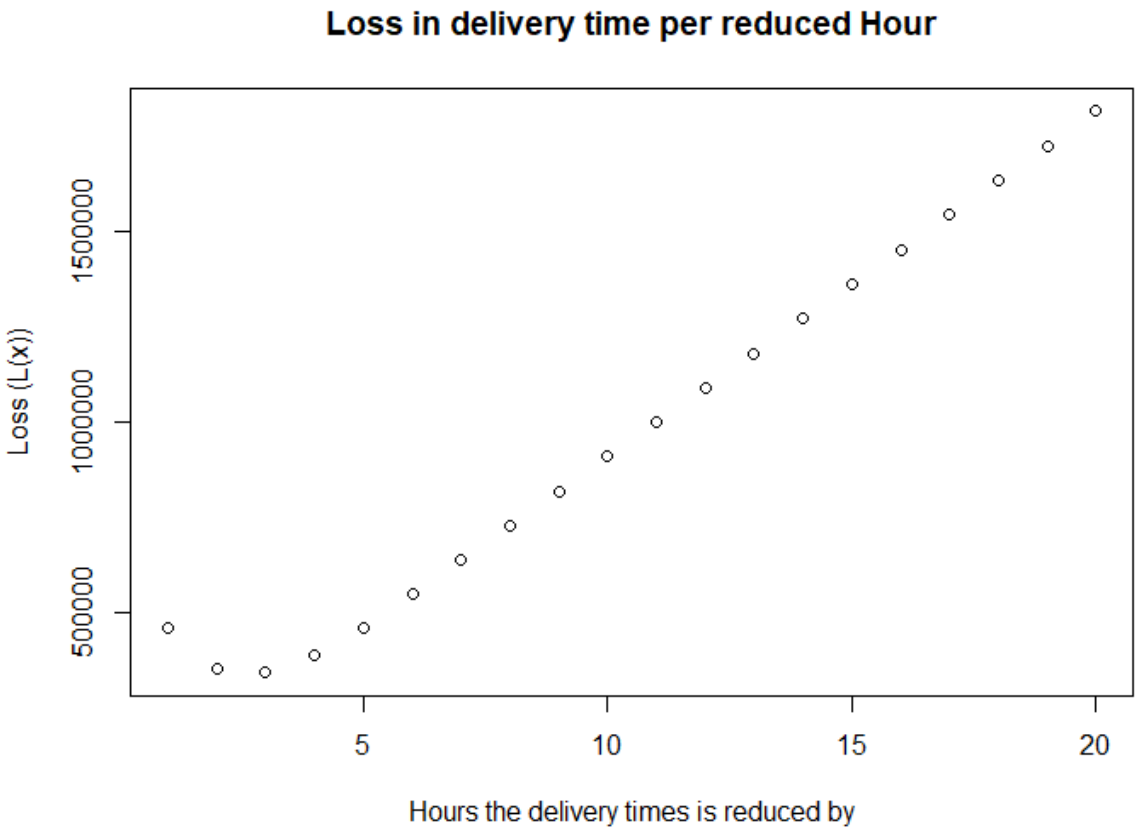


Figure 16: Curve for the Loss in delivery time per reduced hour

4.4 Type 2 error analysis

A type 2 error, otherwise known as a Consumer's error, is also a well-known term in quality engineering. A type 2 error occurs when we accept a false hypothesis. In other words, it is the probability of telling you things are correct, given that they are wrong.; not rejecting the null hypothesis when it is false.

The Null and alternative Hypothesis are the still the same as in Section 4.2. We calculate the probability of making a type 2 error for part A:

Standard Deviation = (UCL-LCL)/6

= (22.97462- 17.77427)/6=0.866725

Likelihood of type 2 error =P(Z<Z2)-P(Z<Z1)

With Z1=(LCL-23)/ (Standard deviation)=(17.77427-23)/(Standard Dev)= -6.029283

And Z2=(UCL-23)/ (Standard deviation)=(22.97462-23)/(Standard Dev)= -0.02928264

Thus P(Z<-0.02928264)-P(Z<-6.029283) =0.4884177=48.83177% (In R)

Therefore, there is a 48.8% chance that a type 2 error will occur for the technology class. This is not too high, and it can be interpreted that 100-48.8%=51.2% of the time, the customers received a product of standard and the process is stable. This should be inspected and improved, because it means that almost half the time, a customer will received a product that is no of standard and this will thus affect our sales and revenue.

