Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

QUALITY ASSURANCE 344

ECSA Graduate Attributes Projects

Vorster, G, Ms

23541229

Contents

List of Figures

Introduction

In this report client data for an online business is given and must be analyzed and manipulated in order to provide some feedback regarding sales. The data is wrangled to ensure that only valid instances form the given dataset will be used for calculations so that those results are accurate. In order to get a better understanding of the interpretation of given data visualization and interpretation is used. Different charts such as s-charts and x-charts are used for statistical process control. Trends and relationships between Classes, Purchase dates , clients and Product Price will be explored.

1. Data Wrangling

Before the dataset can be used, invalid instances need to be removed. The valid and invalid data needs to be separated from one other. After evaluating the data, the following observations is noted:

There are 17, missing values for the Price feature. Instances with missing values should be removed, because Missing values can bias the results of the machine learning model and/or reduce the accuracy of the model.

There are 5 negative values for the Price feature, these values are seen as invalid because for the data to be accurate price ought to be a positive value.

These invalid data will be removed in order to accurately analyze the data, which results in having 179 978 instances.

2. Descriptive features

2.1 The relationship between Age and Class

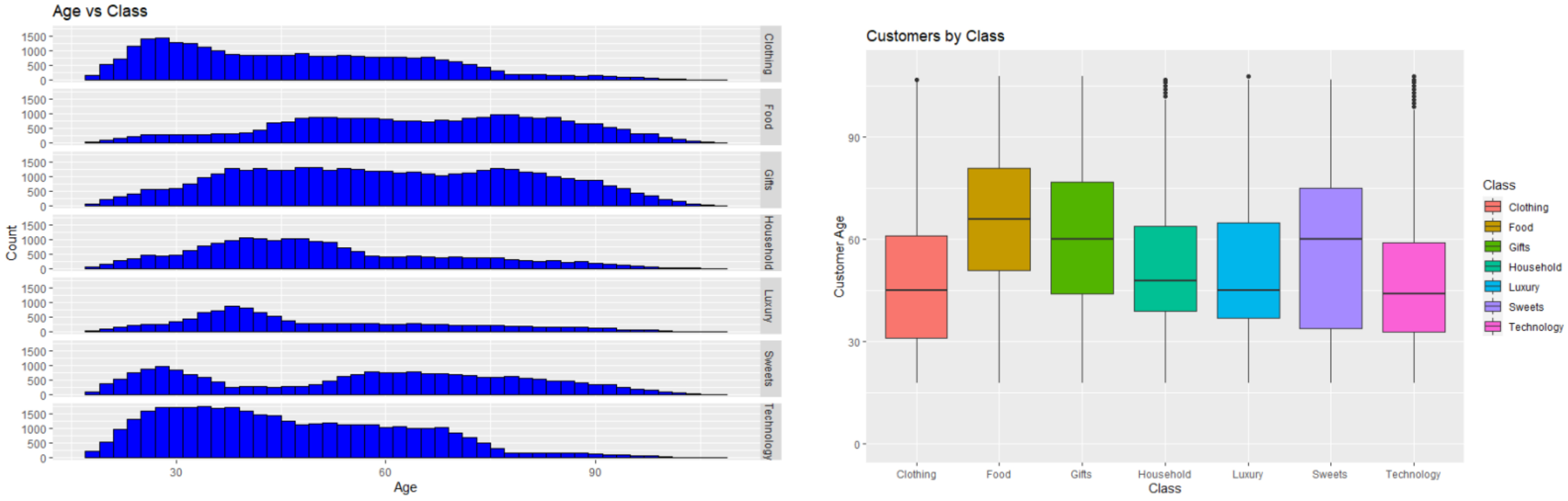


Figure 1-target market by class

The histogram presented by above as well as the boxplot it showcases the weight distribution obtaining how different aged people spend their money. This visualization of variables is important because it can help businesses to identify the target age for their products, which can come in handy when making advertisements.

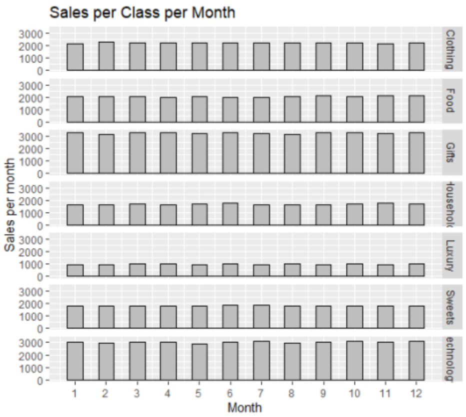
It is noticed that sweets has the largest class distribution of all, with Q1 being around 40 and Q3 at 78.

Of all the classes Technology has the lowest mean of all the classes- a suggestion to the company could be to look into advertising to indented market.

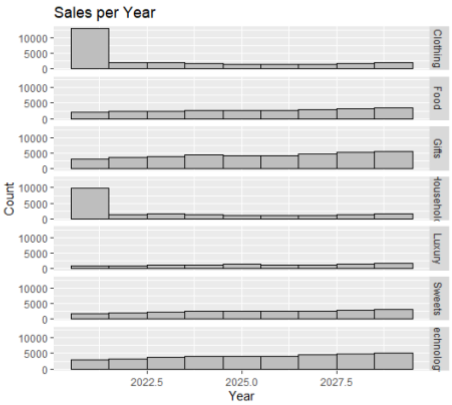
Clothing is very distributed across all ages, this makes sense since clothes are a necessity to all ages. The data can be described as uniformly distributed with a mean age of around 50 years.

Food follows a similar pattern to clothing, with a slightly higher mean.

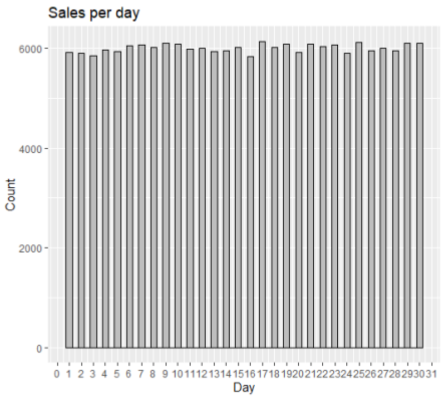
2.2 Sales



A uniform distribution is followed by the sales per month, because there isn’t much variation in sales from January through December as we can see a spike in sales in those months. There is no clear month where sales excel. A smart business suggestion could be to advertise especially for the holiday shoppers that their services is available.



According to this graph sales were at its highest in 2021 and the experienced a sudden drop following 2022 regarding clothing and household. The other classes remained uniformly distributed, but a increase in sales regarding food, gifts, luxury, sweets and technology is evident. Technology has a very positive growth and upward trend. It is recommended that the client should look into stocking more items of the classes that show a positive growth pattern



The uniform distribution presented by sales per day indicate that sales are not influenced by the day. A slight decrease is seen at day 11-16, as well as peaks at 6-10 and 17-26.

2.3 Delivery times

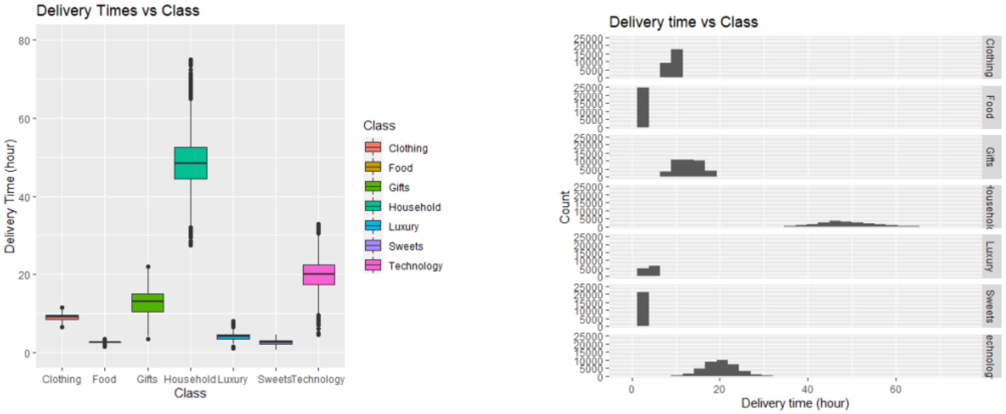
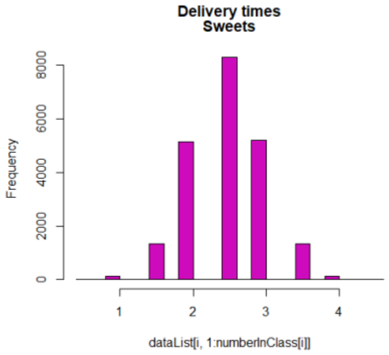


