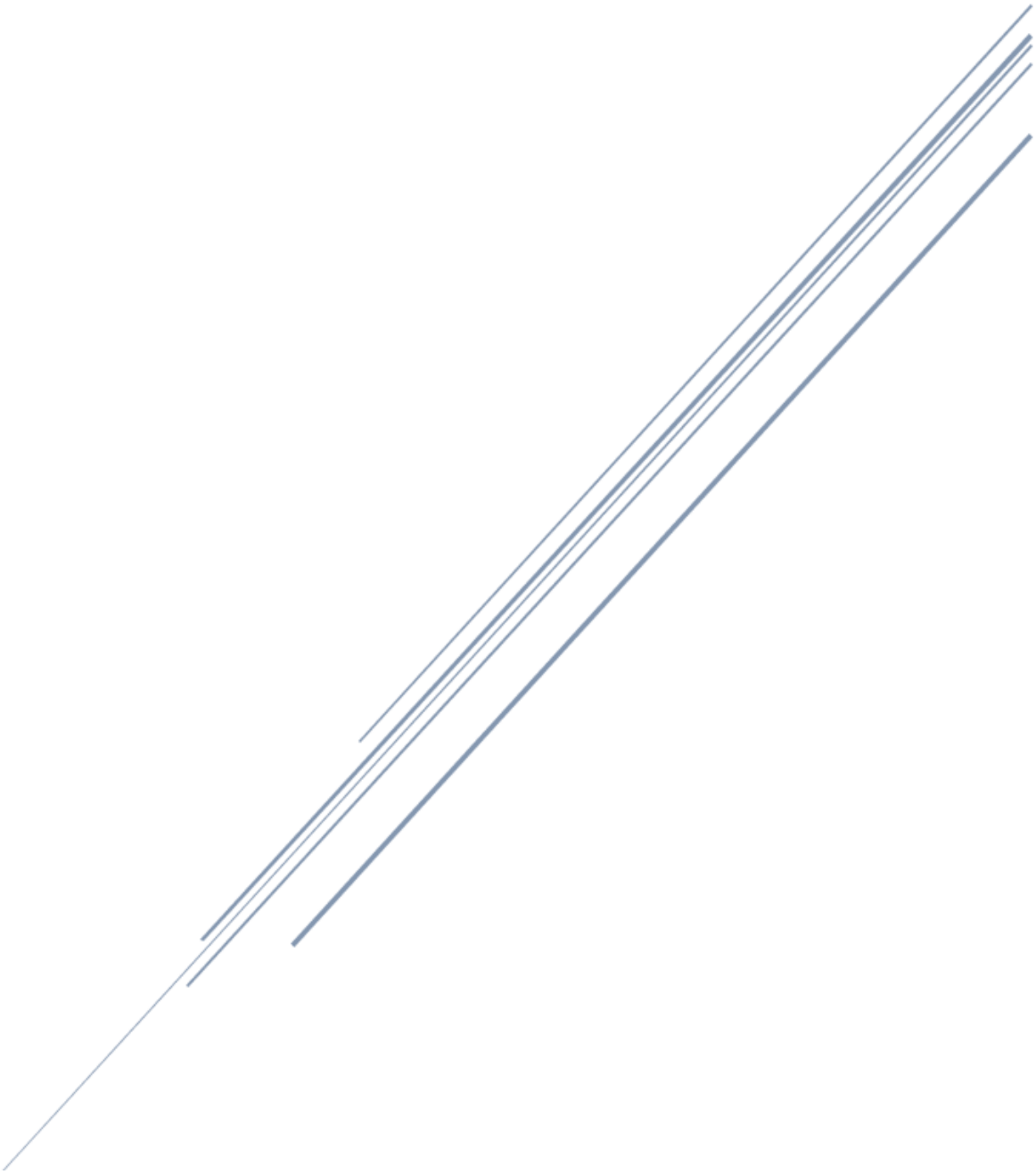
ECSA PROJECT

Kianne Lubbe



University of Stellenbosch
Quality Assurance

Abstract

The following report will analyse sales data from an online business. Data is wrangled to a set of valid data on which performance measures and calculations can be done. Different data comparisons are visually displayed to give a clear understanding of the data distribution. Statistical process control is done by the means of X- and S- charts of the different delivery times for various product classes. Optimizing the delivery times is central to the analytical study, and process control is done to determine if the processes are in control, or not.

Table of contents

Page

Abstract ..... **i**

Introduction **1**

Part One: Data Wrangling **1**

Part 2: Descriptive Statistics **1**

Data Visualization 1

Marketing Analysis 2

Operational Analysis 3

Process Capabilities 6

Statistical Process Control **7**

First 30 Samples 7

Clothing 7

Out of Control Process: 8

Food 8

Out of Control Process: Food 8

Gifts 9

Out of control process: Gifts 9

Household 9

Out of control process: Household 10

Technology 10

Out of control process: Technology 10

Sweets 11

Out of control process: Sweets 11

Luxury 11

Out of control process: Luxury 12

Part 4: 12

Samples outside control limits 12

Type I Error 15

4.3 15

Type II Error 16

Part 5: DOE and MANOVA 16

Part 6: Reliability of Services and Products 18

6.1 Lafrideradora 18

6.2 Magnaplex's process investigation 18

6.3 19

Conclusion 19

References 19

Introduction

Data from an online business selling products in seven different classes undergoes an analysis. In the report data will be cleaned and made fit for use. The goal is to provide the business with information on how it is doing and where improvement can be made.

Part One: Data Wrangling

The original dataset contained many invalid data instances and missing values. This dataset needed to be pre-processed to make it fit for use in a data analysis study.

The dataset was reduced from having 180 000 observations to having 179 978 valid data instances.

Some of the outcomes of the processed dataset is displayed in the table below.

	Count	Cardinality	Min	Q1	Mean	Median	Q3	Max	Standard Dev
Age	179978	91	18.00	38.00	54.56552	53.00	70.00	108	20.38881
Price	179978	78832	35.65	482.31	12294.09837	2259.63	15270.97	116619	20889.15025
Delivery Time	179978	148	0.50	3.00	14.50031	10.00	18.50	75	13.95578

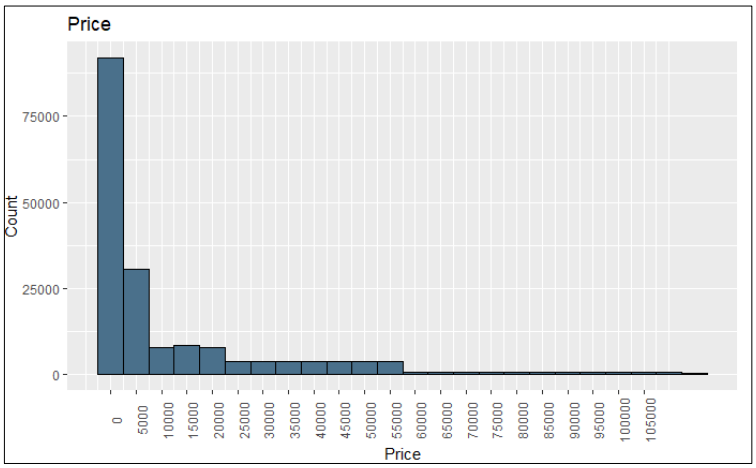
Table 1: Valid data Quality Table

The Data Quality Table reveals a significant age interval ranging from 18 to 108 years. Different age groups will have different desires, different interests, and different ways in which they are influenced to buy products. Considering the different marketing strategies, it will be beneficial to identify which strategies results in the most sales for various age groups. The marketing team will be tasked to ensure that the best way of marketing is used for each age group to attract as much individuals as possible.

