Quality Assurance

ECSA Graduate Attributes Project

Christiaan Johannes Pieterse

23576278

October 2022

Abstract

This report discusses and analyses a client dataset for an online business. The data is processed and wrangled to produce a valid dataset that could be used for analysis, plotting and calculations. The descriptive features are first discussed before any analysis is done. This includes the features, importance of classes, delivery time observations, Marketing, and process capabilities. A statistical process control approach is then used to analyze each class within the dataset. The main purpose of this is to determine whether certain processes are in-control or out-of-control. The delivery process is optimized next. Several aspects and techniques are used to do so. Some of them include looking at the out-of-control samples identified previously. Type 1 and 2 errors are identified alongside the probability of them happening. The optimal delivery times that will reduce cost is calculated. Lastly DOE and MANOVA analysis is done for even more insight between features.

Table of Contents

List of tables

List of figures

Introduction

Data analysis is very important for a company, especially companies with big online aspects. Data analysis gives a company better insight into the smaller details of their company and suggests where room for improvement is. This will help increase the companies profit, customer satisfaction and effectiveness. This report will do a data analysis of the provided dataset and give suggestions for improvements throughout.

Part 1: Data wrangling

The dataset consists of 10 features (columns), each with 180 000 instances (rows). First the data is reordered to be sorted according to increasing values, first in the “Year” feature, then “Month, “Day” and lastly in the “X” feature. Data wrangling entails splitting a dataset into valid- and invalid data. The “Price” feature is used to filter the valid data from the invalid data. Invalid data includes Price instances that are negative or missing. Negative instances exist due to typing errors or within the finance books, a return could be wrongly entered as a sale. The first 10 rows of the two different datasets are shown below in table 1 and 2 below.

X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
463	47101	50	Clothing	1030.86	2021	1	1	9	Recommended
2627	88087	21	Clothing	428.03	2021	1	1	10	Recommended
3374	25418	68	Household	13184.41	2021	1	1	48.5	Website
5288	13566	94	Household	7021.9	2021	1	1	42	Recommended
8182	84692	35	Clothing	475.18	2021	1	1	9	Recommended
9272	46305	72	Clothing	580.98	2021	1	1	8.5	Random
9712	92105	45	Household	6877	2021	1	1	43	Recommended
12163	21614	27	Clothing	513.13	2021	1	1	9.5	Recommended
12195	12174	56	Household	14538.64	2021	1	1	41.5	EMail

Table 1: Table of the valid data

X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
98765	64288	25	Clothing	NA	2021	1	24	8.5	Browsing
54321	62209	34	Clothing	NA	2021	3	24	9.5	Recommended
34567	18748	48	Clothing	NA	2021	4	9	8	Recommended
155555	33583	56	Gifts	NA	2022	12	9	10	Recommended
144443	37737	81	Food	-588.8	2022	12	10	2.5	Recommended
177777	68698	30	Food	NA	2023	8	14	2.5	Recommended
16320	44142	82	Household	-588.8	2023	10	2	48	EMail
56789	63849	51	Gifts	NA	2024	5	3	10.5	Website
19998	68743	45	Household	-588.8	2024	7	16	45.5	Recommended

Table 2: Table of the invalid data

Part 2: Descriptive statistics

2.1 Features

The valid dataset contains 10 descriptive features. Each feature is described and classified in table 3 below.

Number	Feature	Class	Meaning
1	X	Continuous	Original index of the dataset.
2	ID	Continuous	Client/order identification number. Infinite number of combinations.
3	AGE	Continuous	Age of client.
4	Class	Categorical	Type of item sold or the category under which it falls.
5	Price	Continuous	Price of the item sold.
6	Year	Categorical	Year sale is made.
7	Month	Categorical	Month sale is made.
8	Day	Categorical	Day sale is made on.
9	Delivery.time	Continuous	Time taken for item to be delivered in days.
10	Why.Bought	Categorical	How the client heard about the item sold.

Table 3: Descriptive table of the features in the dataset.

2.2 Importance of classes

Each of the classes within the dataset must be analyzed to determine which classes are most popular and which generate the most revenue. Knowing this information will help identify the companies strengths and weaknesses. Focus can be shifted to the more important classes and less important classes can be reconsidered or eliminated.

In figure 1 below the frequency (sales count) of each class is compared to one another. As seen below, Gifts and Technology has the highest frequency with 39149 and 36347 sales respectively. Luxury has the least sales. This figure might indicate that Gifts and Technology are the more important features and Luxury less important. This is not quite true since the revenue generated from each class is not considered.

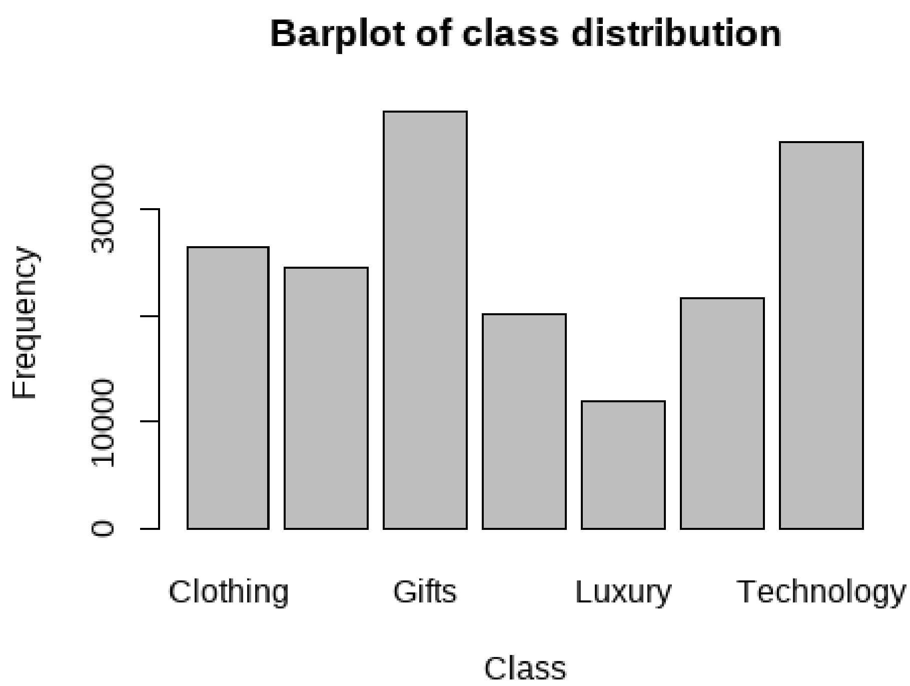


Figure 1: Bar plot of each class's frequency

In figure 2 below the total revenue generated from each class is compared to one another. Technology generates the most revenue and Luxury the second most. Figure 2 is very different to figure 1 due to each class having different prices. More sales from one class does not necessarily mean more revenue when comparing classes. This is due to different average prices within each class and can be seen in figure 3. Take Luxury vs Sweets for instance. Sweets have more sales compared to Luxury, but Luxury has more revenue. This is due to luxury being more expensive items compared to sweets that are relatively cheap. Also observed in figure 3 below, apart from Luxury mentioned above, the Technology (and slightly the Household items) are also on average more expensive than the rest of the classes.

The focus of the company should thus be on Technology and Luxury since they make up 71% of the total revenue. Furthermore, Gifts and Household take up 24% and Food, Clothing and Sweets generate only 5%, which very little revenue even though they had relatively large sales. This indicates that the company should consider advertising the last-mentioned classes, or possibly discount selling them completely.