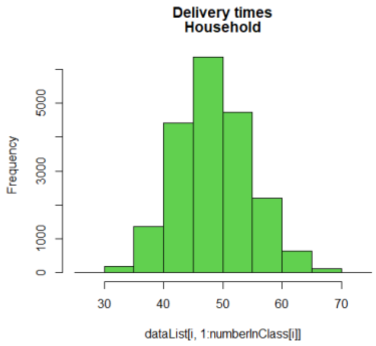
Figure 2-delivery times VS class

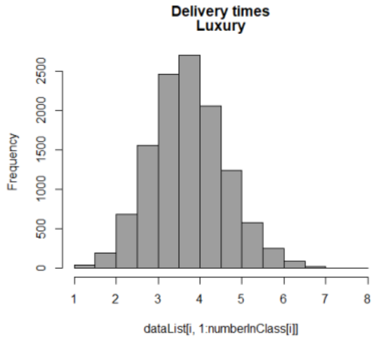
Upon observation of the histogram Household, Technology and Luxury follow normal distribution, while clothing, food and sweets does not.

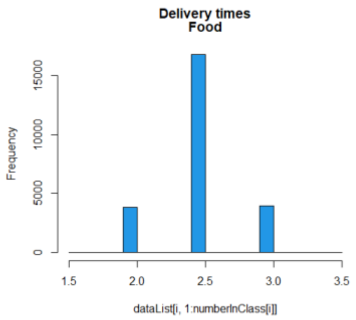
The following are some histograms that for delivery time of every class, some observations we can make is that Clothing, Food and Sweets do not have continuous distributions and deliveries. Clothing and Sweets are made on various days and food is only delivered on days 4,5,6. Normal distribution is shown for instances Household, Technology and Luxury.

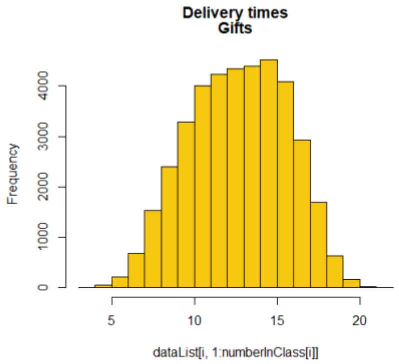
The information presented in Figure 2 is important for a company to make use of to understand the delivery times for certain classes so that they can advise a company to make sure that delivery means are available for those classes that are desired.

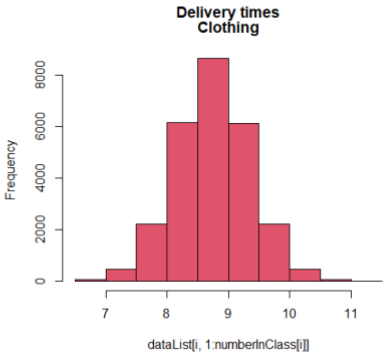


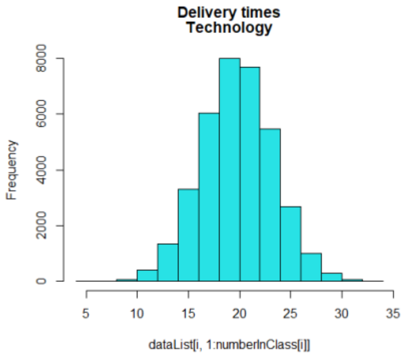












2.4 Delivery times compared to Price

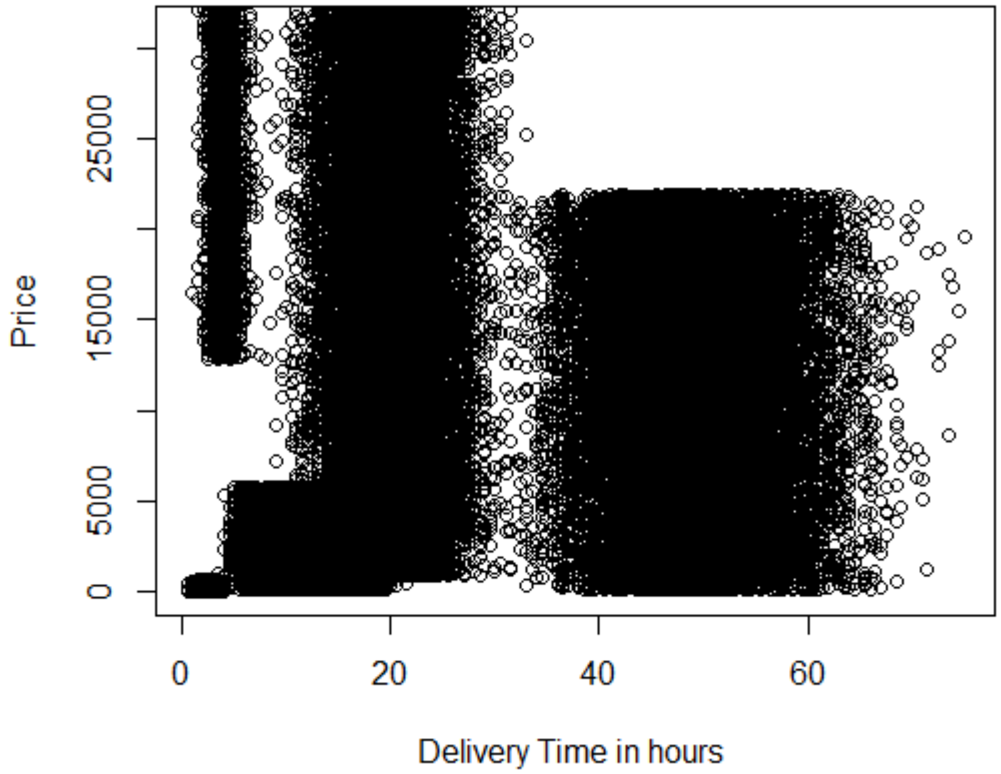
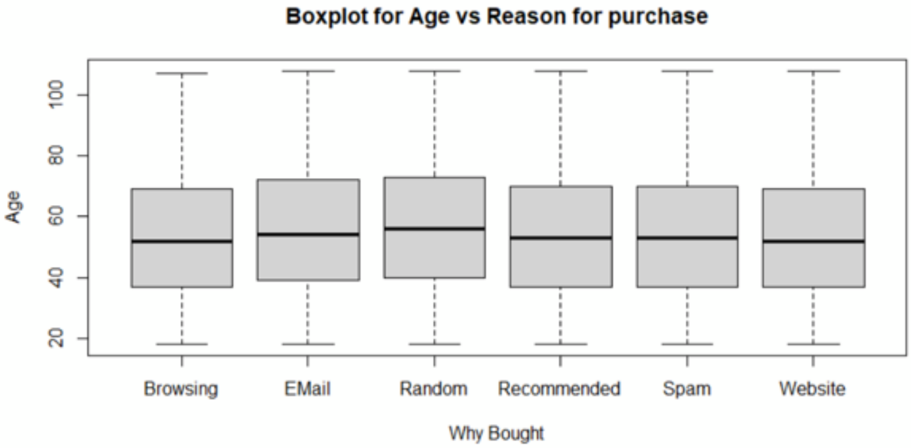