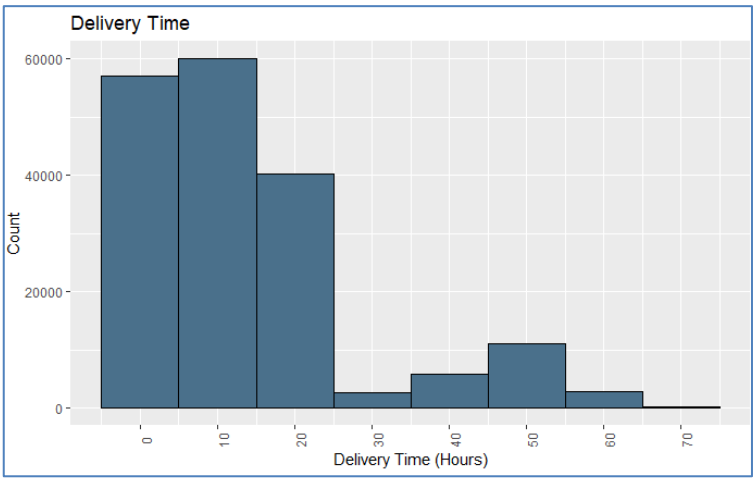
Delivery time also display a significant range. Operation teams should inspect whether all operations on various products is done sufficiently with optimal use of all production lines. Products with different levels of complexity will have longer times to delivery, and this will need to be measured against the cost benefit that the specific product has. Lead time can be evaluated against production intervals, safety stock, and other unique contracted promises.

Part 2: Descriptive Statistics

Data Visualization



Price follows a unimodal skewed right distribution. Most of the products have a low price, and sales above R55 000 are extremely low. Customer demand for lower priced products is much higher.



Delivery Time has a distribution skewed to the right with the majority of delivery times being short.

Marketing Analysis

For marketing analysis, it is important to know which marketing strategy is resulting in the greatest number of sales being made to customers.

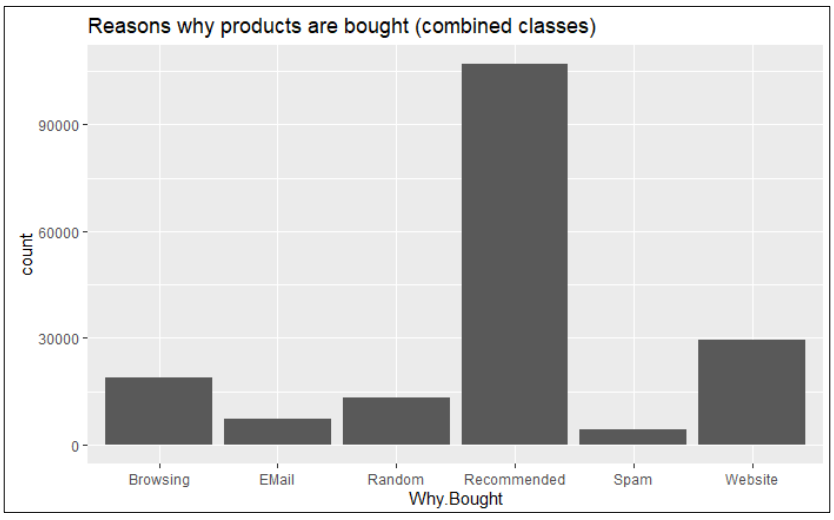


Figure 1:Graph diplaying Why Bought Count

From the results obtained by the graphs most sales were made due to recommendations. This, being a good indication that the reputation of the company is upholding a high standard further encourages the company to maintain their standards of customer relationships by delivering to customers according to expectation.

The company can improve their online platforms to optimize the sales performance resulting from it.

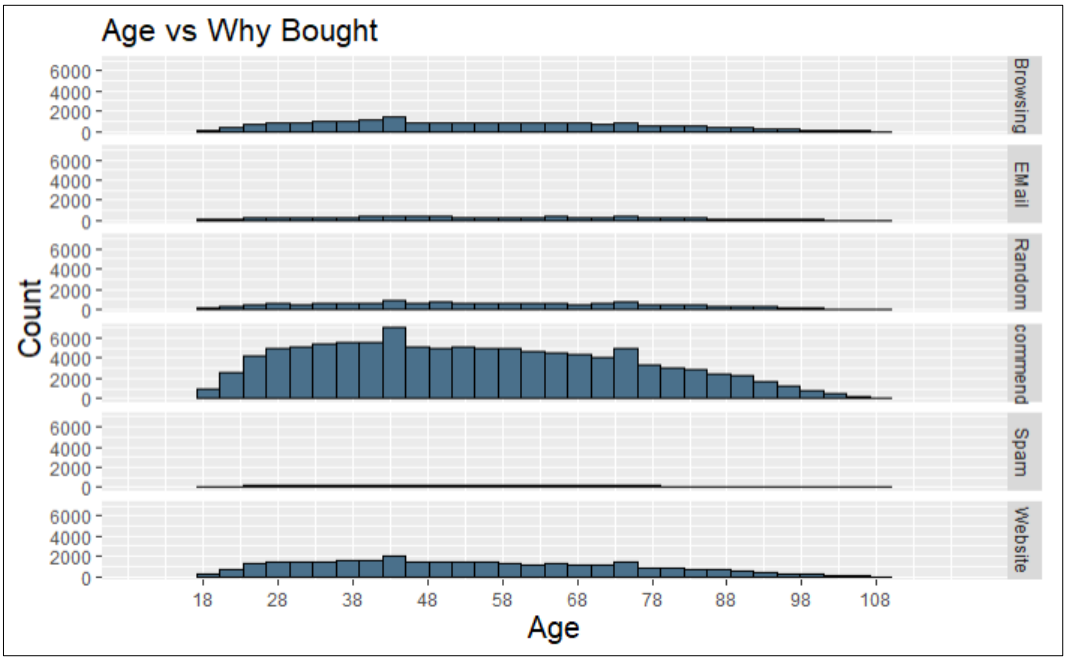


Figure 2: Graph displaying Age vs Why Bought

The distribution is skew to the right with most purchases being made by customers below the age of 58. Reasons can be because of the customer familiarity with online platforms. Older customers have less technological education than younger generations. Marketing strategies to ease the access to the business platform can be investigated.

Operational Analysis

Operational managers should be aware of the number of operations to perform on specific product classes. The demand per class will influence the production levels and the required operation hours per product.

