Quality Assurance 344 Project

Department of Industrial Engineering, Stellenbosch University

Meghan Young, 23714999@sun.ac.za

2022/10/21

Abstract

Client data for an online business is given, which needs to be analysed. In some circumstances, the data must be cleaned or removed from the dataset before it can be used since they are invalid. Since NA values are regarded as incomplete, all data containing them must be eliminated from the dataset. The data will go through numerous steps so that it can be used. Firstly, data wrangling is done in order to clean the data. Next descriptive statistics will be used to extract relevant and important data. Then statistical process control (SPC) for the X&S charts will be given for the different classes; this will show which classes should be examined. In addition, optimisation is done aiming to show which samples of delivery times are out of control. Manova tests are done, which is comparing multivariate sample means. Lastly, the reliability of the services and products is evaluated.

Table of Contents

List of Tables

List of Figures

Introduction

To analyse the client data for the online business, various statistics will be used, and the data will be programmed in Rstudio. The data will first be cleaned by wrangling the data, then descriptive statistics and statistical process control will be used to evaluate the data, the data will also undergo optimisation and manova tests, and lastly the reliability of the products and services will be analysed.

Part 1: Data Wrangling

Before the data can be used to extract meaningful conclusions, it should first be cleaned. When cleaning the data, it is separated into valid data and invalid (incomplete) data. Before separating the data, the data should be analysed, by splitting the features into continuous features and categorical features. Now the data can be evaluated separately.

Continuous features:

The continuous features are “Age”, “Price”, “Month”, “Day”, and “Delivery time”. When evaluating the continuous data, the minimum, first quantile, median, mean, third quantile, maximum, and missing values are calculated.

A table of the continuous data is provided below.

Table 1: Continuous data

Feature	Min.	Q1	Median	Mean	Q3	Max.	Missing values
Age	18.0	38.00	53.00	54.565639	70.00	108	0
Price	-588.8	482.31	2,259.63	12,293.740474	15,270.74	116,619	17
Month	1.0	4.00	7.00	6.521078	10.00	12	0
Day	1.0	8.00	16.00	15.538761	23.00	30	0
Delivery time	0.5	3.00	10.00	14.500050	18.50	75	0

From looking at the table above, issues can be identified. The first issue is that the minimum for the feature price is a negative value which should not be the case. The second issue that can be seen is that there are 17 missing values for the feature Price. These issues are classified as invalid data, which should be removed from the valid data.

Categorical features:

The categorical features are “Why Bought” and “Class”. To analyse the categorical features, the mode, second mode and missing instances are calculated.

Why Bought

Table 2: Categorical Data: Why Bought

Recommended	Website	Random	Browsing	Email	Spam
107,000	29,450	13,122	18,995	0	4,208

From looking at table 2, the mode is “Recommended”, and the second mode is “Website”. This shows that the biggest reason why the product is bought is that it is recommended, and the second

biggest reason why the product is bought is that it is seen on the Website. There are no missing instances.

Class

Table 3: Categorical Data: Class

Sweets	Household	Gifts	Technology	Luxury	Food	Clothing
21,566	20,067	39,154	36,350	11,869	24,588	26,406

From looking at table 3, the mode is “Gifts”, and the second mode is “Technology”. This means that the type of product that is bought the most is “Gifts” and the type of product that is bought the second most is “Technology”.

Now that the data has been evaluated and a summary is made, the missing values can be removed from the data. All the negative Price values are also removed and is part of the invalid data. When all the invalid data is removed, the valid data is left to analyse.

Part 2: Descriptive Statistics

Process capability indices

The valid data set is analysed by using standard descriptive statistics. The process capability indices Cp, Cpu, Cpl and Cpk for the delivery times of technology class items are calculated. This measures the extent of variation process experiences relative to its specification limits.