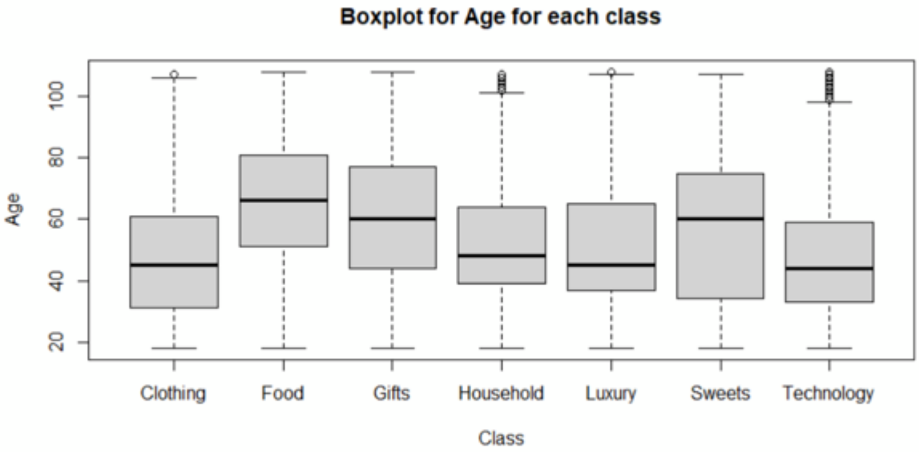
Figure 3-delivery times compared to price

As presented by Figure 3, it is evident that cost ranging higher than R15 000, the delivery time is reduced significantly. It can also be derived that the client has put emphasis on rapid delivery compared to the rest of the products except the low-cost items, that we know to be the food items that require a low delivery time as some items can perish. From hour 40-60 the price is up to about 20000.

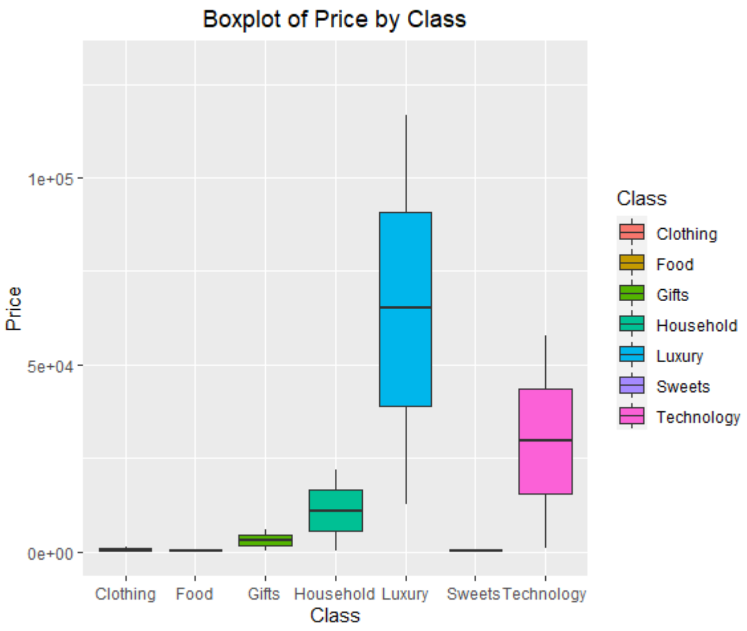
2.5 Boxplots



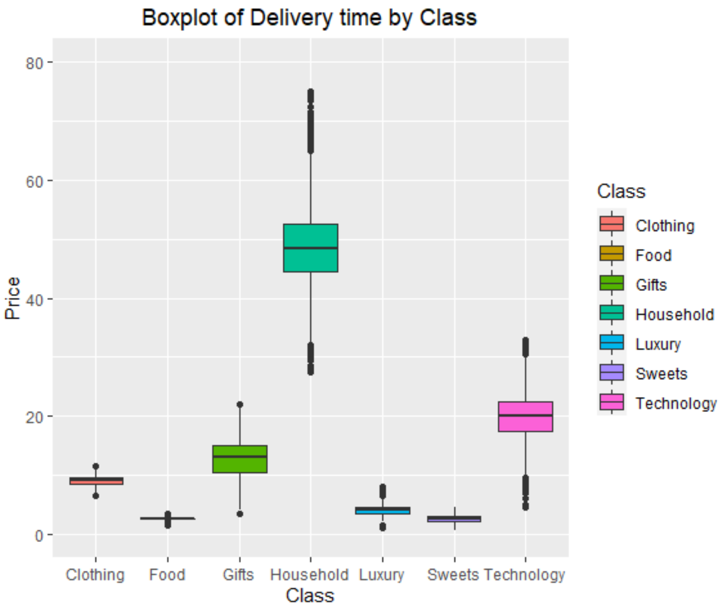
This boxplot shows that people who buys items is not affected by their age as all 6 reasons why people were buying have the same average age. The reason people bought the item is not affected by their age.



This boxplot shows what ages prefer what class of product is needed. Clothing is mostly sold to people ages 45, Gifts is to people ages 60 and Food is sold to people with ages 70 . Household and Luxury is given to people ages 40-45. Technology-ages 45 and Sweets-60.

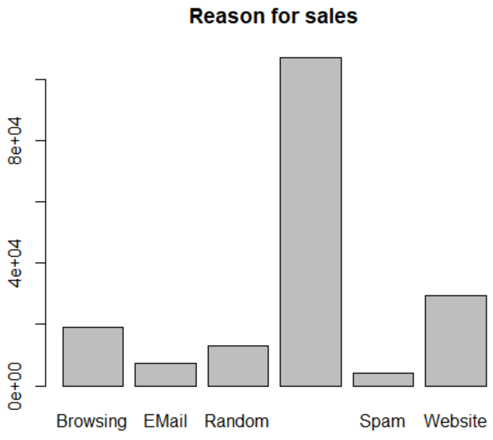


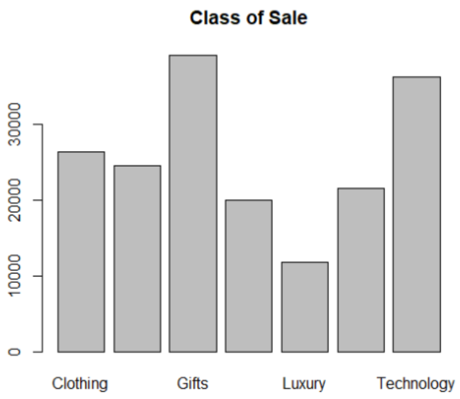
As seen above the lowest price is given to clothing, food and sweets. The luxury class is the most expensive class with a price of R60 000, followed by second most expensive class-Technology with a average of R25 000.



From the above boxplot information can be obtained that Household has the highest delivery times and that the shortage delivery time is food, sweets and luxury. It is clear that the class of the product influence the delivery time.

2.6 Reasons for Sales





2.7 Process Capabilities

Cp and Cpk calculations are used to measure process capabilities, used primarily when a process is under control. The Cp indicate whether a distribution can potentially fit inside a specification. The Cpk indicate the whether the overall average is centrally located, shows how constant you are around your average performance. (pqsystems, n.d.)

USL=24

LSL= 0

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

$Cpu = \frac{USL-X}{3\sigma}$=0.4035293

$Cpl = \frac{X-LSL}{3\sigma}$=1.880884

$Cpk = Min(Cpu, Cpl) = 0.4035293$

If the Cp values are less than 1 it means that the process isn't capable of meeting the specifications. The Cp is 1.142207 meaning that the delivery time is within specifications.

Is the LSL of 0 logical?

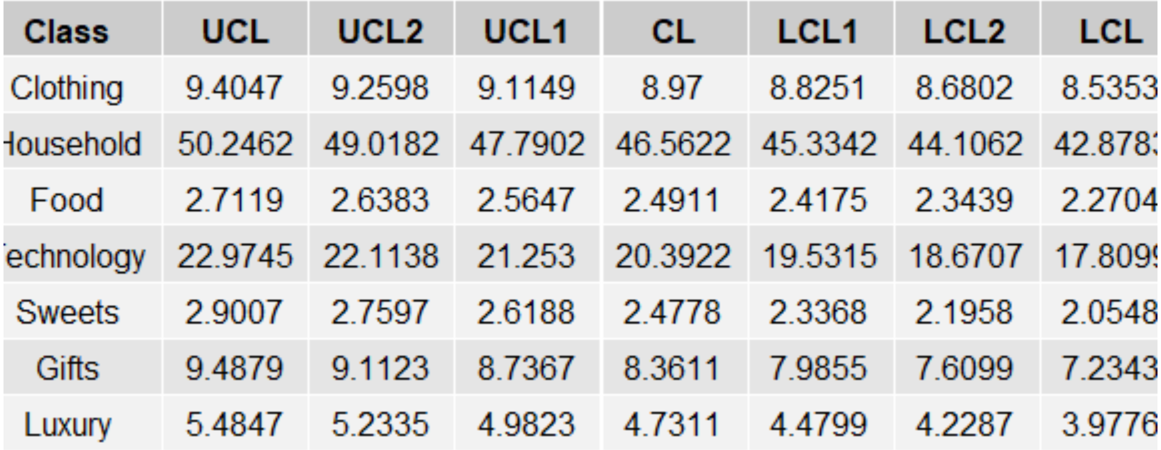
Yes, because the LSL cannot be negative - since it is impossible for the delivery time to be less than 0.