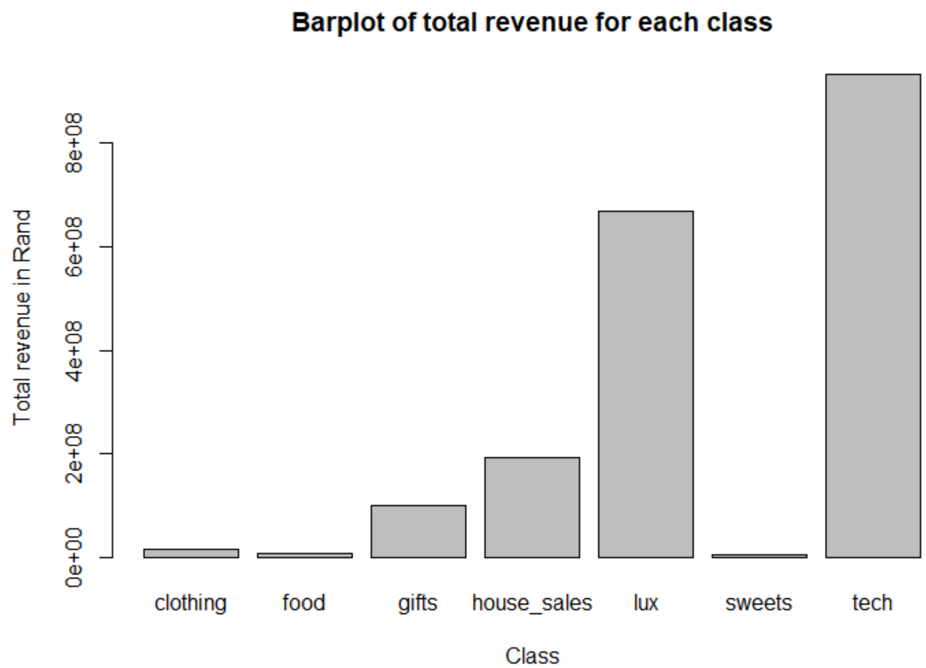


Figure 2: Bar plot of each class's revenue generated

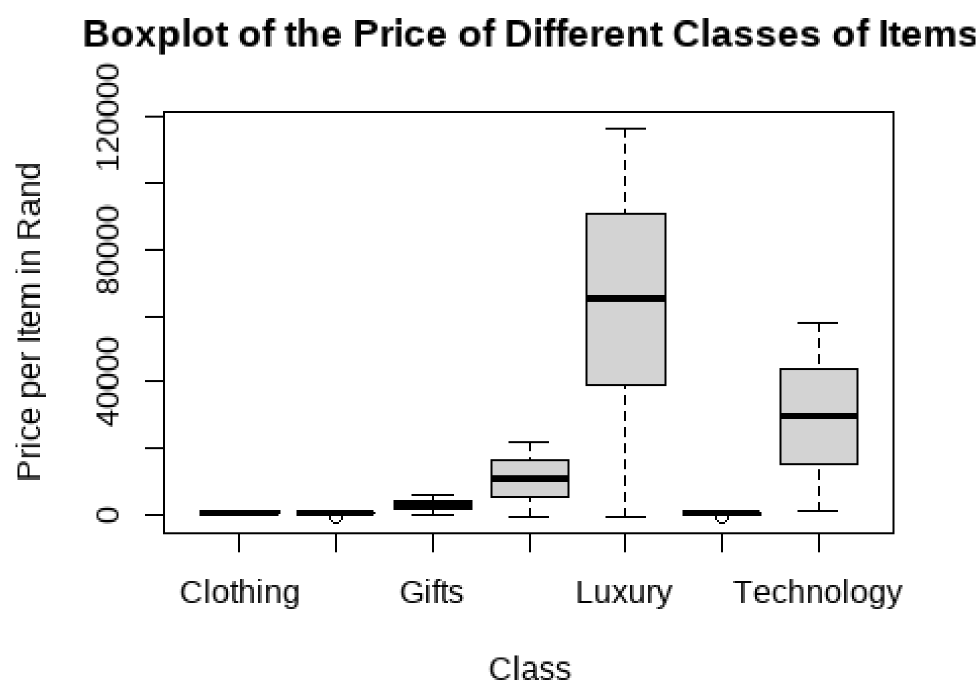


Figure 3: Box plot of each class's price

2.3 Delivery time observations

The delivery time distribution can be seen in figure 4 below. It is observed that it has the highest frequency for the lowest time and a sharp decrease onwards. There is however a spike further to the right, towards the higher delivery times. This indicates that further investigation must be done to decrease or eliminate the spike in delivery times.

Figure 4: Delivery time distribution

Figure 5 below compares the mean delivery time for each class. It is observed that there is a high variation between the mean delivery times. Food items have a very low delivery time, which is expected since food has expiration dates. Luxury items have a low delivery time, which is favorable for the company since luxury items are one of the highest earning classes (as discussed above). Technology has a medium delivery time. This needs to be improved since Technology is the companies highest earning class. Household items have the highest mean delivery time by far, which causes the spike in figure 3 above. The Household items have the 3rd highest revenue generated and thus needs some attention as well.

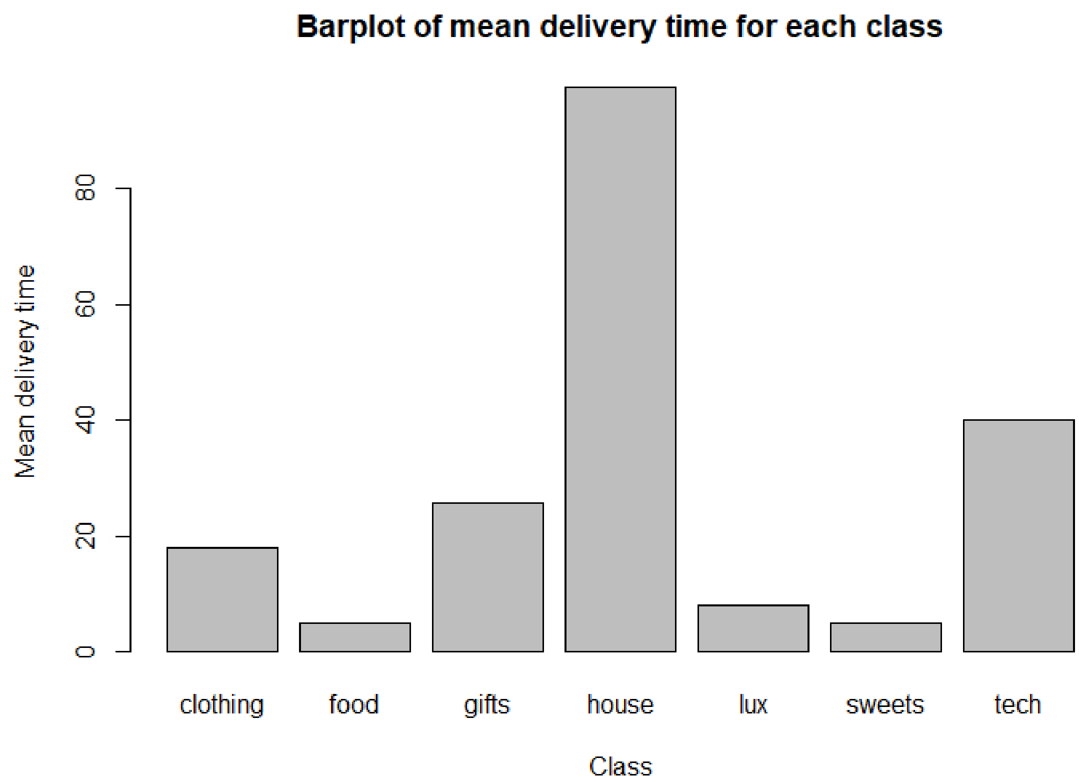


Figure 5: Bar plot of mean delivery time for each class

2.4 Marketing

To improve the companies marketing strategies, a better understanding of the customers reason for purchases is required. Figure 6 below shows the frequency of every customer purchase reason. Recommended has the highest frequency and accounts for 60% of all sales. This means that 60% of the companies sales are due to word-by-mouth and not marketing. The quality of customer service must be one of the main focusses of the company so that they can ensure happy customers and keep a good image via their customers. The website is responsible for the second most sales. This might even link to customers hearing about products via word-by mouth but ordering via the website instead. A good website design will keep customers engaged and interested in buying. Email and spam are the lowest two purchase reasons. The company will thus benefit more if they focus on other marketing strategies that Email or Spam.