Technology’s standard deviation is 3.502 and the mean delivery time for technology is 20.01 hours. With the Upper Specification Limit (USL) of 24 hours and Lower Specification Limit (LSL) of 0, the process capability indices can be calculated. The LSL of 0 is logical because it is the lowest limit a customer will accept; it cannot be a negative value.

Table 4: Process capability indices

Cp	Cpu	Cpl	Cpk
1.142207	0.3796933	1.90472	0.3796933

Cp is the capability potential; it is the measure of the potential capability with the assumption that the distribution of the sample process data is normally distributed and centred within the customer's upper specification limit and lower specification limit. The higher the Cp value, the smaller the spread of the system's output.

Cpu is the process capability based on the upper specification limit. There are numerous delivery times above 24 hours, therefore the Cpu is 0.3769.

Cpl is the process capability based on the lower specification limit. Since the lower limit is zero, the Cpl is very large because there are no delivery times below 0.

Cp and Cpk are used to define the ability of a process to produce a product that meets requirements. Cpk is the minimum between Cpl and Cpu. The higher the Cpk, the better. Cpk less than 1.0 is considered poor, and the process is not capable of delivering within 24 hours. A value between 1.0 and 1.33 is considered barely capable, and a value greater than 1.33 is considered capable.

Data Analysis

The number of rows and columns of the valid data frame is shown in the table below.

Table 5: Valid data rows and columns

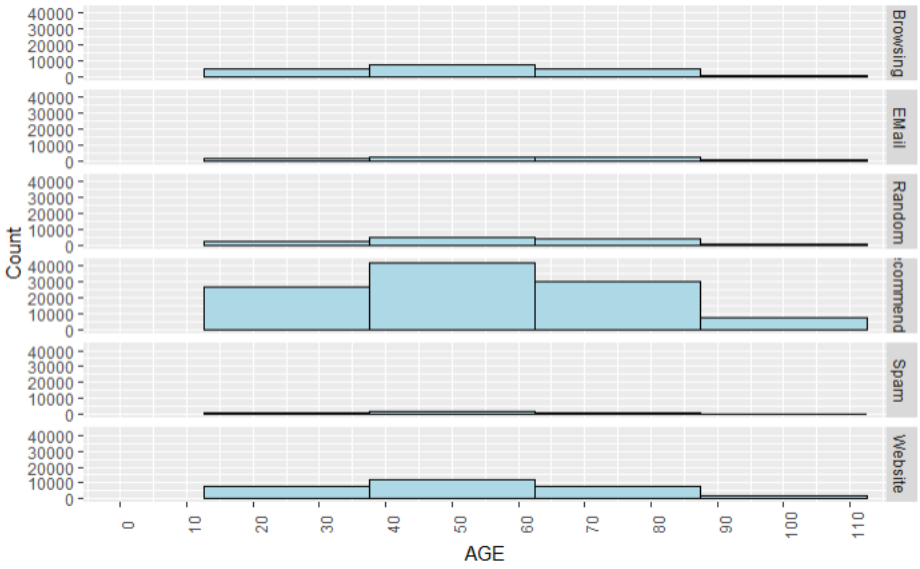
Nr. of rows	Nr. of columns
179,978	10

The continuous and categorical data is analysed separately.

Continuous

Looking at the continuous feature “Age”.