A type 1 error, otherwise known as a Manufacturer's error, is a well-known term in quality engineering. A type 1 error occurs when we reject the null hypothesis(H0) when it is true. In other words, it is a false positive conclusion such as being told something is wrong, given it is correct.

In this specific case, it is the rejecting the null hypothesis that the process is under control. When in fact it is under control. We investigate the likelihood of making a type 1 error when identifying x-bar(mean) samples outside the control limits(A) and when identifying the highest number of consecutive standard deviation values with -0.3 and +0.4 of the control limits.

5. Hypothesis tests with MANOVA

The multi variate analysis of variance (MANOVA) is a tool we can employ to test multiple dependent variables and combining them into a compound or composite variable. This compound variable is then tested against an independent variable

The testing or comparing, of the dependent and independent variable enable us to conclude whether a combination of independent variables (or the independent variables themselves) influence the target variable.

The analysis includes comparing the p-value given by the output with the chosen significance value (or alpha). If the p-value is less than alpha, the MANOVA test is significant, otherwise it is not significant.

An alpha value of 0.05 was selected. Thus, we accept a 5% risk of stating that a relationship exists when no relationship exists. The first MANOVA analysis will be performed as follows:

Null hypothesis:(H0): Price and Class of each purchase do not influence the delivery time

Alternative Hypothesis(H1): Price and Class of each purchase influence the delivery time

```
                Df   Pillai approx F num Df den Df    Pr(>F)
Delivery.time    1 0.039093     3661      2 179975 < 2.2e-16 ***
Residuals   179976
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
 Response Price :
                Df     Sum Sq    Mean Sq F value    Pr(>F)
Delivery.time    1 6.8175e+11 6.8175e+11    1576 < 2.2e-16 ***
Residuals   179976 7.7852e+13 4.3257e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Class :
                Df Sum Sq Mean Sq F value    Pr(>F)
Delivery.time    1  31276 31275.8  7319.7 < 2.2e-16 ***
Residuals   179976 769008     4.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The given P-value is 2.2e-16, this is lower than the alpha value of 0.05. This means that the means for the delivery times are not the same, and thus the price and class have a significant influence on the delivery times. Thus, rejecting the Null hypothesis.

We also used a MANOVA to test the relationship between the features day and month and the target feature Class. It is performed as follows:

Null hypothesis:(H0): Month and day of purchase do not influence the Class of item purchased

Alternative Hypothesis(H1): Month and day of purchase influence the Class of item purchased

```
              Df     Pillai approx F num Df den Df Pr(>F)
Class          6 9.0285e-05   1.3541     12 359942 0.1801
Residuals 179971
```

```
 Response Month :
               Df  Sum Sq Mean Sq F value Pr(>F)
Class           6      87  14.576  1.2219 0.2913
Residuals   179971 2146871  11.929