3. Statistical Process Control (SPC)

3.1 Initial 30 samples

3.1.1 Initialize the X-charts

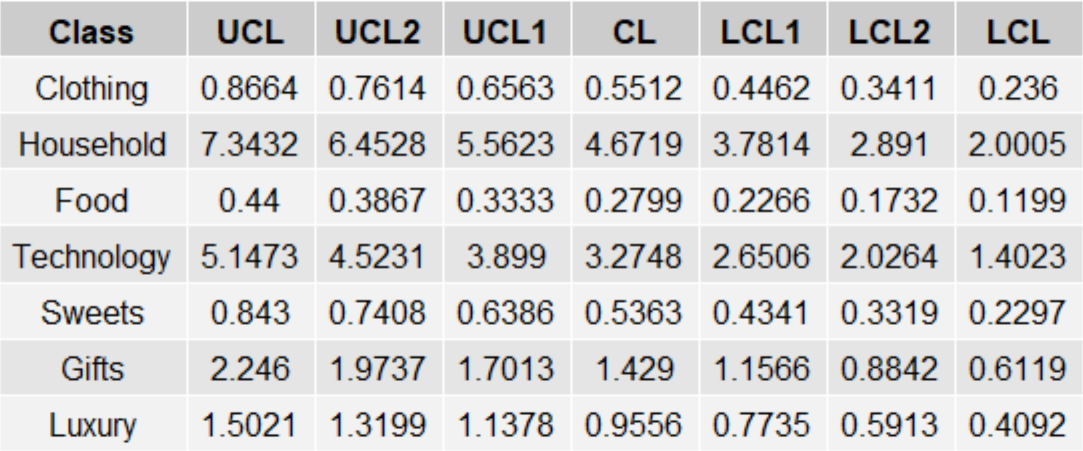
The goal of an x-chart is to specify the mean /average change of a process gradually over a period from a subgroup of values. A subgroup being measurements that is subject to the same operations. It is also used to monitor the effects of process improvement theories. (Radziwill, 2015)



Class	UCL	UCL2	UCL1	CL	LCL1	LCL2	LCL
Clothing	9.4047	9.2598	9.1149	8.97	8.8251	8.6802	8.5353
ousehold	50.2462	49.0182	47.7902	46.5622	45.3342	44.1062	42.878
Food	2.7119	2.6383	2.5647	2.4911	2.4175	2.3439	2.2704
echnology	22.9745	22.1138	21.253	20.3922	19.5315	18.6707	17.809
Sweets	2.9007	2.7597	2.6188	2.4778	2.3368	2.1958	2.0548
Gifts	9.4879	9.1123	8.7367	8.3611	7.9855	7.6099	7.2343
Luxury	5.4847	5.2335	4.9823	4.7311	4.4799	4.2287	3.9776

3.1.2 Initialize the S-charts

An S-chart is a type of control chart that is used to monitor the process variability, the standard deviation upon measuring subgroups. The standard deviation is approximated by the sample moving range. An S-chart provides a better understanding of how the subgroup data is spread out. S-charts are known to be used where there is a large sample size. (Anhoej, 2021)



Class	UCL	UCL2	UCL1	CL	LCL1	LCL2	LCL
Clothing	0.8664	0.7614	0.6563	0.5512	0.4462	0.3411	0.236
Household	7.3432	6.4528	5.5623	4.6719	3.7814	2.891	2.0005
Food	0.44	0.3867	0.3333	0.2799	0.2266	0.1732	0.1199
Technology	5.1473	4.5231	3.899	3.2748	2.6506	2.0264	1.4023
Sweets	0.843	0.7408	0.6386	0.5363	0.4341	0.3319	0.2297
Gifts	2.246	1.9737	1.7013	1.429	1.1566	0.8842	0.6119
Luxury	1.5021	1.3199	1.1378	0.9556	0.7735	0.5913	0.4092

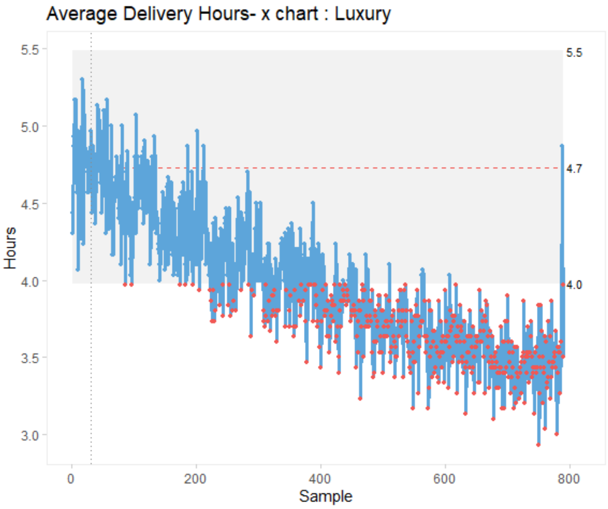
3.2 Delivery Process Control using SPC

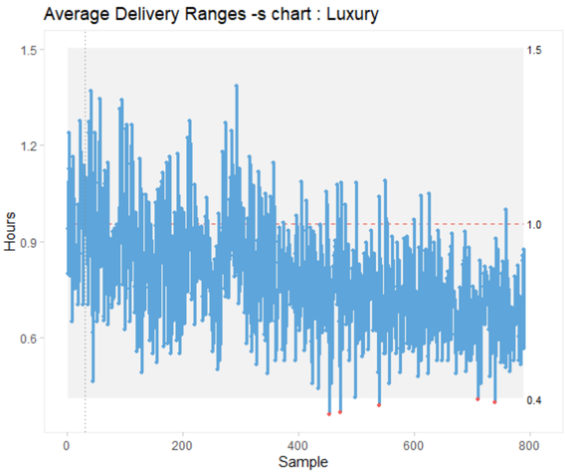
As the mean of the process is being monitored it can be categorized into out of control and in control depending on whether a point is within the limit. Upon inspecting graphs it can be concluded in what category the processes belong.

3.2.1 Out of control process (unstable)

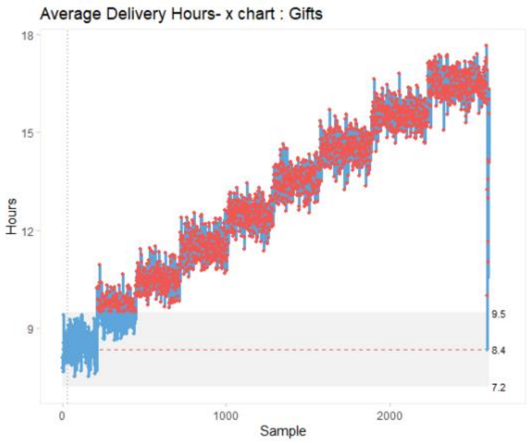
A process is deemed out of control if one or more data points fall outside the control limits - it is the in indication of the presence of non-random variation. The red dots on the charts indicate that it is out of bound. In the charts presented below it is clear that the Luxury, Gifts and Household class are out of control and is thus not a stable process. This can be an indicator that the process needs to be improved.