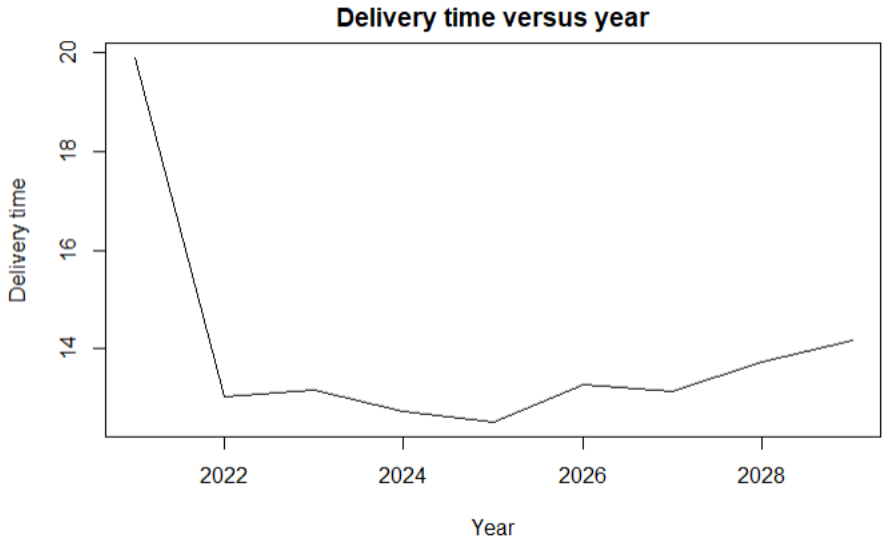
Figure 1: Age distributed between Why bought



The histograms presented above shows all the reasons for buying products (right hand side axis) versus the age of the buyers. The graphs are all distributed over the age of 13-115 years old. For example, looking at “Recommended”, which is the biggest reason for products being bought, the people between the ages of 37- and 62-years old buys has the highest count. All the graphs shows that there is a similar distribution, but just different counts, except for “Email”.

Next looking at the continuous feature “Delivery time”.

Figure 2: Delivery time distributed over the years



Delivery time decreases significantly from 2021 to 2022 and then have small fluctuations. This means that something was done in 2021 to positively impact the delivery times to decrease.

Looking at the Price frequency distribution

Figure 3: Price frequency distribution

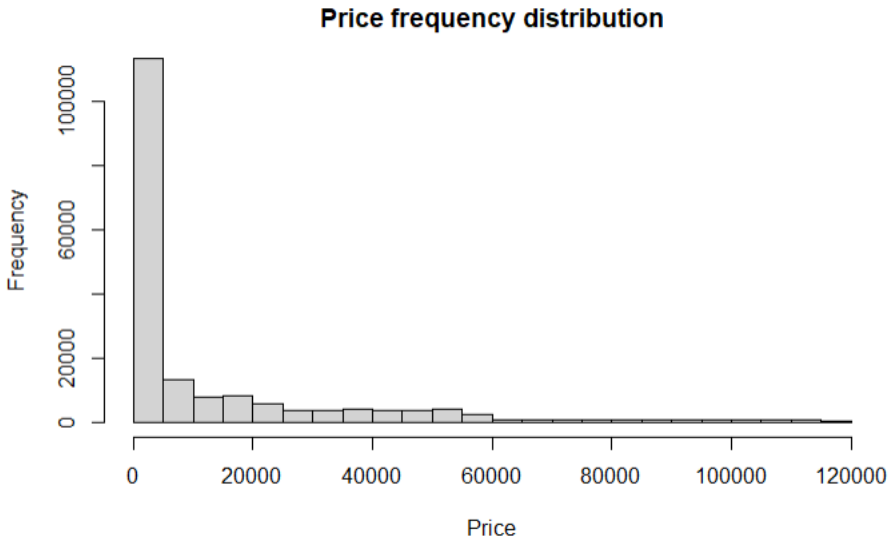
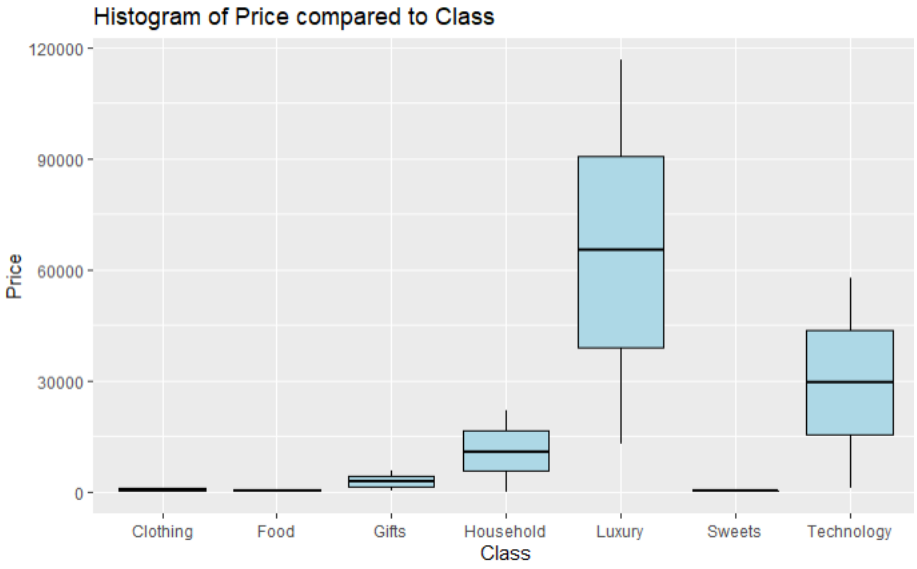