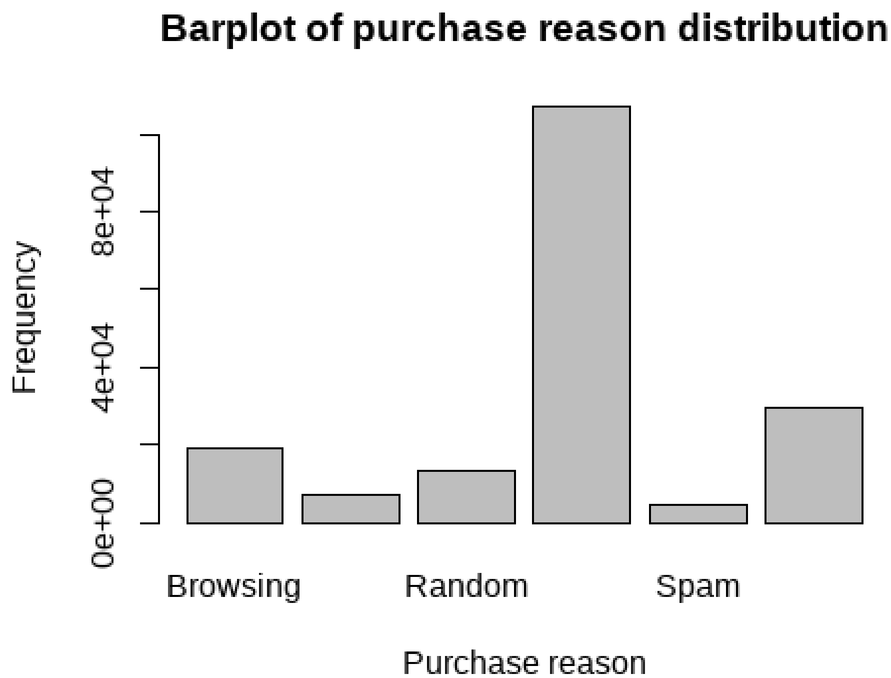


Figure 6: Bar plot of the customers purchase reason

2.5 Process Capability

The process capability of the company, specifically for technology items, is determined to gain possible improvements and insights to the process.

USL = 24 days

LSL = 0 days

$Standard\ deviation\ (\sigma)$ = 3.502

$Mean\ (\mu)$ = 20.011

The process LCL is 0 days since delivery times cannot be negative. Negative delivery times would realistically mean that the delivery occurred before the order was even placed, which is impossible. An LCL of 0 days is thus logical and realistic.

The Process Capability indices (Cp, Cpu, Cpl and Cpk) for the process delivery times of technology class items is calculated and shown below.

C_p = 1.142

C_{pu} = 0.380

```
Cpl = 1.905

Cpk = 0.380
```

The Cp value, in essence, determines how ‘under control’ a process is. The companies Cp > 1 which indicates that the process can reach the goal it is designed for. The Cpk value is equal to the lowest value between the Cpu and Cpl values. The Cpk > 0, thus conforming that the process is under control. (Note: Both the Cpu and Cpl values are > 0, which is a good indication for the Cpk being > 0 and the process being under control.)

Part 3: Statistical Process Control

3.1 Control limits

Before calculating the control limits, the data is first sorted chronologically from oldest to newest as described in Part 1 above (Year, Month, Day, X). The first 450 instances of the ordered data is then used to make 30 samples, consisting of 15 instances in each sample. The control limits for both the mean (x-table) and standard deviation (s-table) are calculated using the 30 samples. The x-table and s-table are shown in table 4 and 5 below.

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	9.589	9.459	9.33	9.2	9.07	8.941	8.811
Household	49.237	47.636	46.035	44.433	42.832	41.231	39.63
Food	2.637	2.569	2.501	2.433	2.365	2.298	2.23
Technology	21.463	20.42	19.377	18.333	17.29	16.247	15.203
Sweets	2.522	2.403	2.285	2.167	2.048	1.93	1.812
Gifts	9.41	8.973	8.537	8.1	7.663	7.227	6.79
Luxury	5.064	4.854	4.643	4.433	4.223	4.013	3.803

Table 4: Control limits of mean from samples.

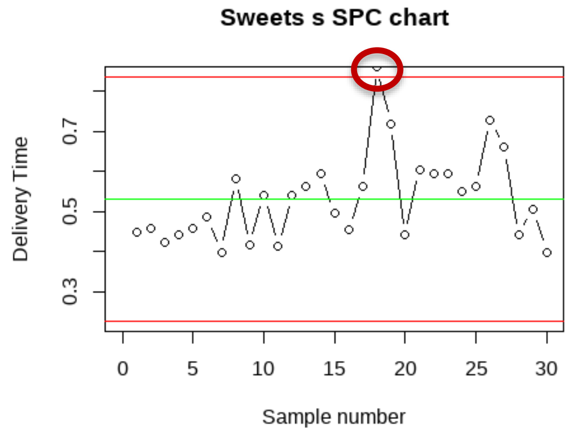
Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	0.775	0.681	0.587	0.493	0.399	0.305	0.211
Household	9.571	8.41	7.249	6.088	4.927	3.767	2.606
Food	0.406	0.357	0.307	0.258	0.209	0.16	0.111
Technology	6.236	5.48	4.724	3.967	3.211	2.454	1.698
Sweets	0.707	0.621	0.536	0.45	0.364	0.278	0.193
Gifts	2.61	2.294	1.977	1.66	1.344	1.027	0.711
Luxury	1.256	1.103	0.951	0.799	0.647	0.494	0.342

Table 5: Control limits of standard deviations from samples.

3.2 S-Charts

S-Charts determine whether the corresponding X-Chart is reliable or not. The S-Charts are thus always considered first. An S-Chart is out of control when it has a value above the UCL or below the LCL, and the corresponding X-Chart is then considered unreliable.

The S-Charts for the 30 samples are shown in figures 7 - 13 below.



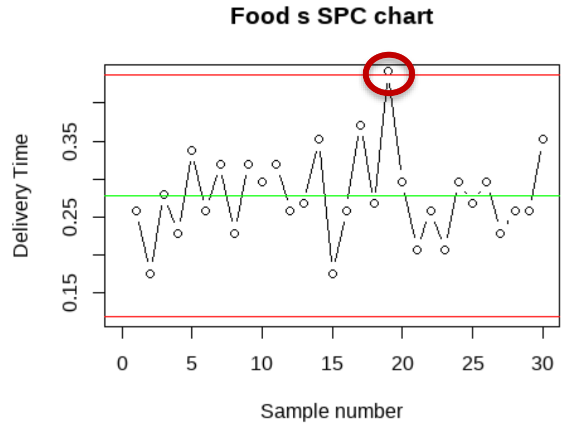
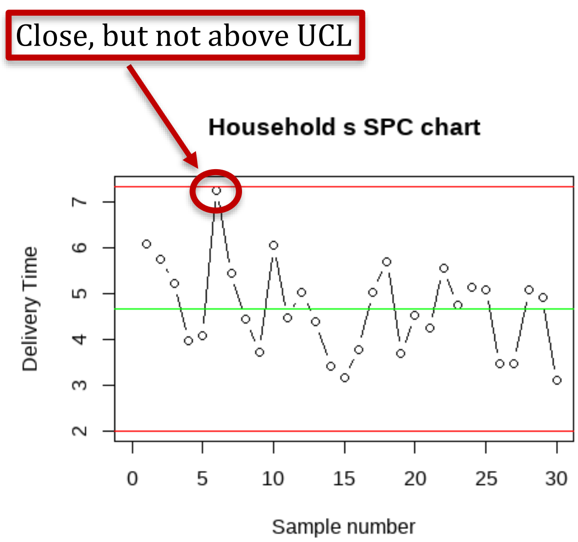


Figure 7: Sweets S-Chart

Figure 8: Food S-Chart

It is observed that the Sweets and Food S-Charts are out of control because they have instances above the UCL (See figure 7 and 8 above). The Sweets and Food X-Charts are thus considered unreliable.