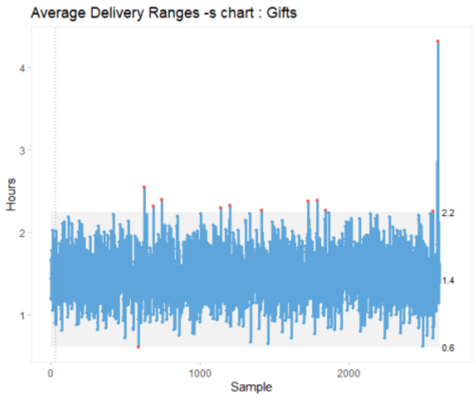
Luxury



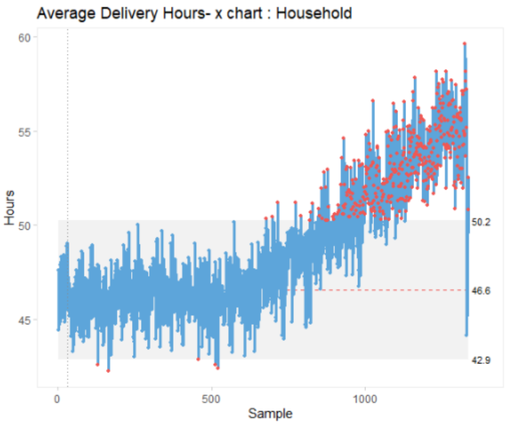


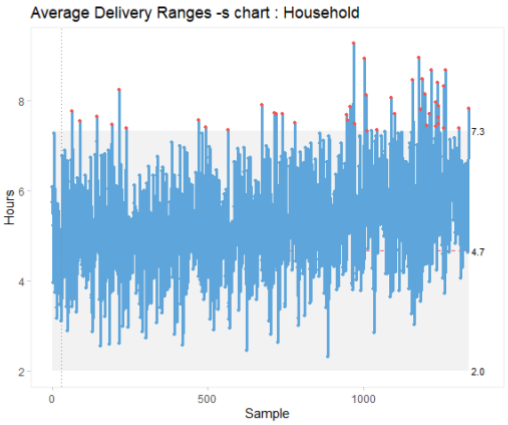
Gifts





Household



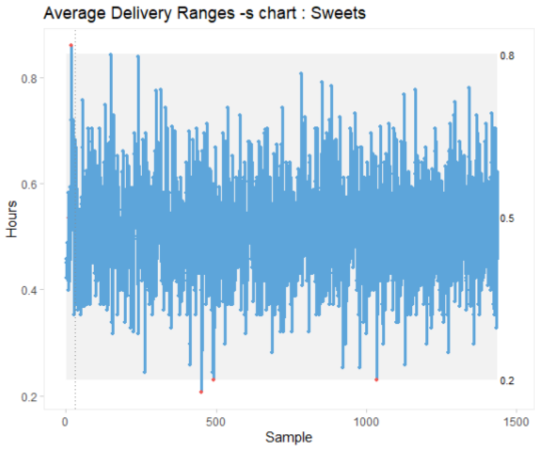


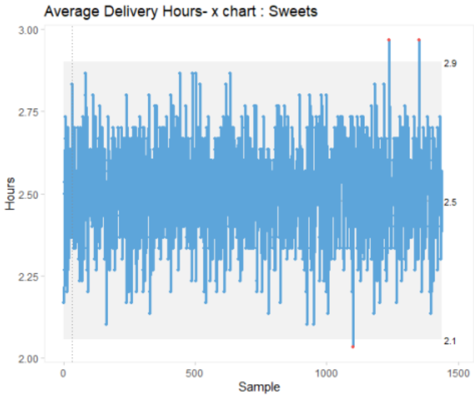
It is evident that gifts delivery time has significantly increased- this could be that the company can no longer support the amount of products that need to be shipped at a specific time.

3.2.2 In control Process (Stable)

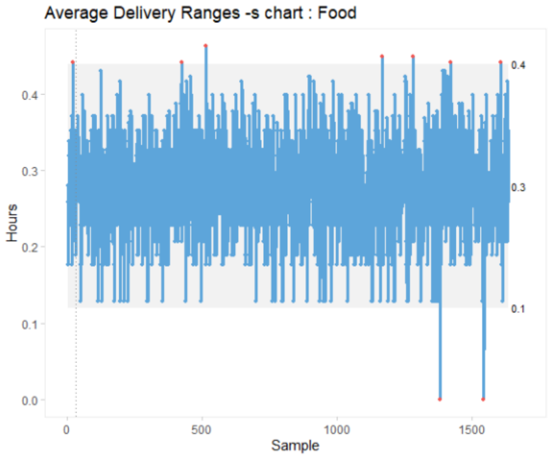
When a process is in control, most of the data points falls inside the bounds. In SPC control can be used to monitor the processes and ensure that desired quality is achieved. The process for Sweets, Food, Technology and Clothing are almost stable, regarding a few outliers that are outside of the control limits. Root cause analysis can be done to these out-of-control instances.