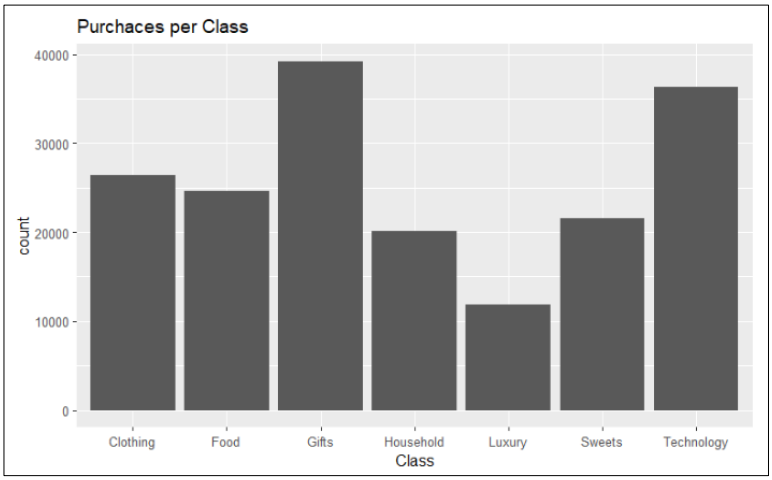
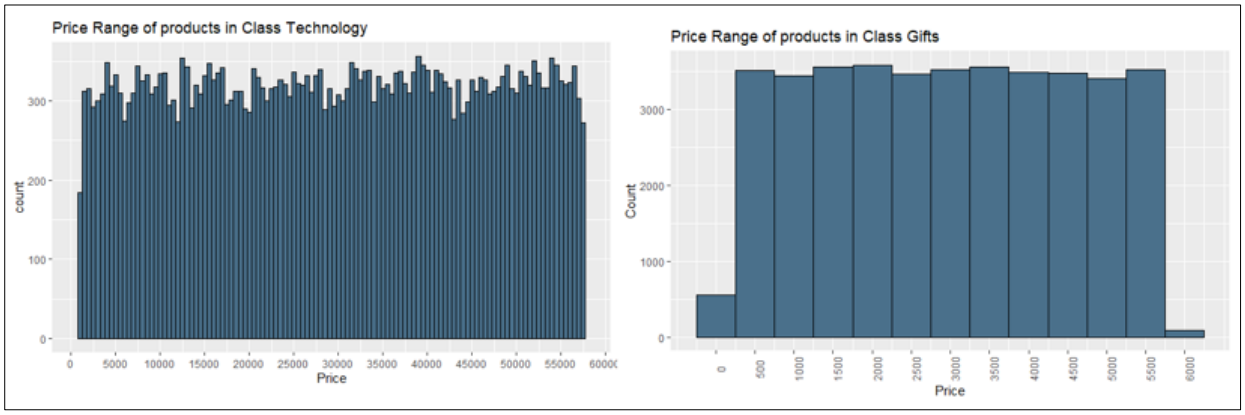


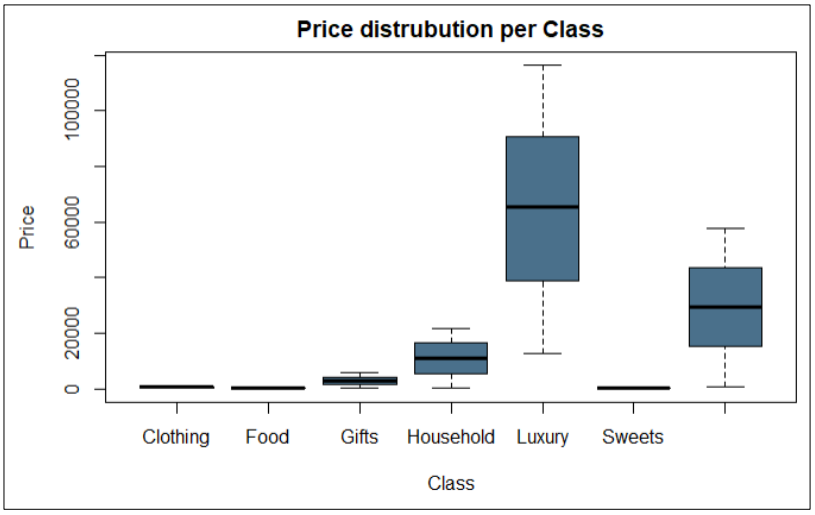
Figure 3: Purchase count per Class

The Purchases per class graph displays that the classes 'Gifts' and 'Technology' has the highest number of purchases, and therefor will have the highest predicted demand.

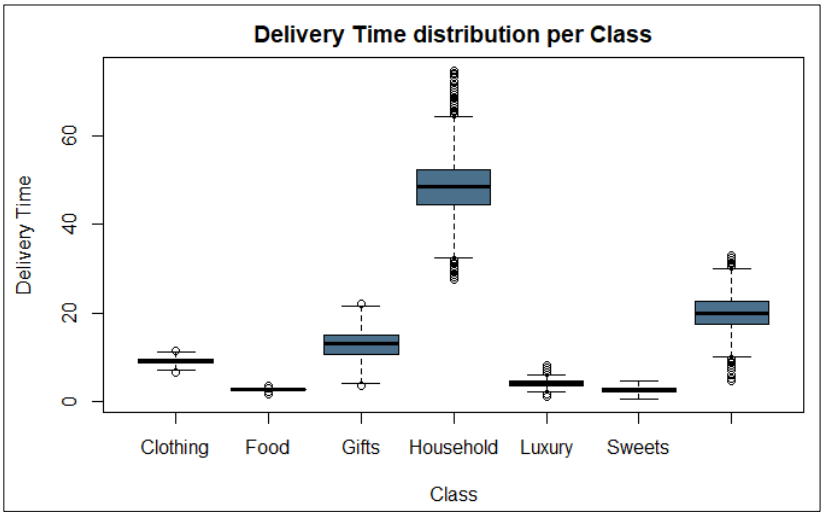
The value of the products being sold, the peak seasons in which demand is required, and the delivery time of products are important to consider when doing an overview of the business.



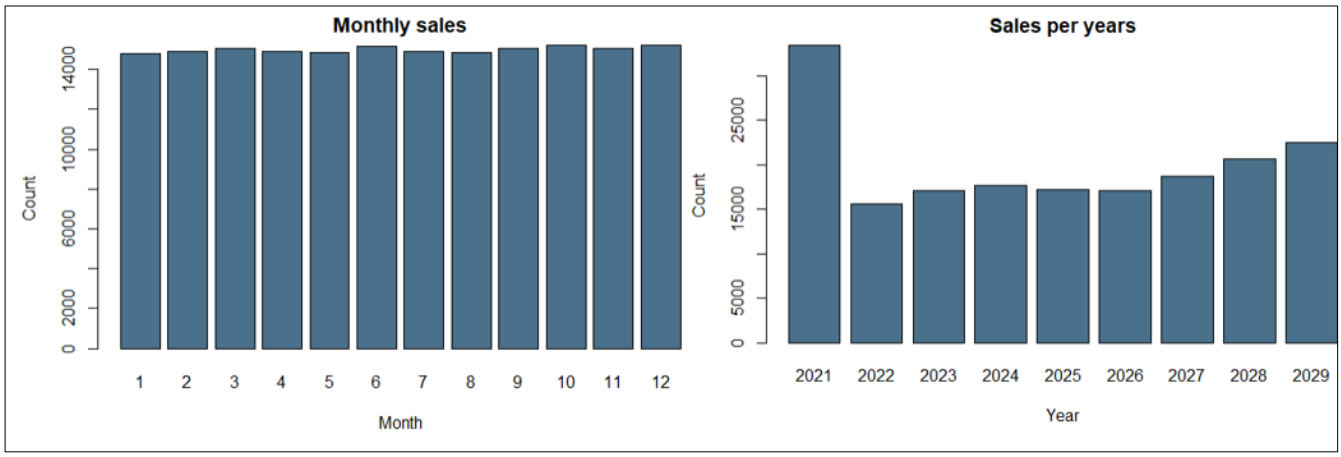
The price range of class 'Technology' items is much higher than that of 'Gifts'. Items with a high monetary value will require a more reliable delivery service, safe storage, and thorough quality control. Operations in time-pressured periods will be assigned to classes with higher priority based on the cost benefit to the company.