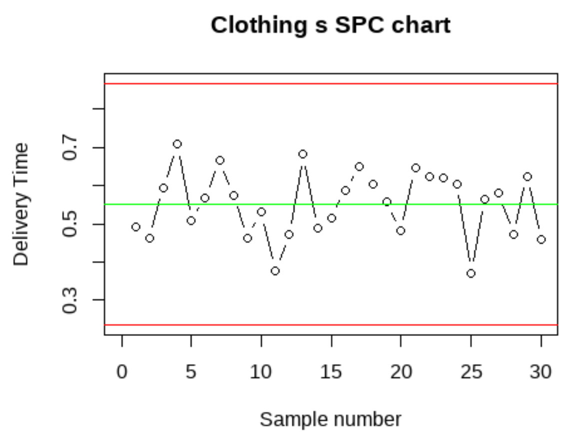


Figure 9: Household S-Chart

Figure 10: Clothing S-Chart

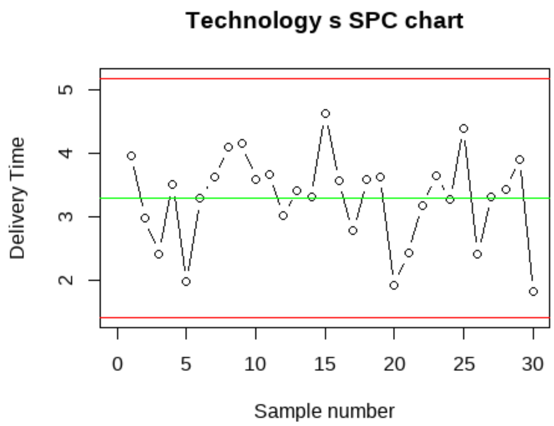


Figure 11: Technology S-Chart

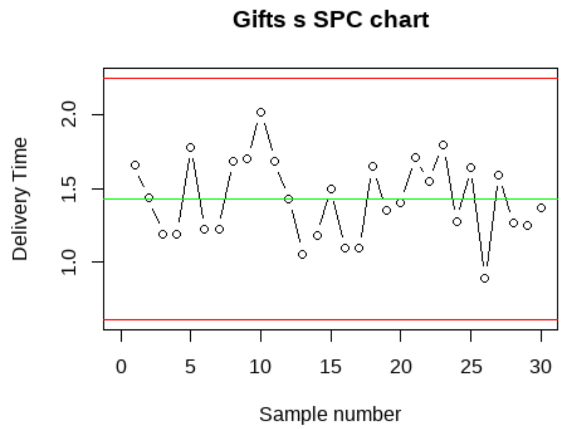


Figure 12: Gifts S-Chart

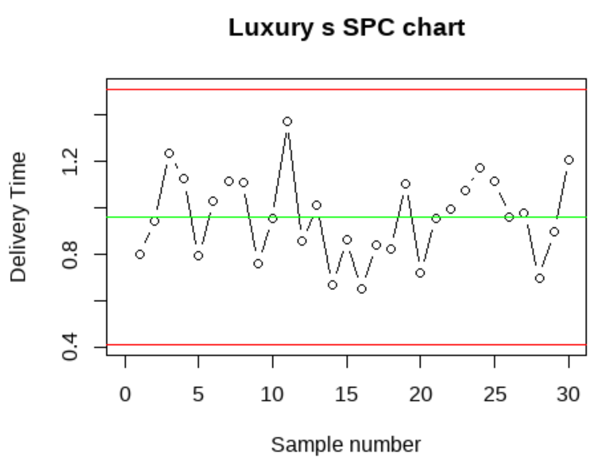


Figure 13: Luxury S-Chart

The Household, Clothing, Technology, Gifts and Luxury S-Charts are in control (see figure 9 to 13 above) and thus the corresponding X-Charts can be considered reliable.

3.3 X-Charts

X-Charts determines whether is in statistical control or not. The delivery process of the business is assessed according to the time (in days) it takes for the purchased item(s) to be delivered to the client. For this dataset, the X-Chart would determine whether the delivery time for each class conforms to customer satisfaction.

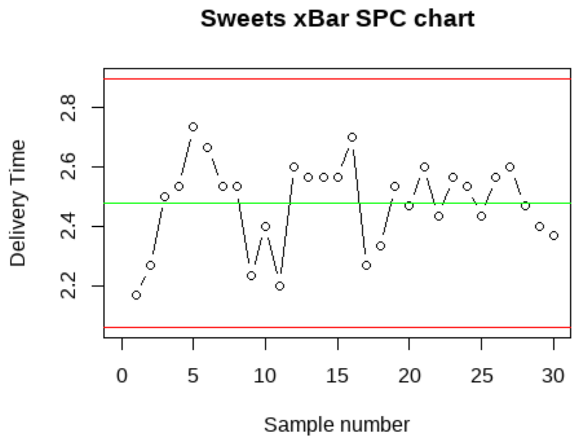


Figure 14: Sweets X-Chart

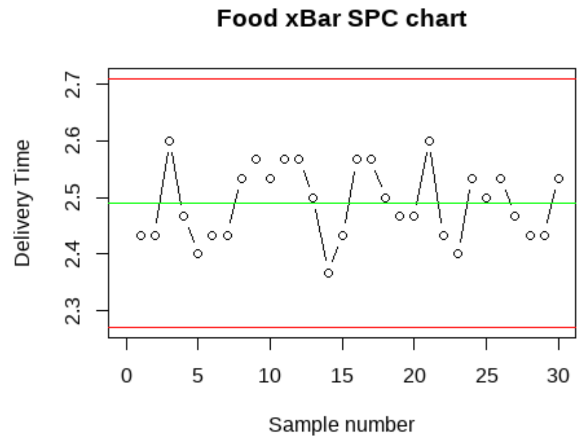


Figure 15: Food X-Chart

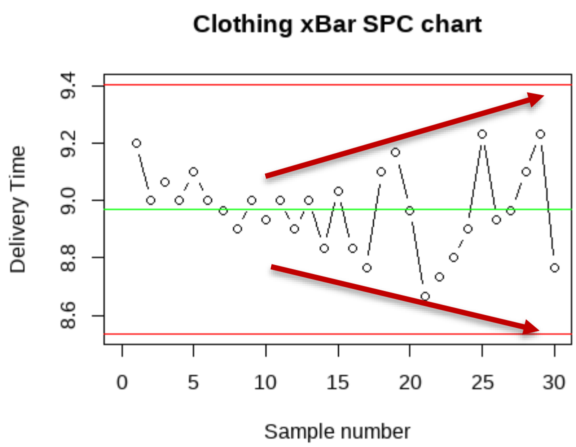


Figure 16: Clothing X-Chart

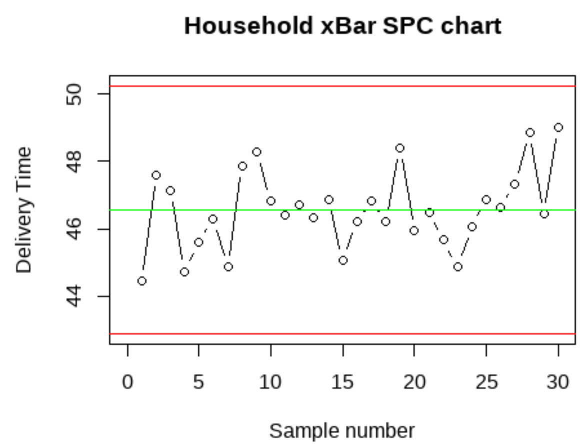


Figure 17: Household X-Chart

The sample numbers in Figure 16 above have small variation from sample 0 to 8. From 9 to 30 the variation increases as indicated with the dark red arrows. This indicates that the process is moving towards an out-of-control system as time increases. This should be closely analysed for future samples.