Figure 3 shows that most of the item's prices are below R5000 and that very few items are above R60 000. When analysing the two extremes, it shows that all the sales of "Sweets", "Food", and "Clothing" fall below the R 5 000 threshold. As expected, all the items priced above R 60 000 are "Luxury" items.

Categorical

Looking at the categorical feature "Class".

Figure 4: Price versus Class

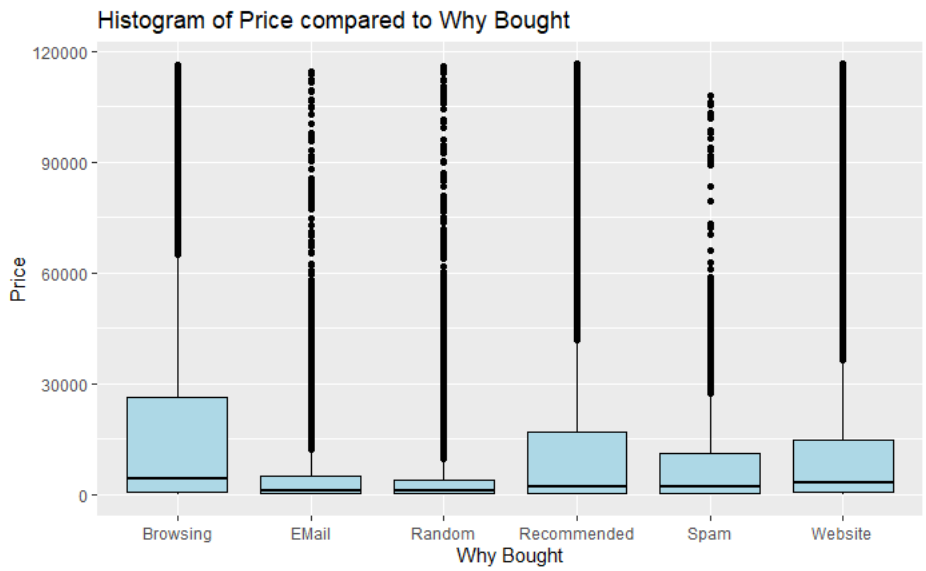


The line in the box shows the median, the median is more or less in the middle of all the box plots, so this means that the Price versus the Class is Normally distributed. This means that Clothing, Gifts, Household, Luxury and Technology are Normally distributed. Food and Sweets median is lower than the box plot meaning that it is skewed right.

There are not many outliers. Luxury are more expensive since its box plot are higher up the plot at the higher prices, and "Clothing", "Food" and "Sweets" are less expensive since their box plots are very low on the plot at the lower prices. Luxury is also more dispersed, so it has the greatest distribution since its box plot are larger, and it also has the highest prices. Luxury is followed by technology, household, and gifts with the highest distribution as well as prices.