The price range are influenced by the different classes. Luxury products has the highest price with an average of R60 000, compared to clothing, sweets and food products with a much lower price average. There are major price differences between the different classes. This will result in different values associated with the classes.



From the result of the plot above it is observed that class 'Household' items have the longest delivery time. Food, luxury, and sweets has the shortest delivery time. It is evident that the class of the product influences the delivery time.



Sales are distributed evenly throughout the year. Sales per month has a uniform distribution. This forces the company to have staff operating during the entire course of the year. Planning for holiday seasons will need to be done in advance to ensure that sufficient productivity is available to keep customers happy.

According to the graph displaying the sales per year it shows that sales were the highest in 2021 with a sudden drop to 2022 and a gradual increase over the years from 2022 to 2029.

Process Capabilities

USL = 24

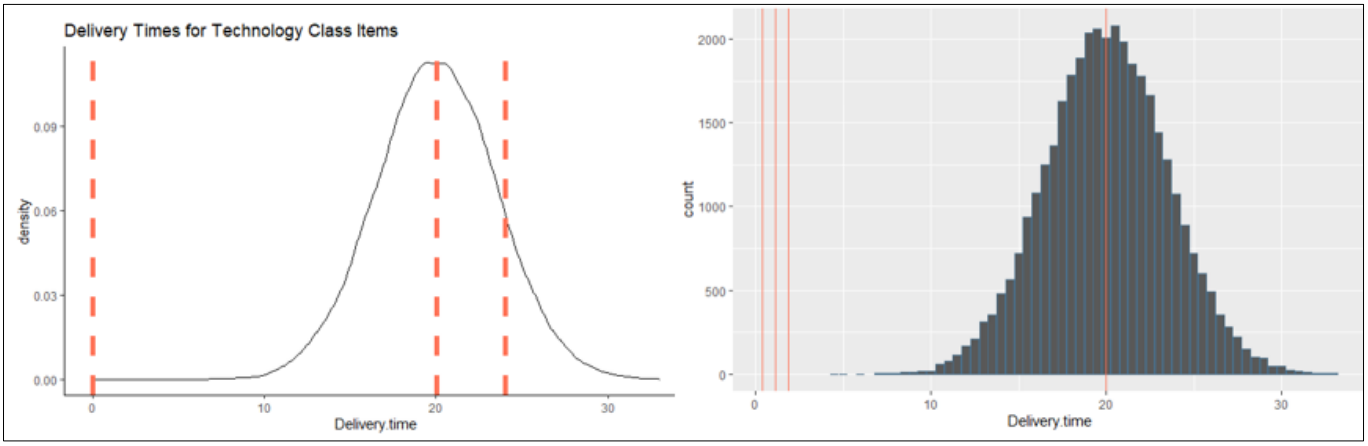
LSL = 0

$$Cp = \frac{USL - LSL}{6\sigma} = \frac{24 - 0}{6 \times 3.501993} = 1.142207$$

$$Cpk = min(\frac{USL - mean}{3\sigma}, \frac{USL - LSL}{3\sigma}) = 0.3796933$$

$$CPU = \frac{USL - mean}{3\sigma} = 0.3796933$$

$$CPL = \frac{mean - LSL}{3\sigma} = 1.90472$$



The delivery time for Technology displays a normal distribution with a mean of 20 hours, and Upper specification limit of 24 hours and a Lower Specification Limit of 0. A LCL is a logical choice, because the delivery time of any product cannot be less than 0.

Cpk is a predictor of the behaviour of the process. The Cpk value of 0.3796933 is low, and the company should aim to get it above 1. A high Cpk value will result in producing fewer defective parts and an improved product performance. The big difference between the Cp and Cpk values also indicates that the process will potentially not entirely fit into specifications. (*Pqsystems*)

Statistical Process Control

First 30 Samples

X Charts are used to indicate mean changes of a process over a period of time.

Class <chr>	UCL <dbl>	U2Sigma <dbl>	U1Sigma <dbl>	CL <dbl>	L1Sigma <dbl>	L2Sigma <dbl>	LCL <dbl>
Technology	22.875070	22.022269	21.169468	20.316667	19.463866	18.611065	17.758264
Clothing	9.405637	9.260355	9.115074	8.969792	8.824510	8.679228	8.533946
Household	50.221298	49.006212	47.791127	46.576042	45.360956	44.145871	42.930786
Luxury	5.459449	5.218452	4.977455	4.736458	4.495461	4.254465	4.013468
Food	2.706814	2.634404	2.561993	2.489583	2.417173	2.344763	2.272353
Gifts	9.483051	9.113353	8.743656	8.373958	8.004261	7.634563	7.264866
Sweets	2.873301	2.741325	2.609348	2.477371	2.345394	2.213417	2.081440

Table 2: X-Chart results

S Charts indicate the standard deviation of a process over a period of time.

Class <chr>	UCL <dbl>	U2Sigma <dbl>	U1Sigma <dbl>	CL.S <dbl>	L1Sigma <dbl>	L2Sigma <dbl>	LCL <dbl>
Technology	5.0324990	4.4358625	3.8392260	20.316667	2.6459531	2.0493166	1.4526801
Clothing	0.8573288	0.7556867	0.6540446	8.969792	0.4507605	0.3491184	0.2474763
Household	7.2627912	6.3818929	5.5009946	46.576042	3.7391979	2.8582995	1.9774012
Luxury	1.4706176	1.2962661	1.1219145	4.736458	0.7732113	0.5988598	0.4245082
Food	0.4273021	0.3766426	0.3259831	2.489583	0.2246640	0.1740045	2.2723531
Gifts	2.2559786	1.9885172	1.7210558	8.373958	1.1861331	0.9186717	0.6512103
Sweets	0.8053533	0.7098733	0.6143932	2.477371	0.4234332	0.3279532	0.2324731

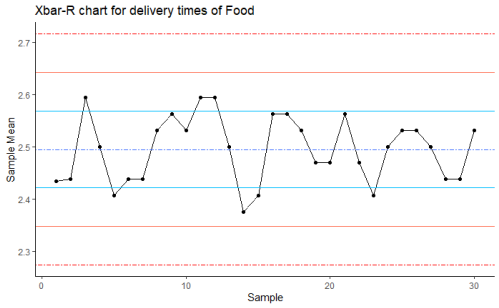
Table 3: S-Chart results

Clothing

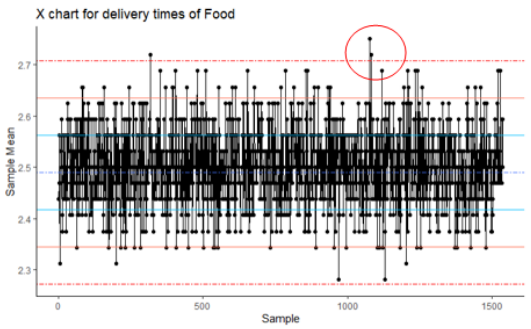
Out of Control Process:

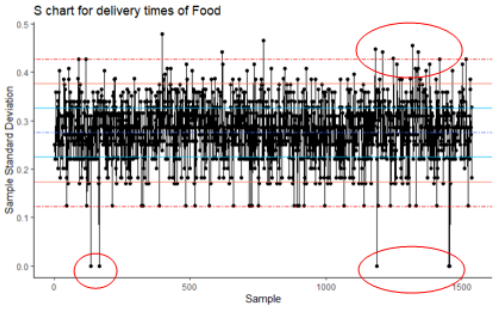
Both charts for the first 30 samples of clothing are within control limits and does not display any concerning deviations. Slight out of control signals are displayed on the X-chat for clothing samples but are overall a stable process. The S-Chart for clothing displays an unwanted deviation from the centreline going over the 1000 Sample line. There are consecutive increasing points, and it is identified as an Out-Of-Control Signal.