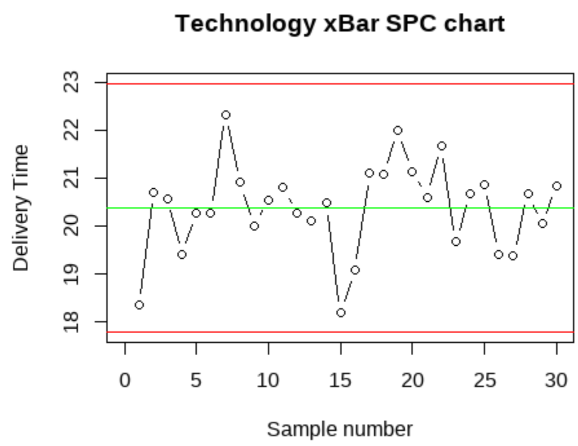


Figure 18: Technology X-Chart

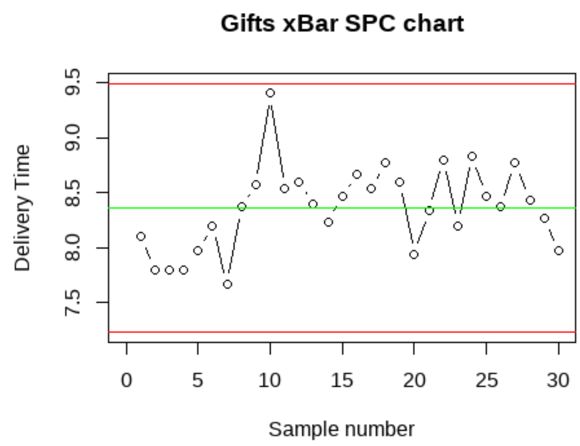


Figure 19: Gifts X-Chart

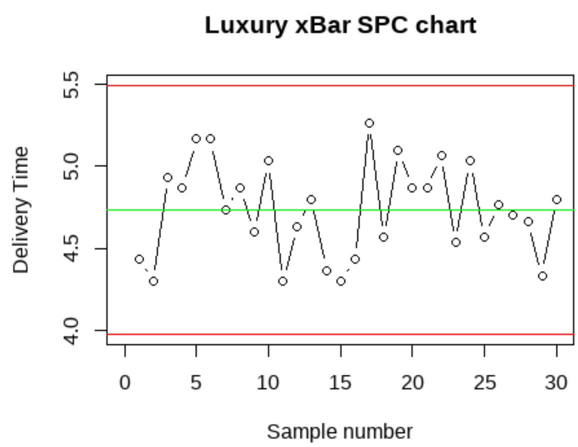
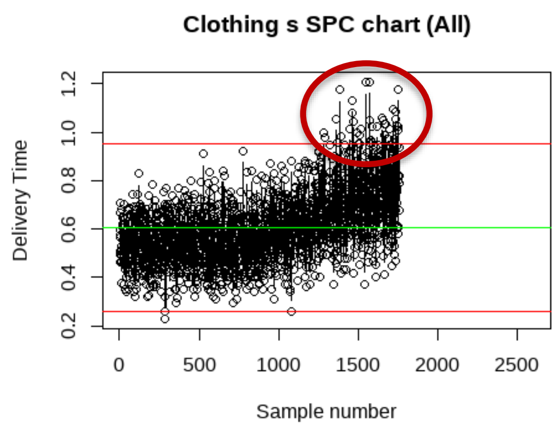


Figure 20: Luxury X-Chart

3.4 S-Chart and X-Chart for all samples

It is found that for the oldest 30 samples all the classes except Sweets and Food are statistically in control. To determine whether the rest of the process is in statistical control, the rest of the samples' S- and X-Charts need to be considered in the same manner previously done. Through doing this we can see how the more recent data performs which will give the company a better understanding of their process from the oldest to the latest data. The resulting S- and X-Charts are shown in figure 21 – 33 below.



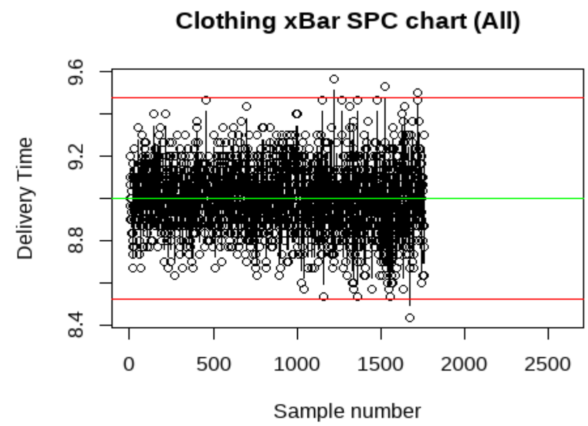
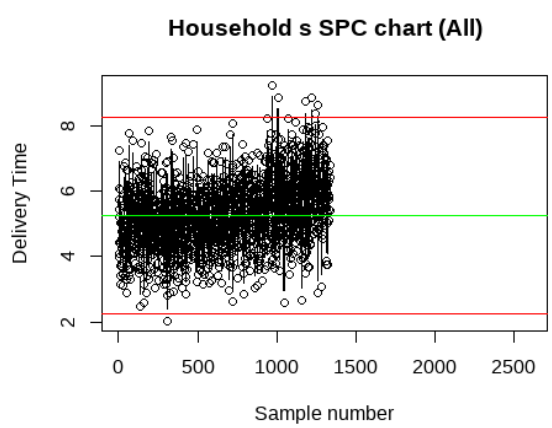


Figure 21: Clothing S-Chart for all samples

Figure 22: Clothing X-Chart for all samples

The Clothing S-Chart clearly shows that the process used to be very well in statistical control but around sample 1300 started experiencing a heavy increase in samples above the UCL. Even though the X-Chart seems to indicate that the process is in control (low number of instances outside of UCL and LCL), it is not reliable since the S-Chart is completely out of control.



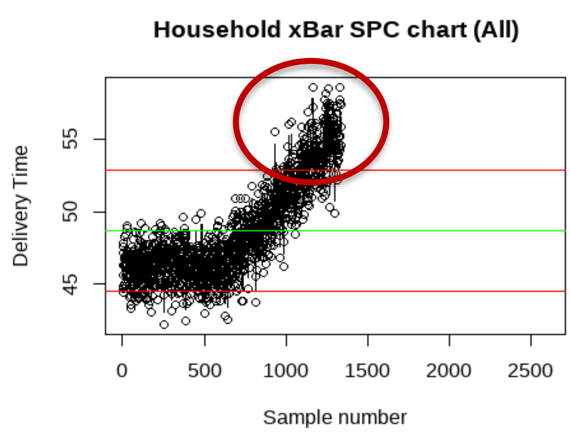
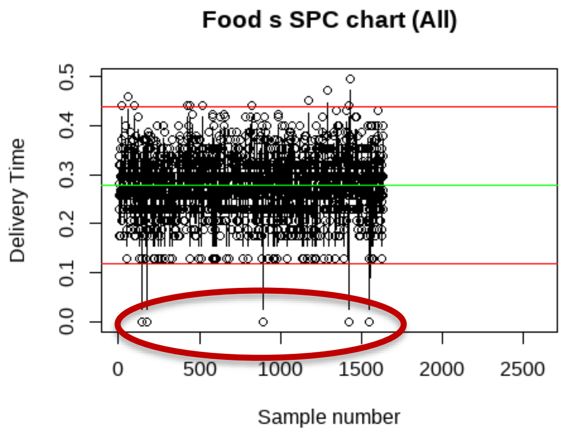


Figure 23: Household S-Chart for all samples

Figure 24: Household X-Chart for all samples

The Household S-Chart seems to be in statistical control and only experiences an increase in out-of-control samples towards the end. The X-Chart on the other hand is clearly out of control with many samples below the LCL in the start and at the end above the UCL. It is also observed that the S-Chart is in control at the start where the X-Chart is out of control.



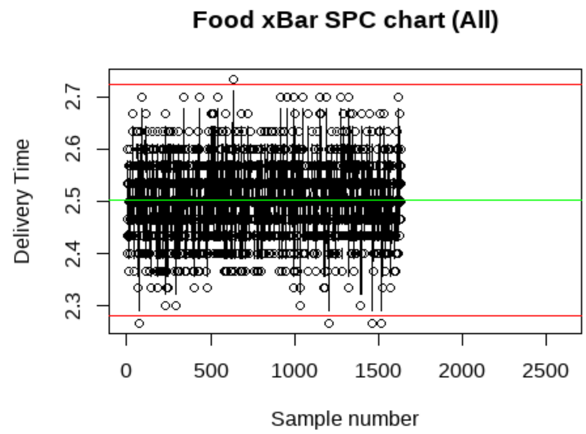
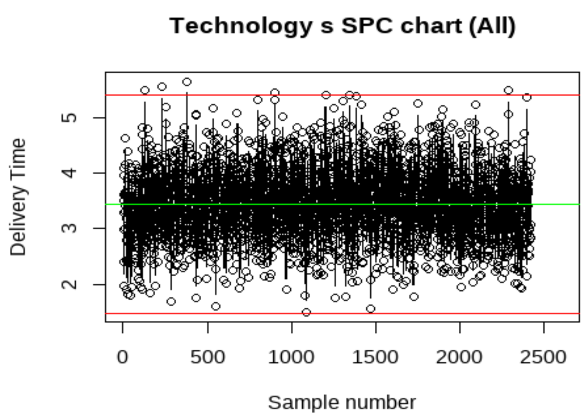


Figure 25: Food S-Chart for all samples *Figure 26: Food X-Chart for all samples.*

The Food S-Chart has quite a few samples very far below the LCL, but they all tend towards zero so could thus be seen as possible outliers. The X-Chart indicates a in-control process with minimal instances above the UCL and below the LCL. The company can be satisfied with the Food delivery process.