Looking at the categorical feature "Why Bought"

Figure 5: Price vs Why Bought



The box plots are low on the plot meaning that 50% of the data is on the lower end of the plot, so 50% is less expensive. The median is lower on the box plot which indicates that the box plots are skewed right. This includes “Browsing”, “Email”, “Random”, “Recommended”, “Spam”, and “Website”. It is visible that the other 50% of the data is above the box plot so more expensive. Browsing is more expensive since 50% of the data is larger up, while email and Random is less expensive but have outliers that are high in price.

Part 3: Statistical Process Control (SPC)

Valid data is all the data without the Na's and the negative “Price” values. So now working with the valid data, a statistical process control (SPC) for the X&S-charts is given. The delivery process times will be used to construct the control charts. 30 samples will be used to calculate the centre lines, outer control limits, the 2-sigma-control limits, and the 1-sigma-control limits for the X&S-charts of the seven processes.

Table 6 is the X-chart that is based on the sample means of the delivery times.

Table 6: X-Chart

class	UCLX	U1Sigmax	U2Sigmax	CLX	L1Sigmas1X	L2SigmaX	LCLX
Clothing	9.404934	9.114978	9.042489	8.970000	8.825022	8.897511	8.535066
Food	2.709458	2.563153	2.526576	2.490000	2.416847	2.453424	2.270542
Luxury	5.493965	4.988359	4.861957	4.735556	4.482752	4.609154	3.977146
Technology	22.974616	21.241168	20.807806	20.374444	19.507721	19.941083	17.774273
Gifts	9.488565	8.736929	8.549020	8.361111	7.985293	8.173202	7.233658
Household	50.248328	47.790924	47.176573	46.562222	45.333520	45.947871	42.876117
Sweets	2.897042	2.617532	2.547655	2.477778	2.338023	2.407900	2.058514

Table 7 is the S-charts that is based on sample standard deviation of the delivery times.

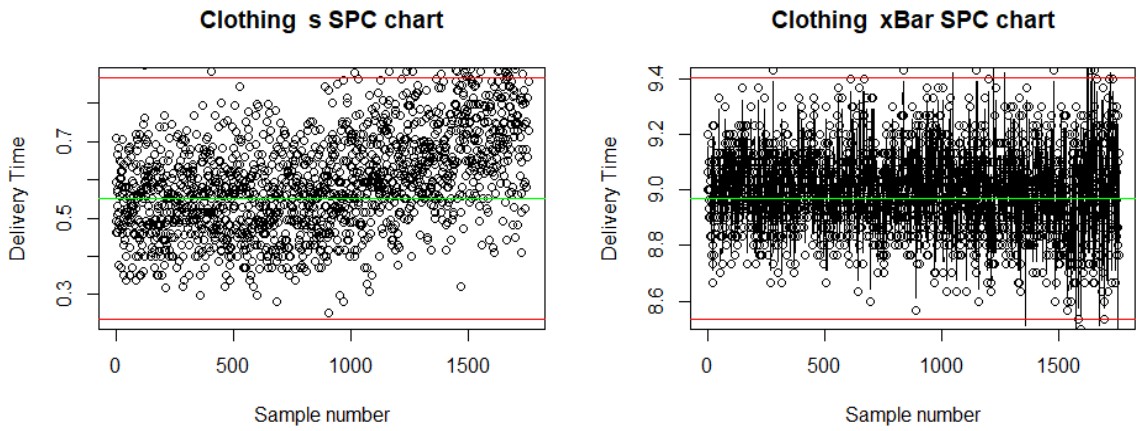
Table 7: S-Chart

class	UCLS	U1Sigmas	U2Sigmas	CLS	L1Sigmas1s	L2Sigmas	LCLS
Clothing	0.8665596	0.6563509	0.6037987	0.5512465	0.4461422	0.4986944	0.2359335
Food	0.4372466	0.3311800	0.3046633	0.2781467	0.2251134	0.2516300	0.1190468
Luxury	1.5110518	1.1445032	1.0528660	0.9612289	0.7779546	0.8695917	0.4114060
Technology	5.1805697	3.9238751	3.6097015	3.2955278	2.6671805	2.9813542	1.4104859
Gifts	2.2463333	1.7014213	1.5651932	1.4289652	1.1565092	1.2927372	0.6115971
Household	7.3441801	5.5626402	5.1172552	4.6718703	3.7811003	4.2264853	1.9995605
Sweets	0.8353391	0.6327039	0.5820450	0.5313862	0.4300686	0.4807274	0.2274333