Food



Out of Control Process: Food

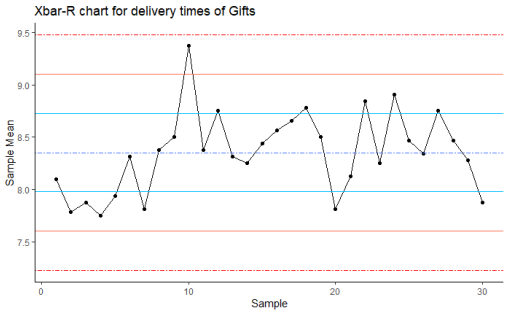


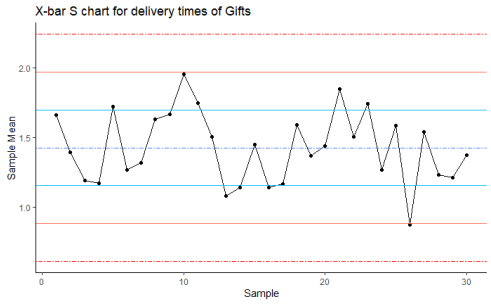


The X-chart of 30 samples for class Food products shows exceptional results within specification limits. Looking at the models for all the samples there are some out of control signals where two out of three samples are above the UCL on the X-chart. The S-chart also displays too many samples above the UCL and LCL. This is concerning because many delivery times are out of specification. A

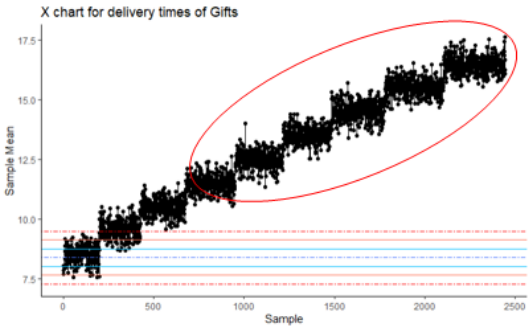
very big number of delivery times are above the 2-sigma line, also revealing some possible problems.

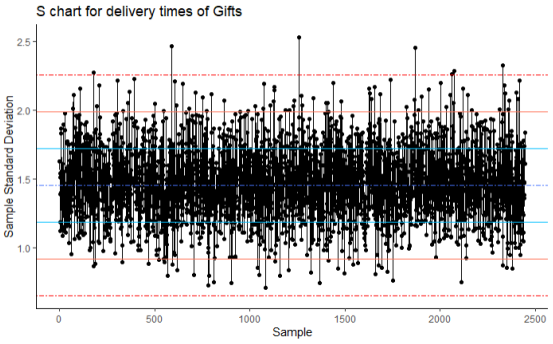
Gifts





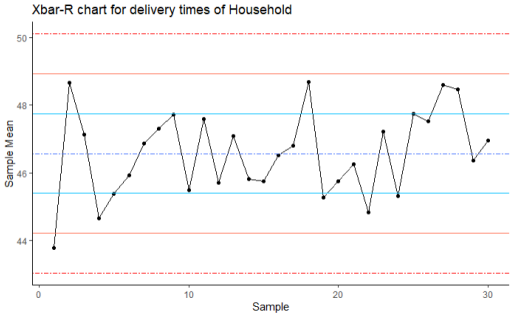
Out of control process: Gifts

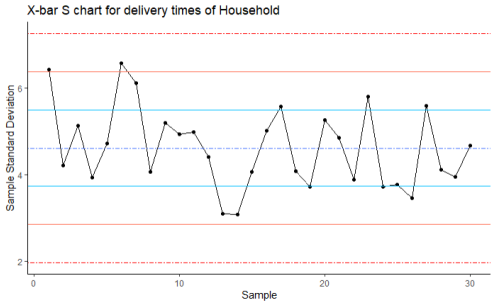




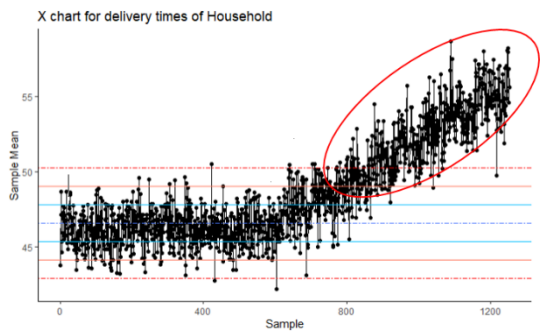
Clear out of control signals are being displayed for products in the class ‘Gifts’. There is an incline deviating majorly from the centre line. Gifts would be classified as a bad product.

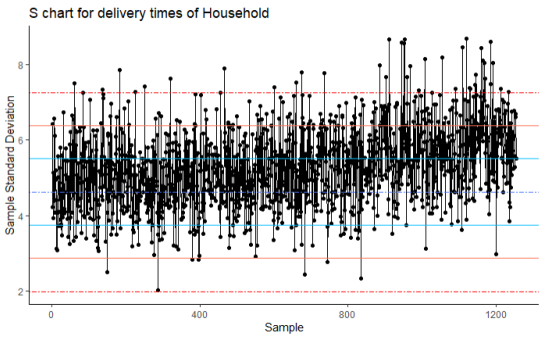
Household





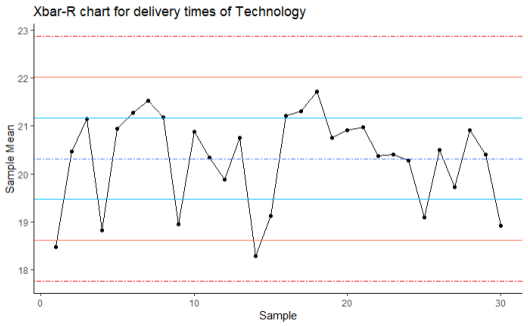
Out of control process: Household

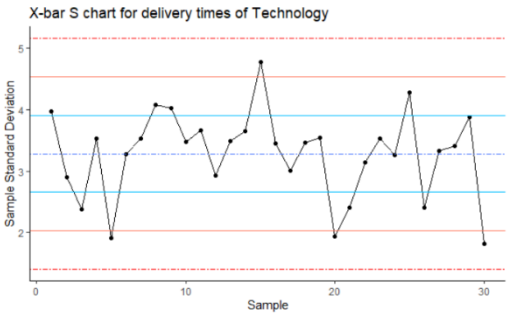




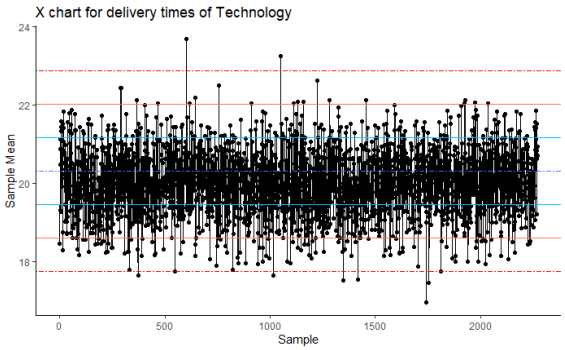
As expected from previous observations the delivery times for Household items are displaying out of control signals with delivery times being extremely high. Re-evaluation on operation processes and deliveries should be done to reduce the delivery times. Household would be classified as a 'Bad' product.