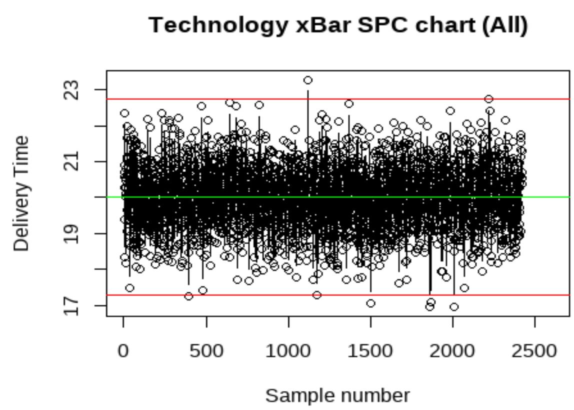


Figure 27: Technology S-Chart for all samples *Figure 28: Technology X-Chart for all samples*

Both the S- and X-Chart for technology shows more desired variation. The center of the S-Chart has moved slightly upwards and now has a few instances above the. The variation is thus more than desired, but this does not mean that the process is entirely out of control. The X-Chart has only one instance above the UCL and a few below the LCL which can be identified as possible outliers. The X-Chart thus indicates that the system is statistically in control. Since Technology is the highest earning class, the company can thus be satisfied with the delivery process being overall in control.

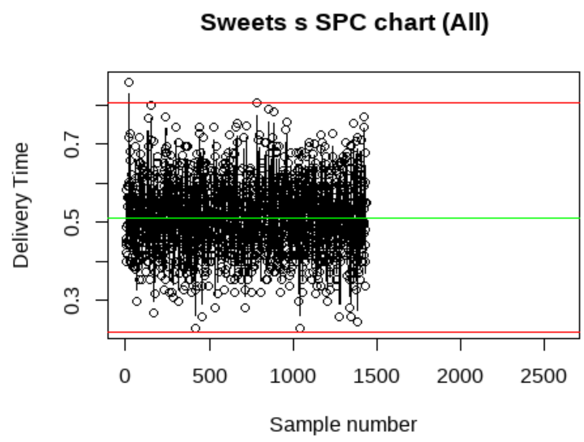


Figure 29: Sweets S-Chart for all samples

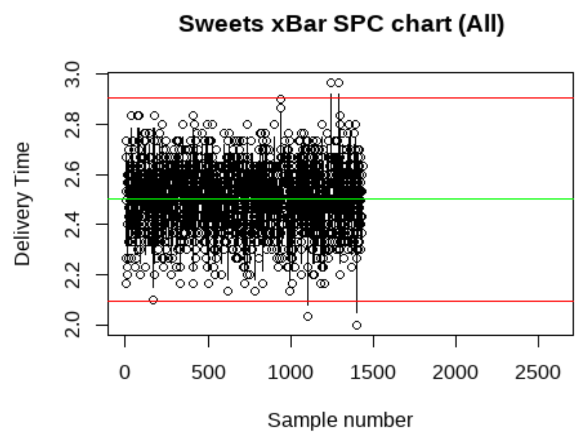


Figure 30: Sweets X-Chart for all samples

Both the S- and X-Charts are statistically in control with minimal outliers outside of the control limits. Sweets has on average a low revenue and delivery time so the company can be satisfied with the performance.

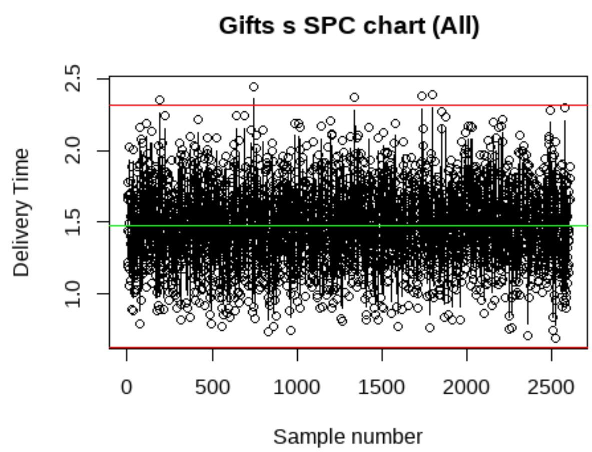


Figure 31: Gifts S-Chart for all samples

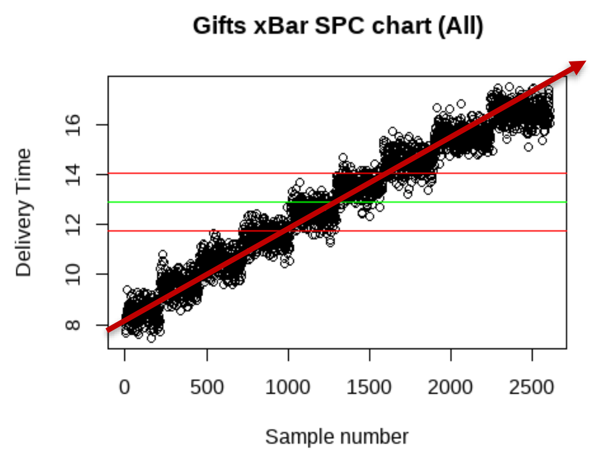


Figure 32: Gifts X-Chart for all samples

The S-Chart for Gifts seems in control with minimal samples above the UCL. The X-Chart is clearly out of control. A pattern is recognized with a clear increase in delivery time every 250 samples. This indicates that the delivery time increases every year by about 2 days. Gifts is one of the higher earning classes so the company needs to address this problem.

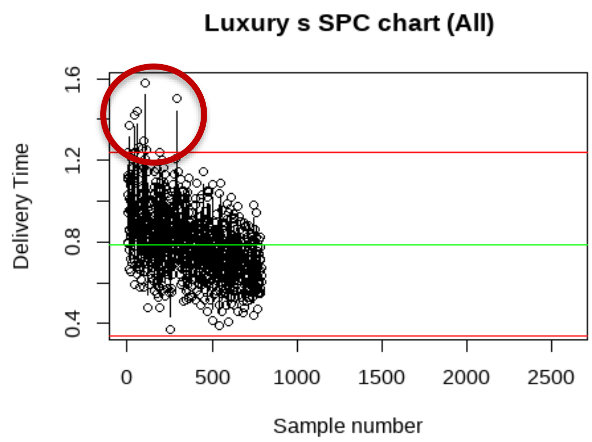


Figure 32: Luxury S-Chart for all samples

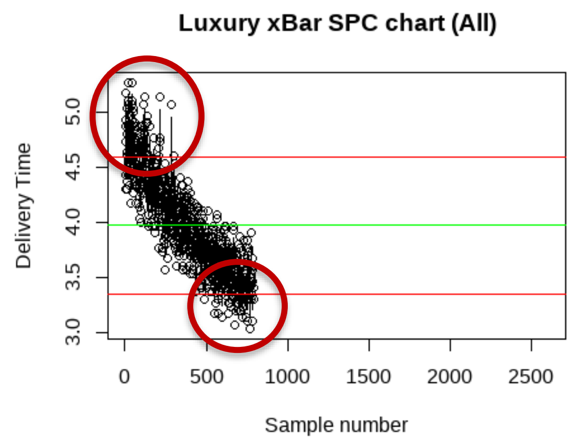


Figure 33: Luxury X-Chart for all samples

The Luxury S-Chart has quite a few samples above the UCL at the start but seems to be very controlled towards the end. The X-Chart, however, does not indicate the same. There are many out of control samples at the start and a lot towards the end as well, where the S-Chart only had a few at the start and none towards the end. Both the charts have a clear downward trend which can be good since the company would like to decrease delivery times, but the process is out-of-control due to this reason which is not good for the company. Luxury has the second most total revenue and thus requires attention to move the process towards statistically in-control.

Part 4: Optimizing the delivery process]

4.1 Out of control samples

4.1 A: A Sample means outside the outer control limits

Table 6 below shows the out-of-control samples for each class. It indicates the first three instances, last three instances and the count of all instances. It is observed that the classes that are in statistical control, only have a few instances out-of-control. Technology has the 2nd least out of control instances. This is very favorable for the company since Technology is the highest earning class. Luxury is the second highest earning class but has almost three times more of out-of-control instances than Technology. This requires more attention. Gifts have more out-of-control instances than all the other classes combined. This requires more attention.