The following plots are the sample mean and sample standard deviations of the delivery times plotted for every class on the X&S SPC Charts. The outliers assigned should be investigated and analysed.

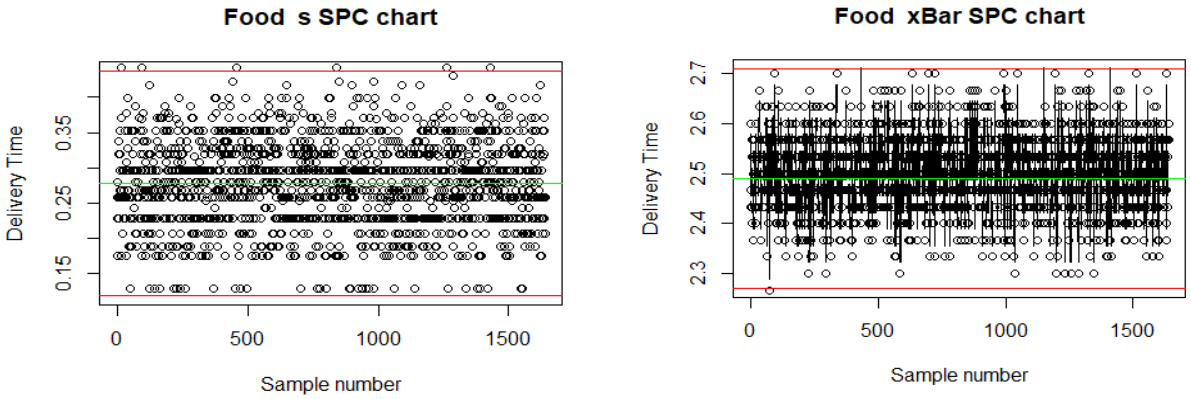
Figure 6: x&s SPC-charts:

Clothing x&s SPC-charts:



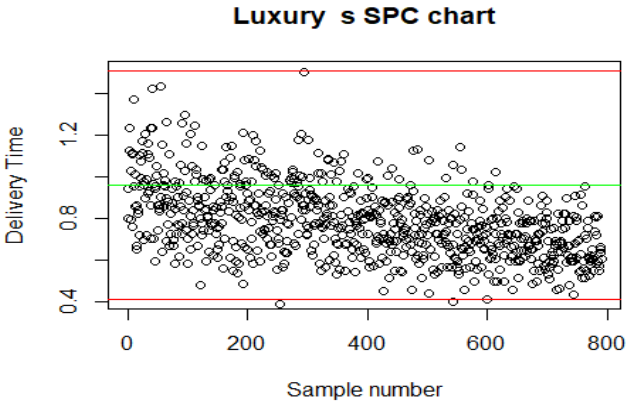
The s SPC chart shows no outliers below the lower control limit but shows increasingly more outliers from sample number 1000 and up. The x SPC chart is more or less consistent throughout the samples but shows some outliers above the outer control limit and very few below the lower control limit after 1300 samples.