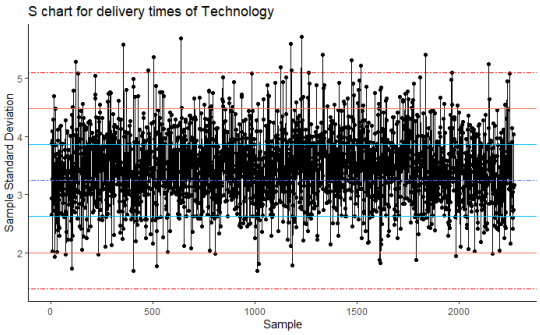
Technology





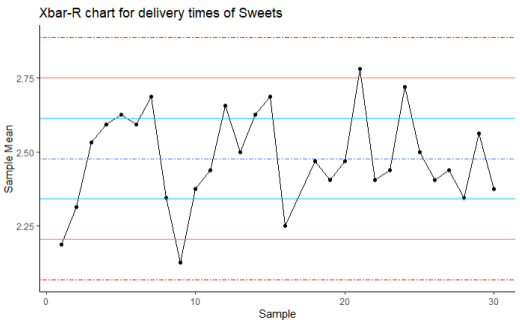
Out of control process: Technology

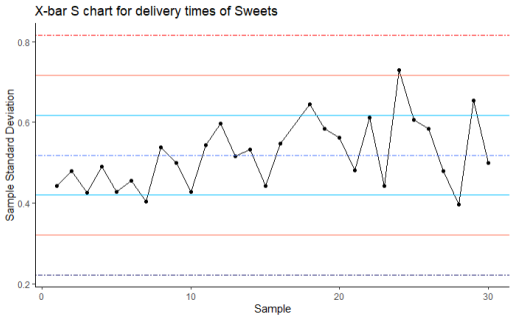




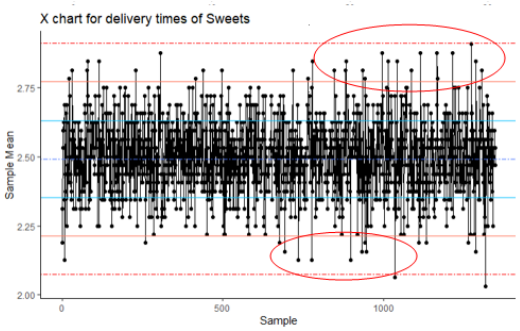
The delivery times of technology is well within specification limits, with a few deviations and outliers. The X-chart reveals a mostly stable system with some instances out of the control limits.

Sweets



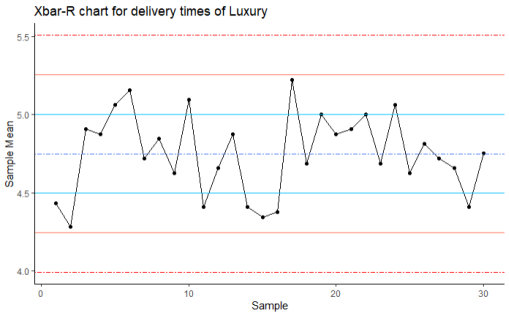


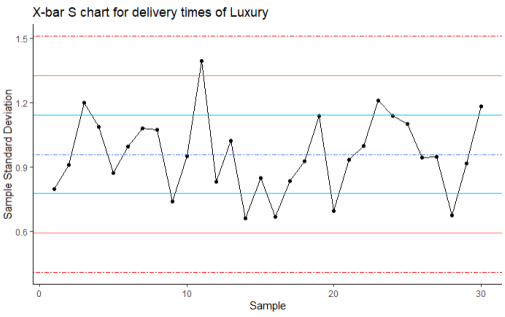
Out of control process: Sweets



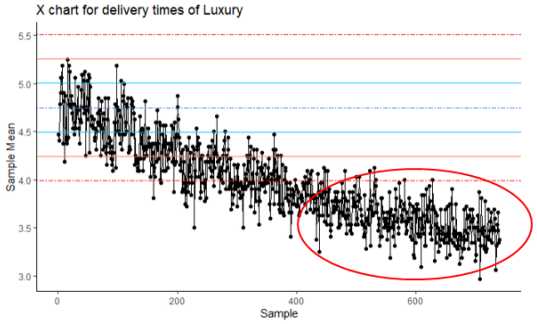
Both the X- and S- charts are within all control limits. The X-chart displays out-of-control signals, as many consecutive points are lying outside of the two-sigma line. Delivery-time control can be investigated to see why so many samples are deviating from being within specifications. Although many samples are outside of the two-sigma line, all of them are below the UCL in the X-chart.

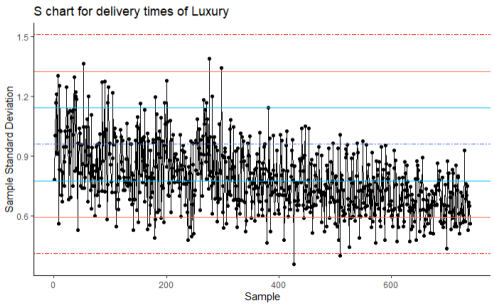
Luxury





Out of control process: Luxury





The X-chart reveals out-of-control signals with a downward tendency. The majority data does not fit onto the model created with the 30 samples. Although these samples are out of the control specifications, because it is below the LCL it won't be considered a major problem due to the delivery times being faster. An investigation should still be done to recalculate averages, centrelines, and to update the process control model. S-chart is also tending downwards, and most samples are below the mean line. The model will need to be reconsidered to account for shorter delivery times.

Part 4:

Samples outside control limits

It is identified that the classes 'Gifts' and 'Household' displays major out of control signals. The following table revels the total amount of samples outside of the UCL and LCL respectively.

Classes	first	second	third	Third_Last	Second_Last	Last	Total
Clothing	282	837	1,048	1,644	1,653	1,723	13
Household	679	693	720	1,335	1,336	1,337	392
Food	432	1,149	1,408	432	1,149	1,408	3
Technology	643						1
Sweets	942	1,243	1,294	1,243	1,294	1,358	4
Gifts	213	216	218	2,607	2,608	2,609	2,287
Luxury	none						1

Table 4: Out of control samples above UCL

Classes	first	second	third	Third_Last	Second_Last	Last	Total
Clothing	1359	1,574	1,587	1,677	1,695	1,756	7
Household	252	387	643	252	387	643	3
Food	75						
Technology	37	345	353	1,933	2,009	2,071	18
Sweets	none						0
Gifts	none						0
Luxury	142	171	184	789	790	791	440

Table 5: Out of control samples below LCL

From the tables above it is observed that Gifts have 2287 lot of out-of-control samples above the Upper specification limit. The company will have to investigate or re-model to identify where the problem is lying. Household has 395 out-of-control samples, which is also concerning.