Class	1st	2nd	3rd	3rd last	2nd last	Last	Count
Clothing	17	21	22	1756	1759	1760	202
Household	370	450	484	1335	1336	1337	481
Food	38	93	110	1537	1621	1629	35
Technology	7	19	22	2412	2413	2414	128
Sweets	4	5	6	1434	1436	1437	665
Gifts	213	216	218	2607	2608	2609	2308
Luxury	5	6	17	789	790	791	353

Table 6: Out-of-control samples

4.1 B: Consecutive samples of s-bar between -0.3 and +0.4 control limits

Table 7 below shows the maximum number of consecutive samples between -0.3 and +0.4 sigma control limits, as well as the final sample in this range for each class.

Class	Number of consecutive samples	Ending sample number
Clothing	12	83
Household	17	1316
Food	14	1329
Technology	17	1566
Sweets	11	188
Gifts	12	1907
Luxury	9	373

Table 7: Consecutive samples

4.2 Type 1 Error

A Type 1 error, also known as a manufacturing error, is when the true null hypothesis is rejected. In the case of this dataset, it means that and actually in-control process id found to be out-of-control. This will result in unnecessarily further investigation.

4.2.1 Probability of type 1 error for 4.1 A

```
(1-pnorm(3))*2*100 = 0.2699796

The probability of making a type 1 error is thus 0.270%
```

4.2.2 Probability of type 1 error for 4.1 B

```
pnorm(1.2, lower.tail = FALSE) + pnorm(-0.5) = 0.4236072

The probability of making a type 1 error is 42.361%. This means that there is
a 42.361% chance that an actual in-control process will be found to be out-of
-control.
```

4.3 Optimal Delivery Time

Figure 34 below shows the total cost associated with the decreased delivery time. The graph has a negative slope from 1-3 hours and a positive slope from 3-15 hours. This indicates that the lowest cost is at 3 hours. The delivery times should thus be decreased by 3 hours to reduce the total cost with R 880 156.2.

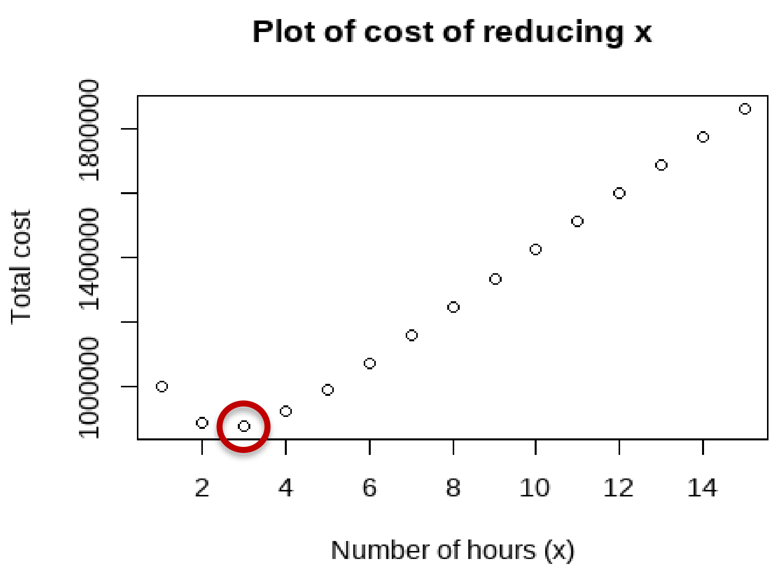


Figure 34: Plot of total cost as delivery time is decreased

4.4 Type 2 Error

A type 2 error, also known as a customers error, is when the false null hypothesis is accepted. In the case of this dataset, it means that an actual out-of-control process is incorrectly identified as in-control.

```
Probability = (pnorm(21.463, mean=23, sd=(21.463 - 15.203)/6) - pnorm(15.203,
mean=23, sd=(21.463 - 15.203)/6))*100 = 7.041%
```

The probability of a type 2 error is 7.041% which is also indicated in figure 35 below. The black graph represents the original graph with a mean of 18.33 and the blue graph on the right represents the new distribution with a mean of 23. 7.041% of the new normal distribution falls between the upper and lower control limit resulting I the occurrence of type 2 errors.

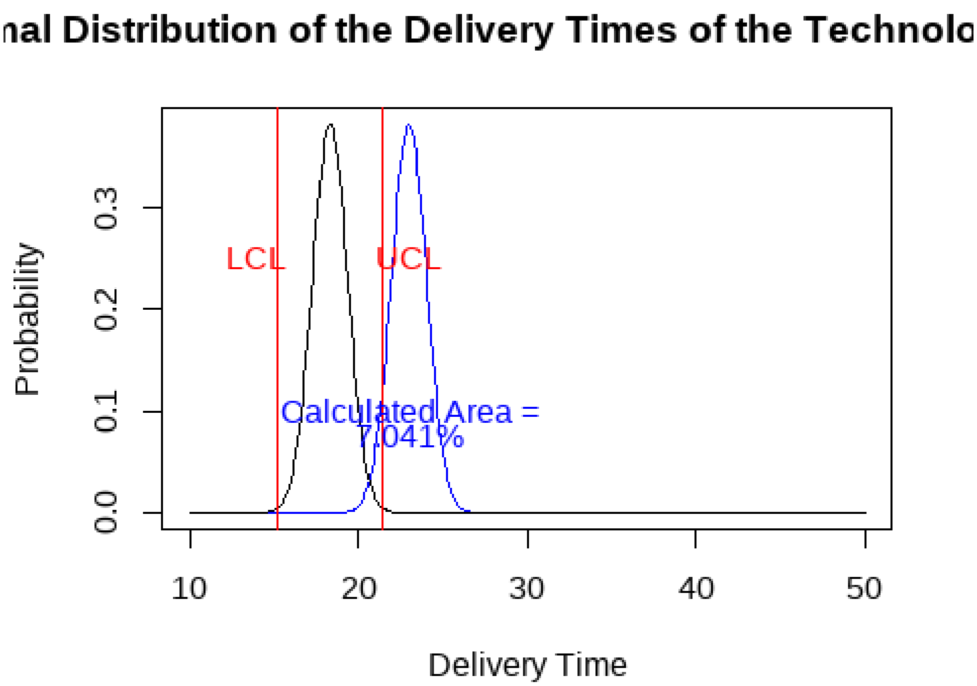


Figure 35: Normal distribution of the delivery times for Technology

Part 5: DOE and MANOVA

MANOVA is a technique that is used to determine the impact of different features on one another. The Price, Year and Delivery time classes are investigated. The constructed MANOVA has the following hypotheses:

Null hypothesis (Ho): The means of the features are equal. This implies that the features mentioned (Price, Year and Delivery time) does not have an impact on Class.

Alternative hypothesis (H1): At least one means differ from the rest. This implies that at least one of the features mentioned (Price, Year and Delivery time) have an impact on Class.

The Manove results are as follows:

```
               Df Pillai approx F num Df den Df    Pr(>F)
Class           6 1.7582    42466     18 539862 < 2.2e-16 ***
Residuals 179954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting p-value is 2.2e-16. This means that the null hypothesis must be rejected and thus it is concluded that at least one of the features (Price, Year and Delivery time) have an impact on Class. The Price, Year and Delivery time will thus be different for varying classes.

Figure 36 below shows the boxplot of all the MANOVA features. Their means and distributions are not similar and thus the null hypothesis must be rejected.

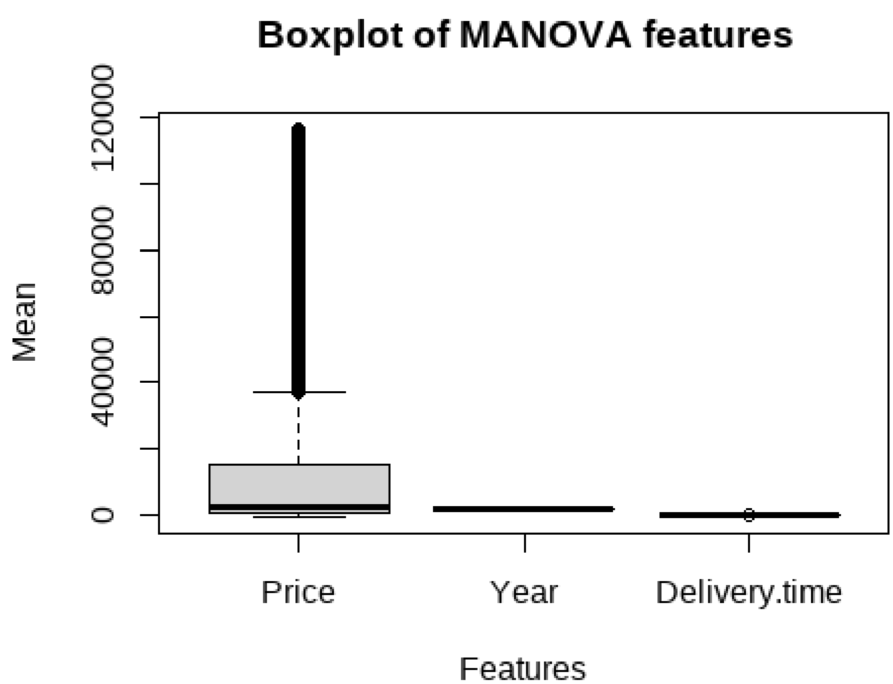


Figure 36: Boxplot of the MANOVA features

Part 6: Reliability of the service and products

6.1 Lafrideradora