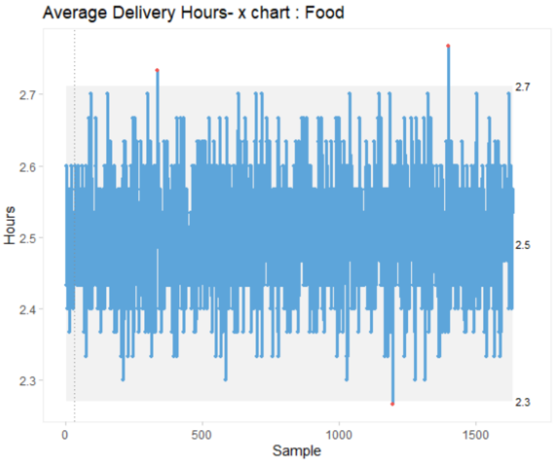
Sweets



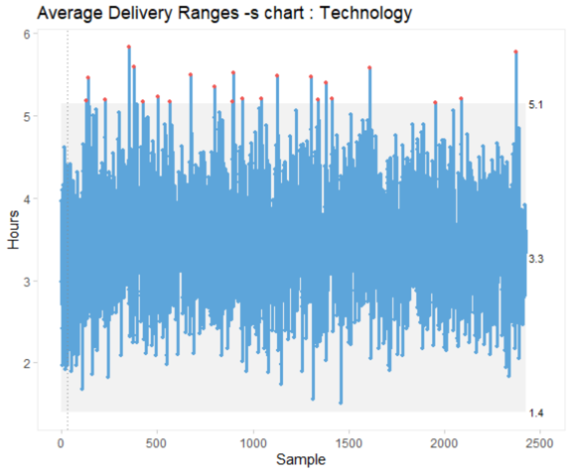


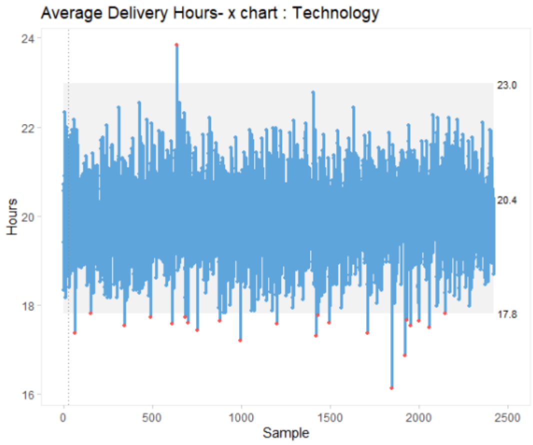
Food



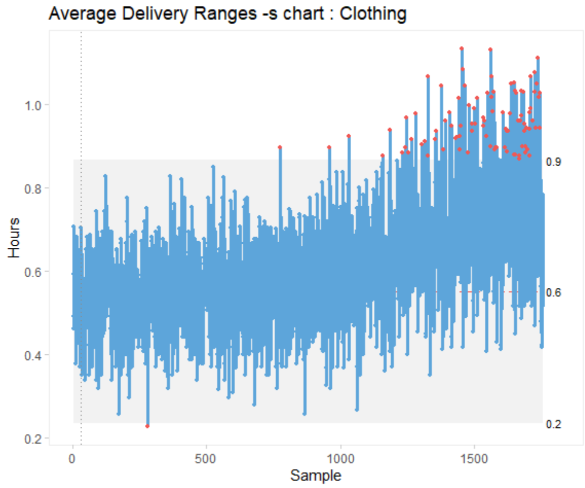


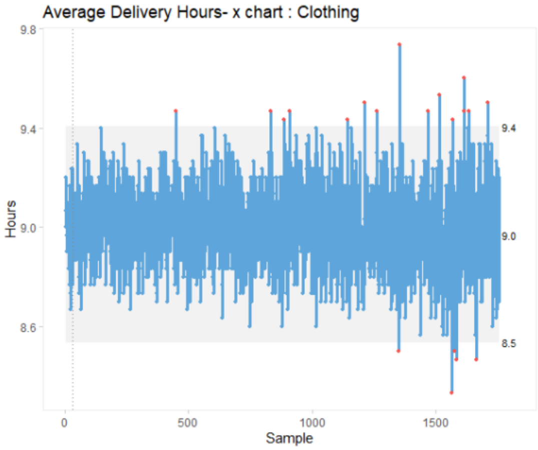
Technology



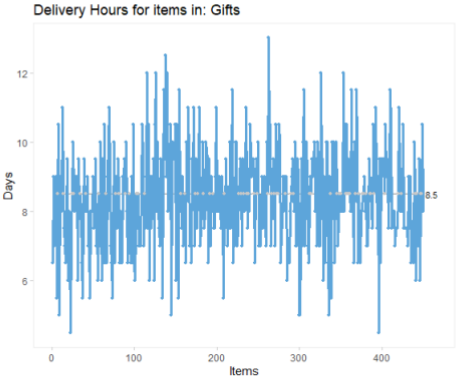


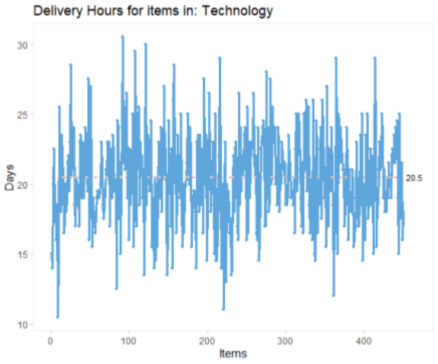
Clothing

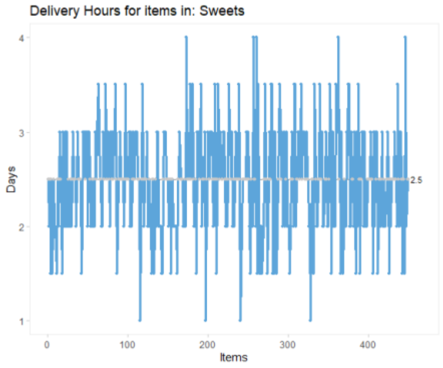


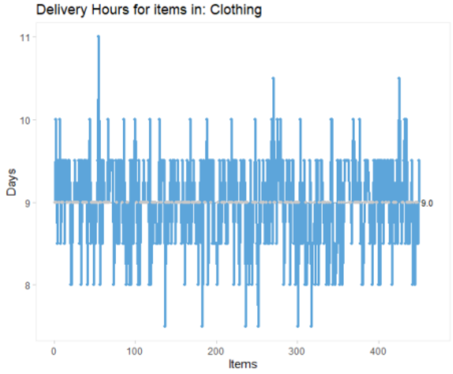


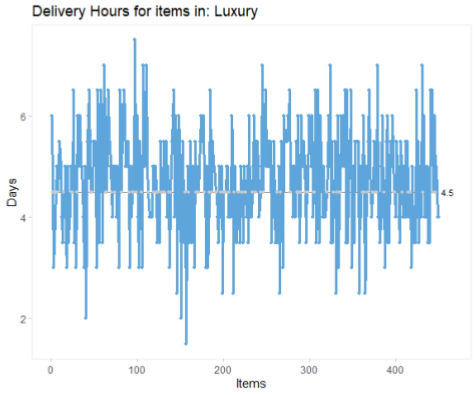
3.3 Delivery Hours for items

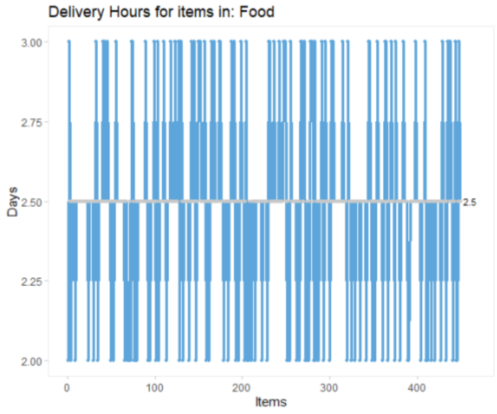


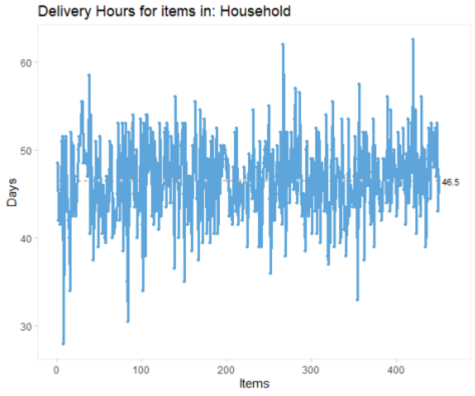






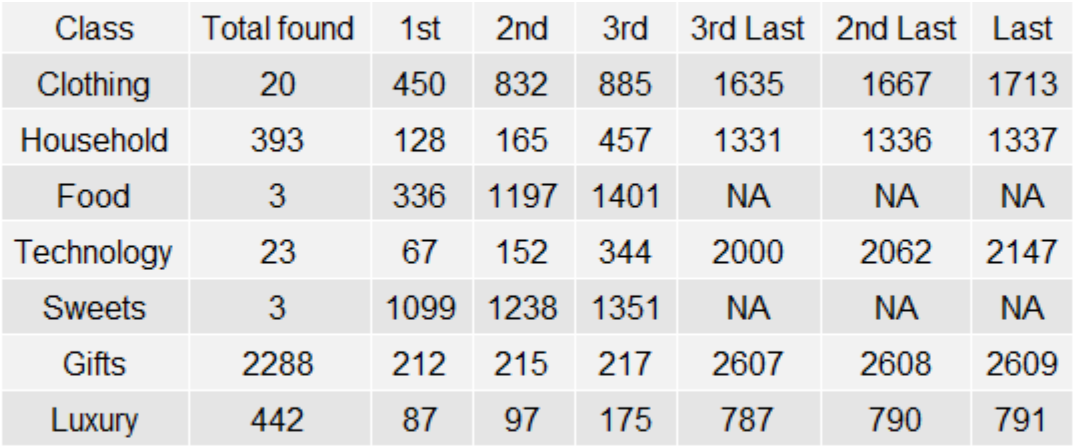






4. Optimizing the delivery processes

4.1 X-Chart analyzation

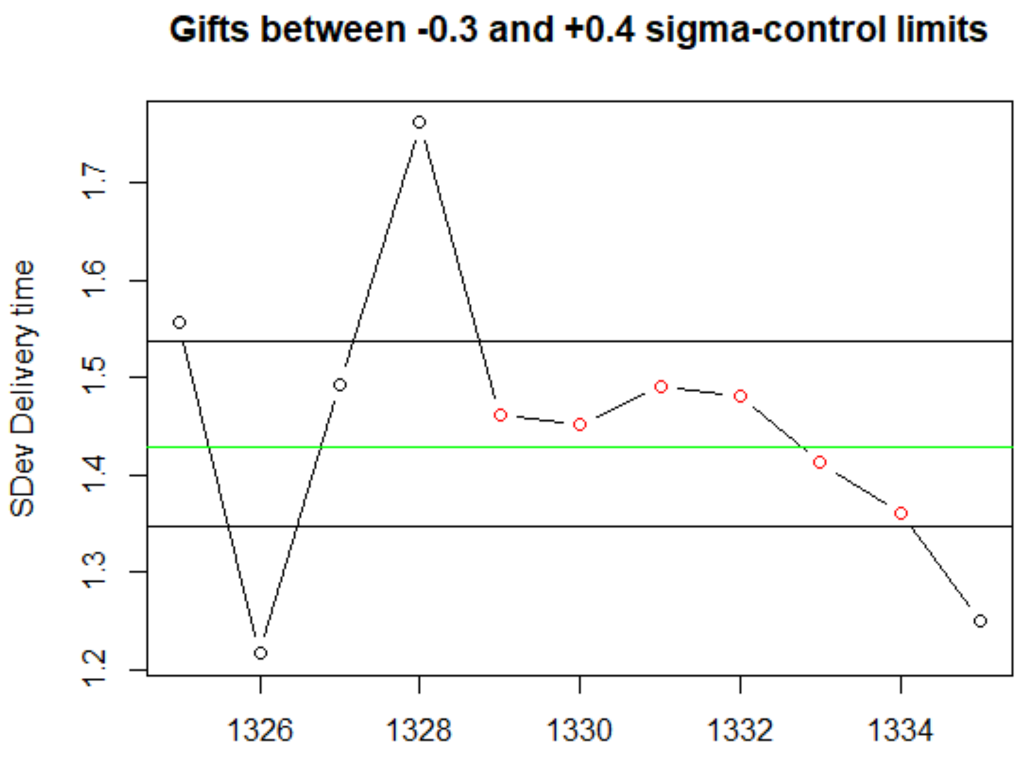


Class	Total found	1st	2nd	3rd	3rd Last	2nd Last	Last
Clothing	20	450	832	885	1635	1667	1713
Household	393	128	165	457	1331	1336	1337
Food	3	336	1197	1401	NA	NA	NA
Technology	23	67	152	344	2000	2062	2147
Sweets	3	1099	1238	1351	NA	NA	NA
Gifts	2288	212	215	217	2607	2608	2609
Luxury	442	87	97	175	787	790	791

Figure 4-A: Sample means outside the Outer Control Limits

From Figure 4 , it can be seen that Gifts has a big problem with outliers, this can affect delivery times significantly.