Possibilities for out of control signals could be due to items being large and harder to deliver, multiple items ordered by the same customer and having to

deliver all at once, or problems on the company's side, such as breakdown, which need to be investigated and corrected.

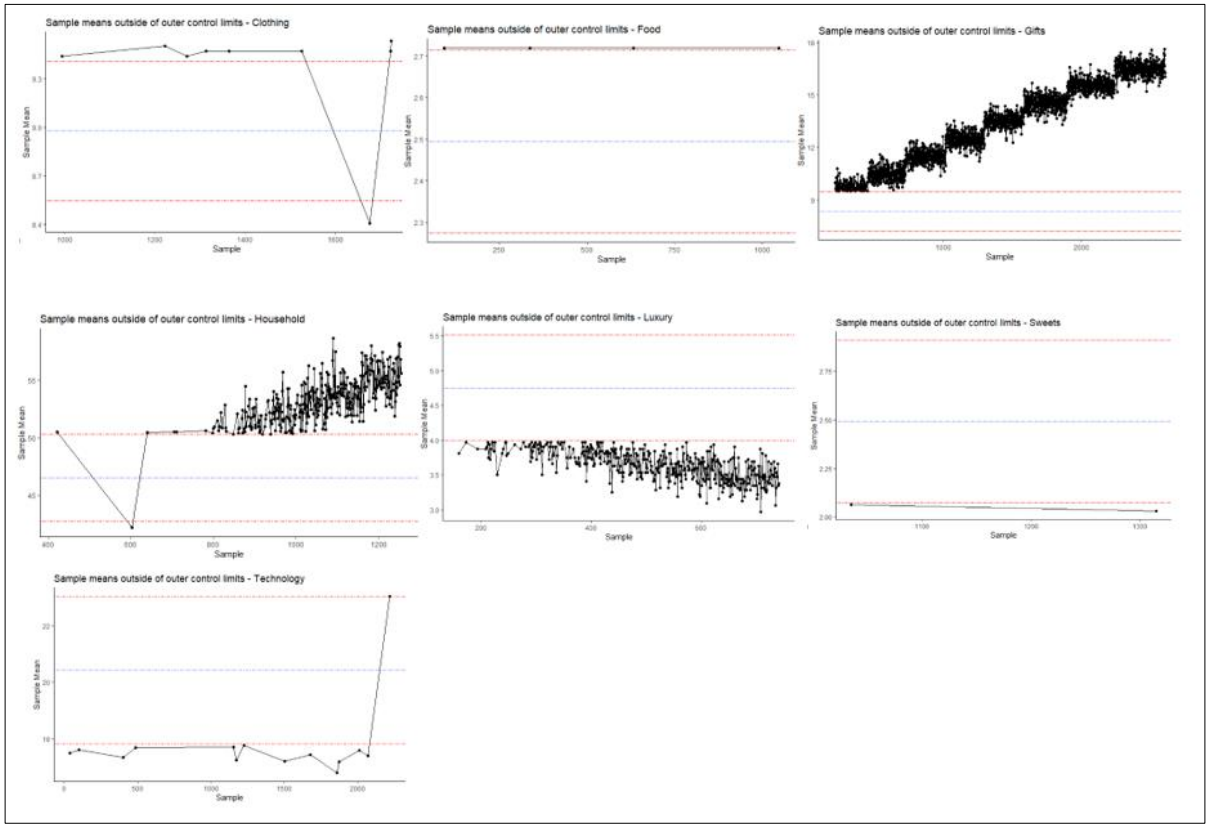


Figure 4: Graphs displaying Out of Control Samples

Class	Most Consecutive (amount)	Last Sample
Clothing	505, 506,507,508,509 (5)	1636
Food	436, 437, 438, 439, 440	1637
Gifts	1647, 1648, 1649, 1650, 1651, 1062, 1063 (7)	2606
Household	430:444 (15)	1254
Luxury	106, 107, 108, 109 (4)	731
Sweets	638:646 (9)	1347
Technology	1660:1677 (18)	1467

Figure 5: Most Consecutive samples of s-bar

Type *I* Error

A type *I* error, also known as a manufacturing error, occur when an out-of-control process is identified when the process is in control. It is the probability that a process gets identified as being out-of-control, when in reality the process is in control.

4.1.A	0.2699796%
4.1.B	27.33%

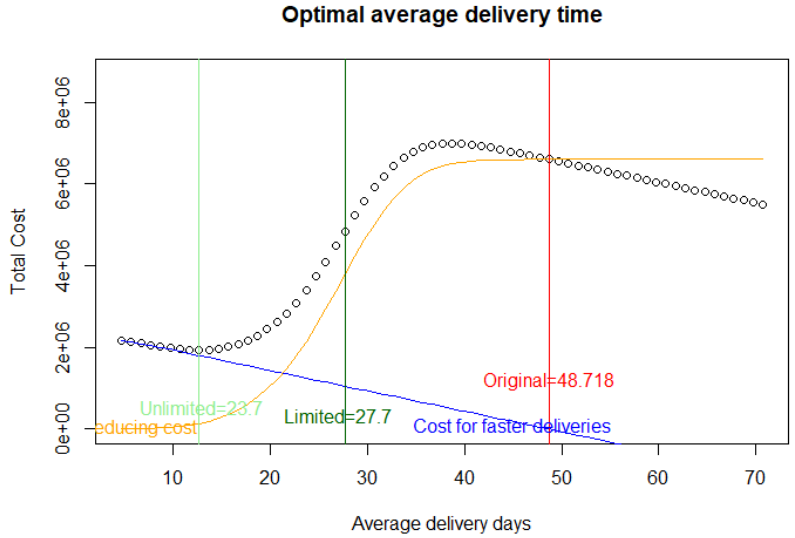
The probability for making a type *I* error of finding a X-bar sample outside the control limit is 0.2699%. The outer control limits are the -3 and +3 sigma control limits.

4.3

For items in the class 'Technology' with a delivery time exceeding 26 hours a cost of R329 per item-late-hour is associated. A cost of R2.50 per item per hour is required if the average time is to be reduced by one hour. The graph shows the optimal (minimum) point, where loss is minimized by a reduction in delivery time.

This results in a minimum loss of R340 870 corresponding with a delivery time reduced by 4 hours.