 Response Day :
               Df   Sum Sq Mean Sq F value Pr(>F)
Class           6      668 111.302   1.488 0.1777
Residuals   179971 13461680  74.799
```

Seeing that the P-value given by the output is 0.1801, which is larger than the alpha value of 0.05. This means that means of the target feature are the same. In other words, we accept our null hypothesis and conclude that there is no corelation between the day and month and the class of the item purchased

Part 6: Reliability of the service and products.

6.1 Problem 6 and 7 of chapter 7 (page 363)

Question 6

We have specifications that we sent to our subsidiary, Lafrigeradora, for a company. The blueprint specification is 0.06+ - 0.04 centimeters(cm). It costs $45 to scrap an item that is outside the limits of the specifications. We were tasked with determining the Taguchi Loss function.

The Taguchi Loss function is given by:

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

With **L**=loss incurred, **k**=a constant, **x**=actual value or size of product, **T**=Target value.

In this question:

$$k = \frac{45}{(0.04)^2} = 28125$$

Thus: $L = 28125(x - 0.06)^2$

Question 7

A team formed to study the refrigerator part in Question 6, the team eventually found a way to reduce the scrap cost to $35. We first determined the Tugachi Loss function:

a)

$$k = \frac{35}{(0.04)^2} = 21875$$

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

b) The process deviation is shown to be 0.027cm, and thus (x-m) = 0.027:

$$L = 21875(0.027)^2 = \$15.94$$

6.2 System reliability

(Problem 27 of Chapter 7)

The Question and Problem statement: "Magnaplex, Inc. has a complex manufacturing process, with three operations that are performed in series. Because of the nature of the process, machines frequently fall out of adjustment and must be repaired. To keep the system going, two identical machines are used at each stage; thus, if one fails, the other one can be used while the first one is repaired."

We have been given the reliabilities of the respective machines, namely:

Machine A- 0.85

Machine B- 0.92

Machine C- 0.90

a) The system reliability is to be calculated. We assume that only one machine at each stage is used.

System reliability = 0.85 * 0.92 * 0.90 = 0.7038

b) Now we want to see how the reliability will improve if we have two machines per stage:

Probability that both As fail = (1 - 0.85) * (1 – 0.85) = 0.0225

R(AA)=Combined reliability for both As = 1 – 0.0225 = 0.9775

Probability that both Bs fail = (1 – 0.92) *(1 – 0.92) = 0.0064

R(BB)=Combined reliability for Bs = 1 – 0.0064 = 0.9936

Probability that both Cs fail = (1 – 0.90) *(1 – 0.90) = 0.01

R(CC)=Combined reliability for Cs = 1 – 0.01 = 0.99

Total Reliability = R(AA)*R(BB)*R(CC)= 0.9775*0.9936*0.99= 0.96153156

Improvement thus is by 36.62%

6.3 Delivery Process Analysis

Here, we are tasked with answering a practical question: How many days per year can we expect reliable delivery times? The business currently has 21 vehicles and 21 drivers available. To ensure a reliable process, we must investigate the reliability of the vehicles together with the reliability of the drivers.

In this investigation, we used a binomial distribution for the vehicles and the drivers respectively. Our task was to solve for the probabilities of "successes" (i.e., 0 defects) in both cases.

P(Reliable delivery vehicles on a day)= 0.8615412

P(Reliable delivery driver on a day)= 0.9344269

We can then calculate the days in a year (365 days) in which zero vehicles and zero drivers will be defective as:

P(Reliable Delivery)= P(Reliable delivery vehicles on a day) * P(Reliable delivery driver on a day)* 365

= 0.8615412 * 0.9344269 *365

= 293.8422 days= 2

If we were to increase the number of vehicles to 22, the probability that there are zero defective vehicles becomes:

P(Reliable delivery vehicles on a day)= 0.8615661

Thus, we can calculate the number of days in a year where we will face no defects again:

P(Reliable Delivery)= P(Reliable delivery vehicles on a day) * P(Reliable delivery driver on a day)

= 0.8615661 * 0.9344269 *365

= 293.8508

We can thus see, that adding an additional vehicle as does very little in terms of the number of days in which we will have a reliable delivery

7. Conclusion

In this report, we analysed the dataset representing the sales of the online business with respect to variables or factors pertaining to these sales. These variables describe the details pertaining to these sales. Upon analysis, suggestions and patterns were drawn from the graphs and figures generated, in order to advice the company on which aspects of the business should seek to be improved.

Prior to analysis, data cleaning and wrangling was performed. This included removing missing values and invalid values. Since these may greatly affect the results of the analysis, it is important to only work with valid data. The data was subsequently ordered in ascending order to analyse the quality of the data with respect to control limits. These control limits showed us which sale categories need urgent attention by taking samples of the sales and observing which one of their standard deviations and means are out of control. Many of the process need extensive investigation and improvement.

The following analysis was understanding the likelihood of type 1 and type 2 errors occurring and translating these into what it means for the business. A MANOVA was also performed to understand the relationship between features and how this can be maximised to ensure maximised profit. We finally examined the reliability of the business service providers(transport of food) and also optimised this delivery process to maximise the businesses profitability.

8. References

The experts at dummies. (2022) 'How to use the Z-Table', 8 August.

Available at: https://www.dummies.com/article/academics-the-arts/math/statistics/how-to-use-the-z-table-147241/ (Accessed:22 October 2022)

RCoder(2020)'Normal Distribution in R', 9 September.

Available at: https://r-coder.com/normal-distribution-r/ (Accessed:22 October 2022)

Bhandari,PB.(2022) 'Type I and Type II Errors, Differences, Examples, Visualizations', 2 September.

Available at: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/ (Accessed:22 October 2022)

QIMacros(2018) 'Control Limits are the key to Control Charts', 15 June.

Available at: https://www.qimacros.com/free-excel-tips/control-chart-limits/index2.php (Accessed:22 October 2022)