B: Most consecutive samples of s-bar between -0.3 and +0.4 sigma control limits

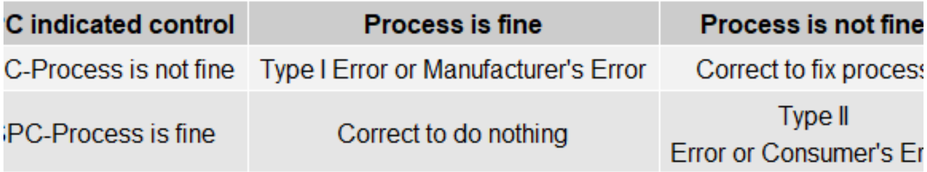


The main goal for a delivery systems to stay between -0.3 and +0.4 sigma control limits so that management can be satisfied with this accurate delivery system.

4.2 Type I Error

A type I error is the rejection of the null hypothesis when the null hypothesis is true. In this instance it can also be defined as when a process is deemed out of control. The following table showcase the probability of a type I error, regarding A and B. (Banerjee, 2009)

Rule	Probability	Probability %
A	0.00269979606326019	0.269979606326019
B	0.131659416692719	13.1659416692719



C indicated control	Process is fine	Process is not fine
C-Process is not fine	Type I Error or Manufacturer's Error	Correct to fix process
PC-Process is fine	Correct to do nothing	Type II Error or Consumer's Er

Figure 5-Type I & Type II Errors

4. 3 Optimizing the delivery process

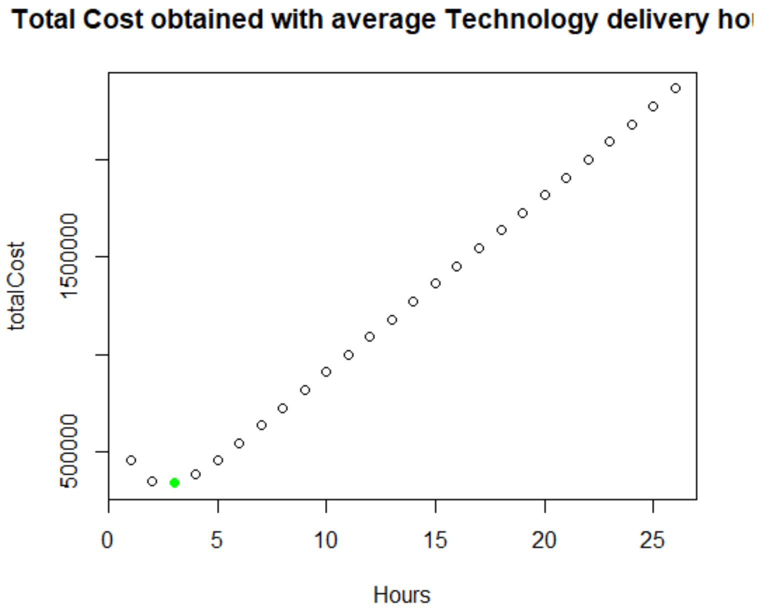


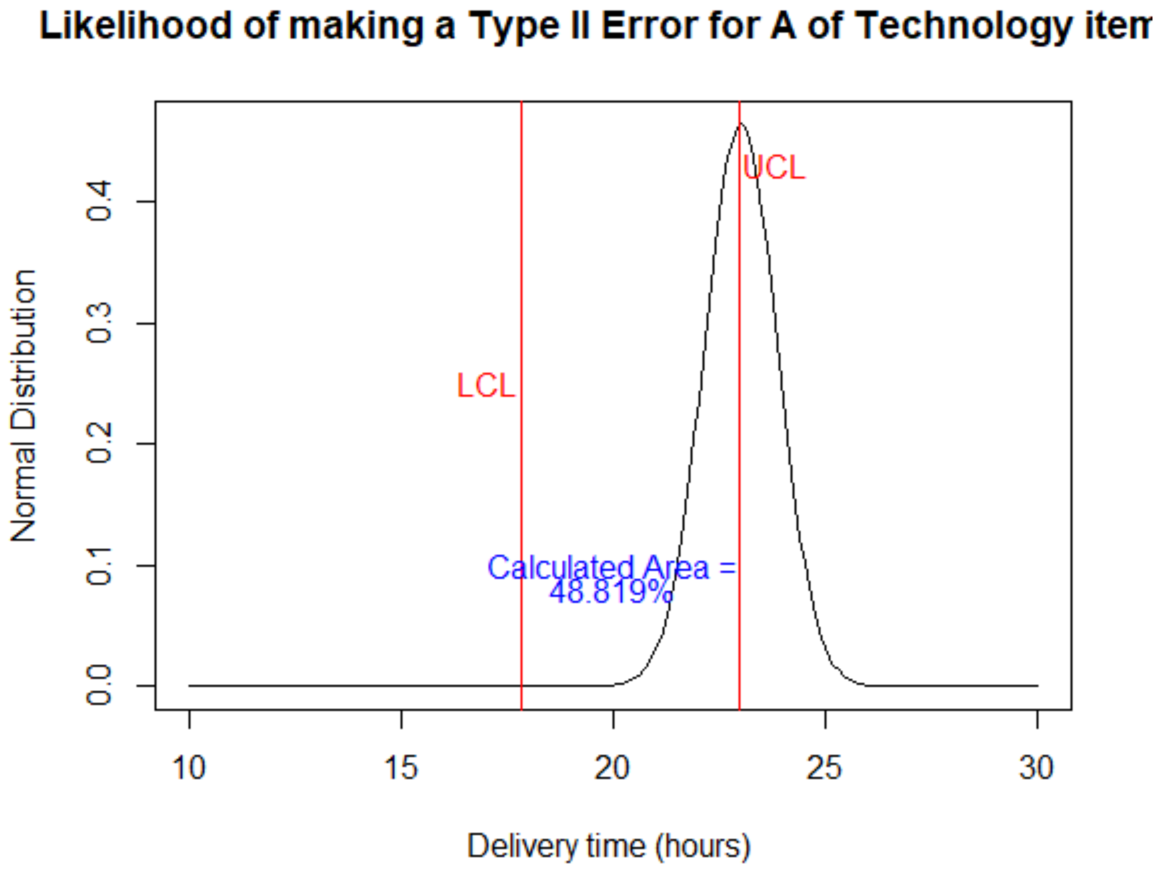
Figure 6-total cost obtained with average Technology delivery hour

Figure 6 visualize the total cost obtained with average Technology delivery hours , it can be seen that the total cost increase with hours.