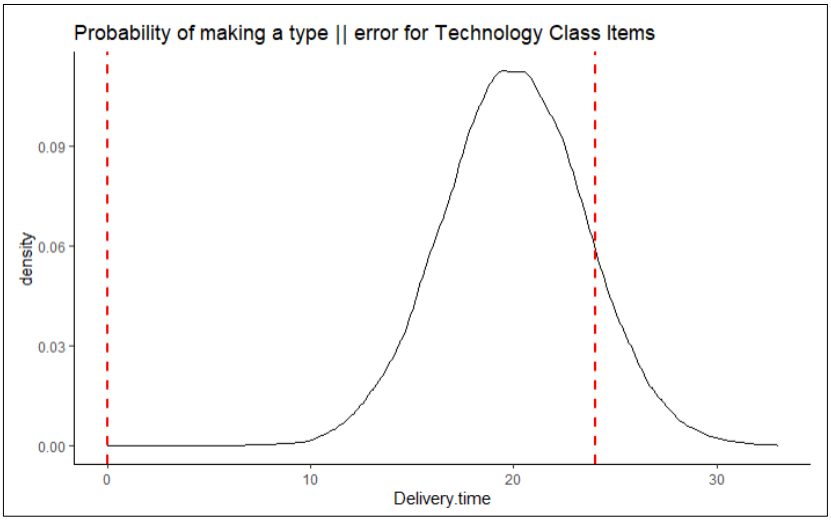
For best profit results the current delivery time needs to be reduced by 4 hours.



Optimal average delivery time in days will be 12.71859.

Type II Error

A type II error occur when the process is said to be in control, but in reality, fails to meet requirements.



Part 5: DOE and MANOVA

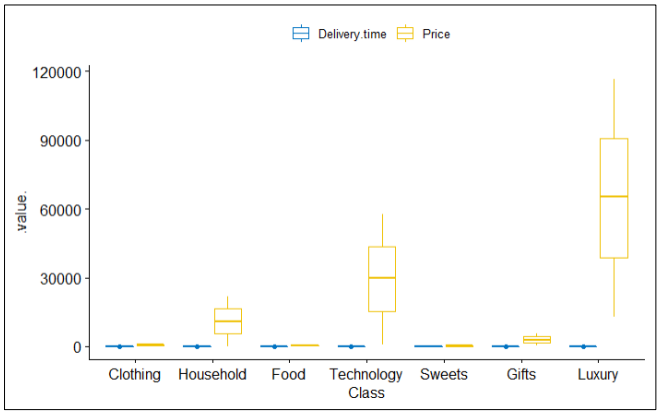
The One-way MANOVA test indicates whether there are differences between independent groups on more than one continuous dependant variable. The following MANOVA is set up to indicate whether the class of an item has an influence on the delivery time and price of an item.

$H_{0,delivery\ time}$: The class of an item has no significant influence on the delivery time of the item.

$H_{1,delivery\ time}$: The class of an item does have a significant influence on the delivery time of the item.

$H_{0,price}$: The class of an item has no significant influence on the price of the item.

$H_{1,price}$: The class of an item does have a significant influence on the price of the item.

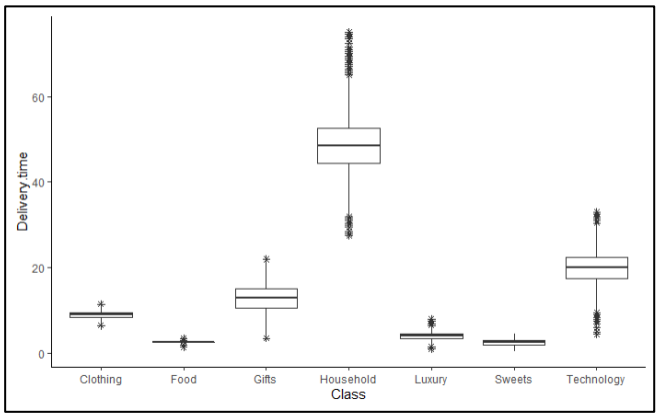


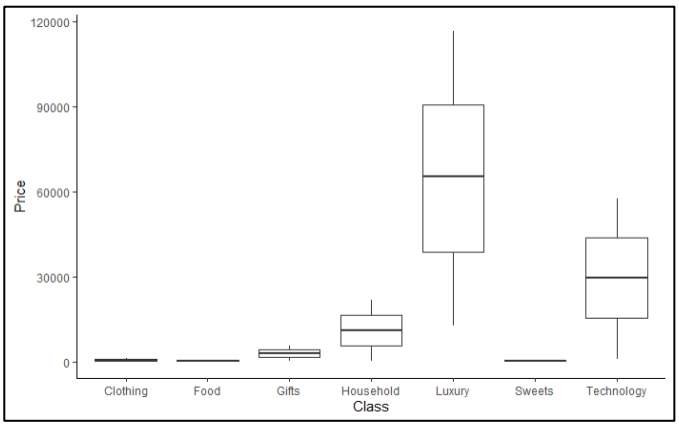

```
              Df Pillai approx F num Df den Df    Pr(>F)
Class          6 1.6797   157291     12 359942 < 2.2e-16 ***
Residuals 179971
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-values for both instances are 2.2e-16 which is significantly small (<<0.05). The null hypotheses in both cases will be rejected.

The conclusion is that the class of an item does have an influence on the delivery time, as well as the price, of the item.

The following graphs confirms the outcome of the MANOVA test.





In both these cases it is clear that both the delivery time and the price of the individual classes displays a significant difference.

Part 6: Reliability of Services and Products

6.1 Lafrideradora

Problem 6

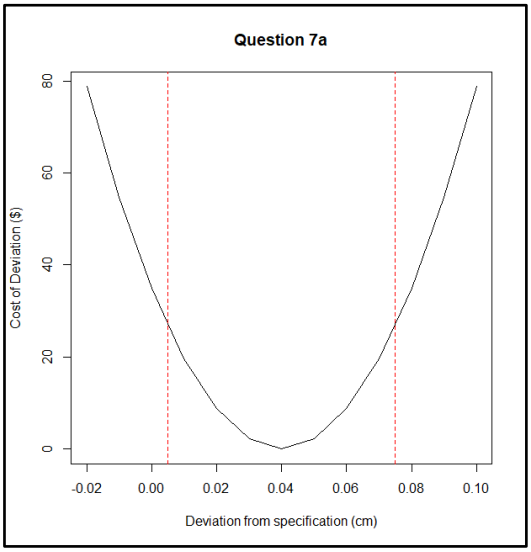
$L = k(y - m)^2$

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

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

Problem 7

7.a



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

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

7.b

$L = 21875(0.027)^2 = 15.95$

Management should decide is they are content with current specifications, as tighter specifications will result in an increase in cost.

6.2 Magnaplex's process investigation

A probability analysis was done to see what the consequences will be if the production line is reduced to only one machine per process. The results showed that the probability for the production line to be fully working will be only

70.38%, compared to the current probability including a backup machine per process, of 96.153%.

A 70.38% reliability is very low considering Technology items to be of very high demand, as well as costly. This will not benefit the company's responsiveness or the profit, and a big risk will be faced if Magnaplex loses their backup machines.

6.3

If the business has 20 delivery vehicles available, with a requirement of 19 to be operating at any given time for reliable service then we would expect 364 days to have reliable delivery given the past days recordings.

Conclusion

The original data table was processed for validity of data instances that was used in the data analysis. The analysed data was visually evaluated and used to ultimately improve the delivery time of the online business. Important information was drawn from the analysis with the goal to maximise profit, the marketing and operation of the company. The probability of different types of errors, and various other calculations were made to see the influence of different decision variables on each other.

Two classes displayed major out of control signals, namely Household and Gifts. The online business can take action to improve and ultimately optimize their total profit.

References

Pqsystems.com. Available at: https://www.pqsystems.com/qualityadvisor/DataAnalysisTools/capability_4.6.3.php (Accessed: October 19, 2022).