```
L = k(y-m)^2
This means L = 45  when (y-m) = 0.04
k = 45/(0.04^2)
To plot make a vector between -1 and 1 by increments of 0.01 for y.
```

```
y = seq(-1,1, by = 0.01)

The Taguchi loss is thus:
L = 28125*(y - 0.06)^2
```

Figure 37 and 38 below displays the Taguchi loss function. The bigger the deviation, the more expensive it becomes. The goal should be thus to keep the deviation as low as possible. That is why the minimum point of the function is close to 0.

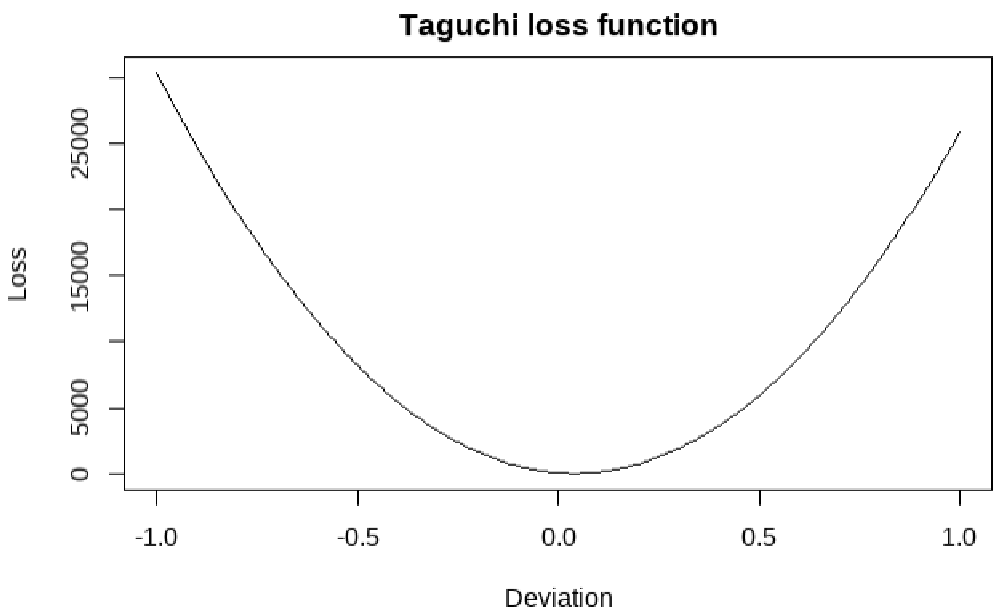


Figure 37: Taguchi loss function

```
# b

k = 35/(0.04^2)
y = seq(-1,1, by = 0.01)

The Taguchi loss is thus:
L = 21875*(y - 0.06)^2
```

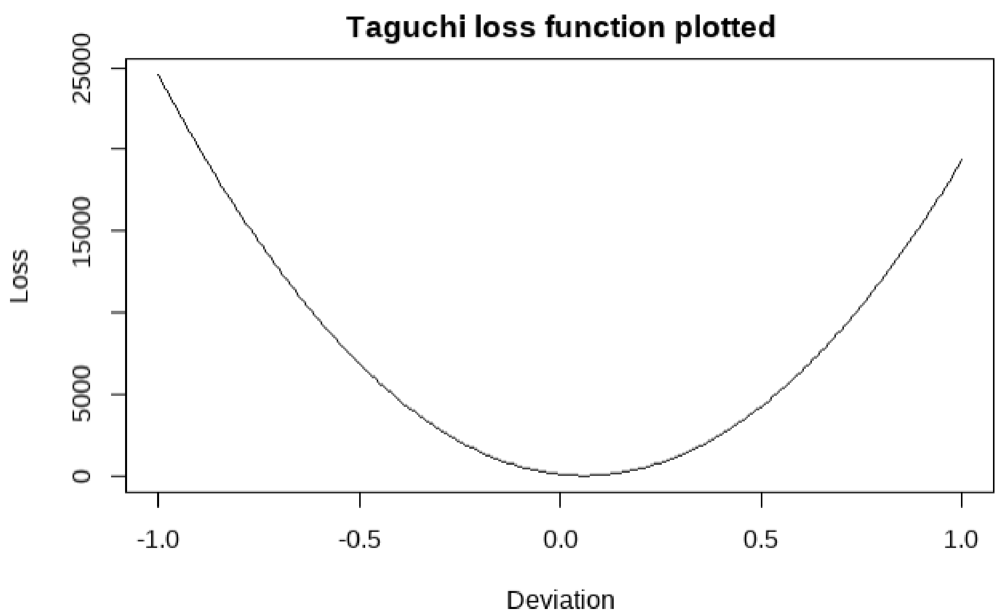


Figure 38: Taguchi loss function

```
# b
L = 21875*0.027^2 = 15.95
```

6.2Magnaplex

```
A = 0.85
B = 0.92
C = 0.9

# a. Analyse the system reliability, asuming only one machine at each station
.
reliability_a = 0.85*0.92*0.9 =0.7038

# b,
#reliability = (A1 or A2) AND (b1 or B2) AND (C1 or C2)
AorA = A + A - A*A = 0.9775

BorB = B + B - B*B = 0.9936

CorC = C + C - C*C = 0.99

reliability_b = AorA*BorB*CorC = 0.9615316

Improvement = reliability_b - reliability_a = 25.773%

The process improved by 25.773% from a to b.
```

6.3 Reliable Delivery

```
days = c(1:365)
reliability1 = ((190+22)/1560)*((1560-1)/1560)
dbinom1 = dbinom(days, 365, reliability1, log=F)
which.max(dbinom1) = 49

Number of days per year that has expected reliable delivery times with 20 veh
icles = 49 days

reliability2<-((190+22+3)/1560)*((1560)/1560)
dbinom2 = dbinom(days, 365, reliability2, log=F)
which.max(dbinom2) = 50

Number of days per year that has expected reliable delivery times with 21 veh
icles = 50 days
```

Conclusion

After analyzing the sales data, a definite better understanding of the business functions has been acquired and recommendations are made based on findings throughout the analysis. Some of the big findings and recommendations include:

Sweet, Food and Clothing earn very little revenue compared to the other classes. The sales of these products need to be reconsidered. Removing these products will enable the company to focus, grow and improve on some of the other better performing classes, such as Technology and Luxury.

Since Technology and Luxury is the highest earning classes, their control charts are closely examined. It is found that Luxury delivery process is out-of-control, but it is slowly moving towards in-control which is positive for the company. The delivery process for Gifts, Household and clothing are severely out of control.

It is found that increasing the Technology delivery time by three hours will reduce costs with R880 156.2.

The MANOVA analysis yielded that Class by other features such as Delivery time, Year and Price.

Overall, it is found that the companies focus need to be on Luxury and Technology

References

fch808.github.io. (n.d.). *Data Analysis with R - Exercises*. [online] Available at: http://fch808.github.io/Data-Analysis-with-R-Exercises.html.

www.sthda.com. (n.d.). *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.

cran.r-project.org. (n.d.). *Index of /web/packages/MANOVA.RM/vignettes*. [online] Available at: https://cran.r-project.org/web/packages/MANOVA.RM/vignettes/ [Accessed 20 Oct. 2022].