4.4 Type II Error

A type II error is a term used in statistics that describes the error that occurs when one fails to reject a null hypothesis that is actually false (Hayes, 2022). The Type II error occurs when the company think that delivery will be on time, but in reality the product is being delivered late.

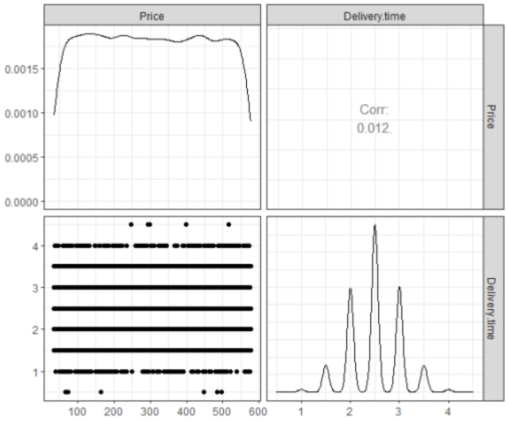
The red lines in the graph below represents the outer control limits. The likely hood of making a Type II error is calculated as 0.48819. This is also the area of the graph that falls between the outer limits.

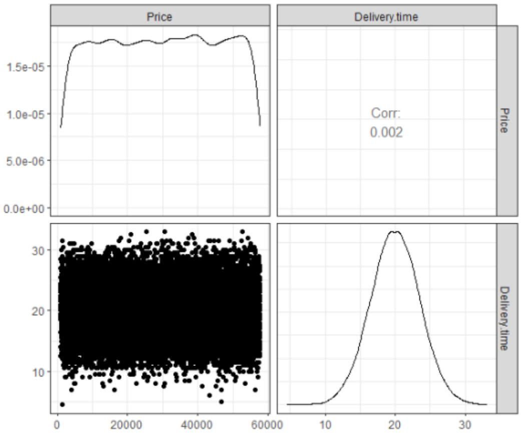


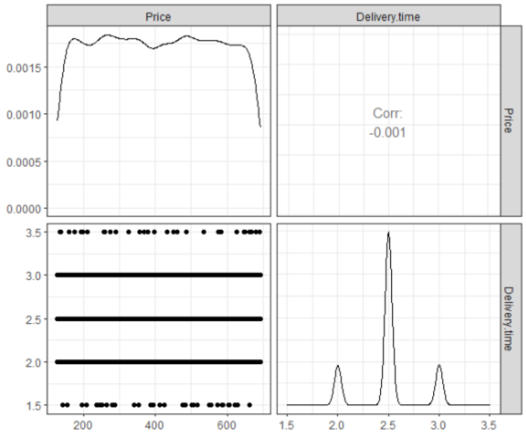
5. DOE and MANOVA

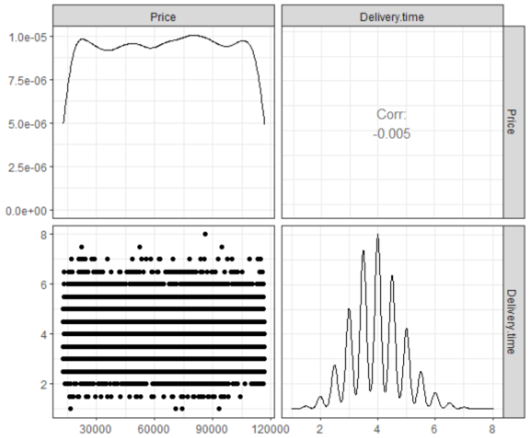
Design of experiments (DOE) is defined as a branch of applied statistics that deals with planning, conducting, analysing, and interpreting controlled tests to evaluate the factors that control the value of a parameter or group of parameters (asq, n.d.). MANOVA is used to determine whether multiple levels of independent variables on their own or combined with one another have an effect on the dependent variables.

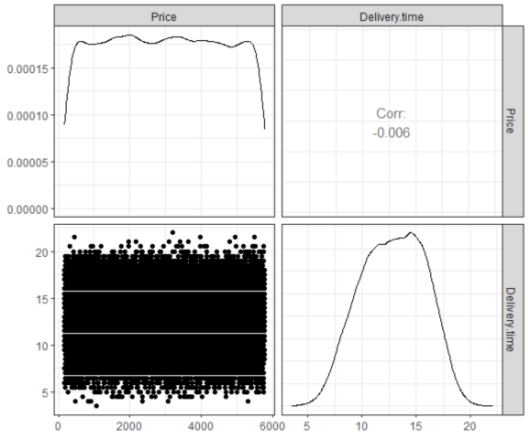
With the MANOVA test we want to determine if there is any difference in the delivery time and price between the different classes. If this is true , it will be explored how price and delivery time is effected (dependent variables). If p<0.05 reject. The linear relationship between price and delivery time is described in the graphs below.











6. Reliability of the service and products.

6.1 Problem 6 and 7

Problem 6:

$$\boldsymbol{L(x) = k(x - T)^2}$$

In order to determine k :

$$45 = k(0.06 + 0.04 - 0.06)^2$$

k=28 125

Taguchi loss function: **L(x)=28 125(x-0.06)^2**

Problem 7:

$$\boldsymbol{L(x) = k(x - T)^2}$$

Determining k:

$$35 = k(0.06 + 0.04 - 0.06)^2$$

k=21875

Taguchi loss function: **L(x)=21875(x-0.06)^2**

B) Sub 0.027 into x value

$$L(0.027) = 21875(0.027 - 0.06)^2$$
$$= \$23.82$$

6.2 Problem 27

A)

One machine at each stage

$Reliability = Reliability\ (Machine\ A) \times Reliability(Machine\ B) \times Reliability(Machine\ C)$
$Reliability = 0.85 \times 0.92 \times 0.90$
$Reliability = 0.7038$

B) Two machines at each stage

$$Reliability = Reliability(set\ A) \times Reliability(set\ B) \times Reliability(set\ C)$$

$$Reliability = (1 - (1 - 0.85)^2) \times (1 - (1 - 0.92)^2) \times (1 - (1 - 0.90)^2)$$

$$Reliability = 0.9615$$

Running the machines in parallel, will reduce amounts of breakdowns.

6.3 Days per year

The following results were obtained using Binomial and Uniroot functions in R.

Vehicles	361.5046
Drivers	0.003223914
Days per year	264.352
Vehicles and drivers	360.7471

It is important to note that drivers results may vary depending on how accurate the driver is.

Conclusion

Analyzing the data after it being wrangled to ensure no invalid data is present there are a few conclusions that can be drawn. The company's main source of revenue is obtained through Technology Products; thus the company must focus on improving the supply chain, marketing and the logistics to keep/expand this growth to Technology and better the others. The information presented can be used to optimize and improve the company's delivery process. Useful information that could be used in improvements was obtained by using the information presented in the x and s charts. The company can use the charts delivered to have optimal delivery time to maximize profit and customer satisfaction as well as calculation the probability of different types of errors.

Bibliography

Hayes, A., 2022. *Investopedia.* [Online]
Available at: https://www.investopedia.com/terms/t/type-ii-error.asp
[Accessed 18 October 2022].

pqsystems, n.d. *pqsystems.* [Online]
Available at: https://www.pqsystems.com/qualityadvisor/DataAnalysisTools/capability_4.6.3.php
[Accessed 20 October 2022].

Radziwill, N., 2015. *r-bloggers.* [Online]
Available at: https://www.r-bloggers.com/2015/11/control-charts-in-r-a-guide-to-x-barr-charts-in-the-qcc-package/
[Accessed 20 October 2022].

Anhoej, J., 2021. *r-project.* [Online]
Available at: https://cran.r-project.org/web/packages/qicharts/vignettes/controlcharts.html
[Accessed 16 October 2022].

Banerjee, A., 2009. *National Library of Medicine.* [Online]
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996198/
[Accessed 15 October 2022].

asq, n.d. *asq.* [Online]
Available at: https://asq.org/quality-resources/design-of-experiments
[Accessed 18 October 2022].