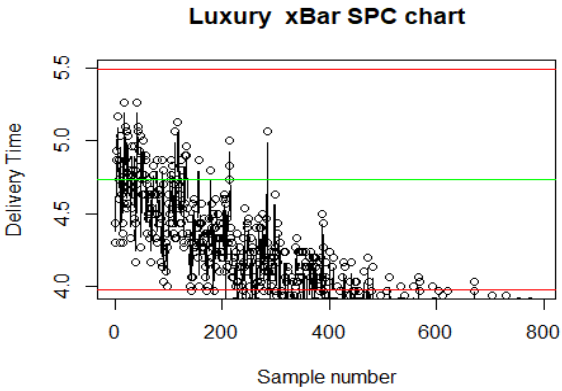
Food x&s SPC-charts:



The s SPC chart shows some outliers above the upper control limit and no outliers below the lower control limit. The x SPC chart shows only one outlier below the lower control limit and is in control. This outlier can be normal since there will always be some outliers.

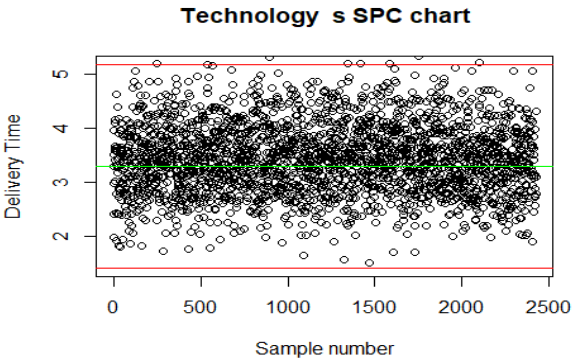
Luxury x&s SPC-charts:

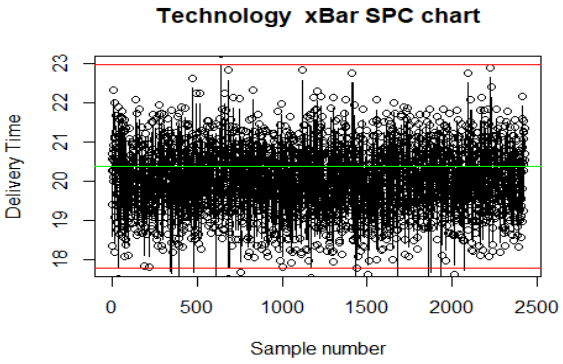




The s SPC chart show that the delivery times decreased every year with very few outliers, which is good. The x SPC charts also shows that the mean delivery times have decreased significantly, with many outliers. The management team should have noticed this, and it should be analysed.

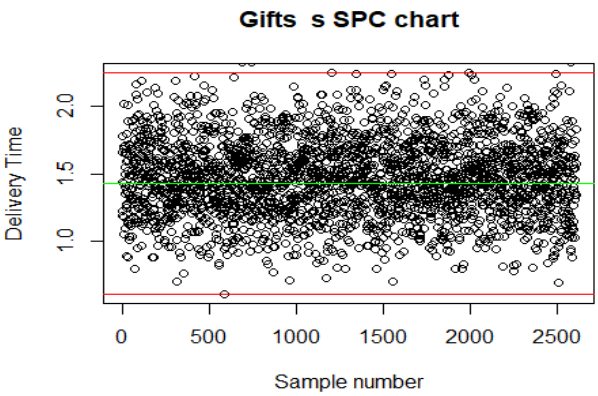
Technology x&s SPC-charts:

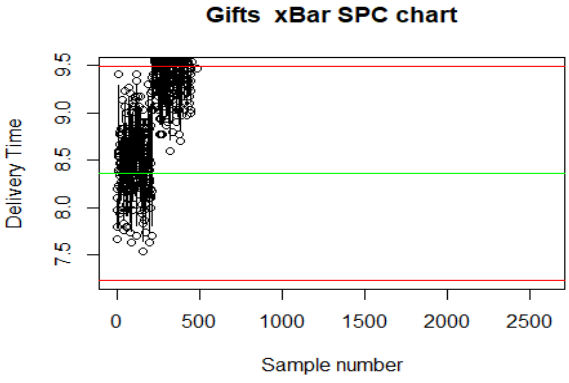




Both the x&s SPC chart for technology shows that the delivery times are very consistent with very few outliers.

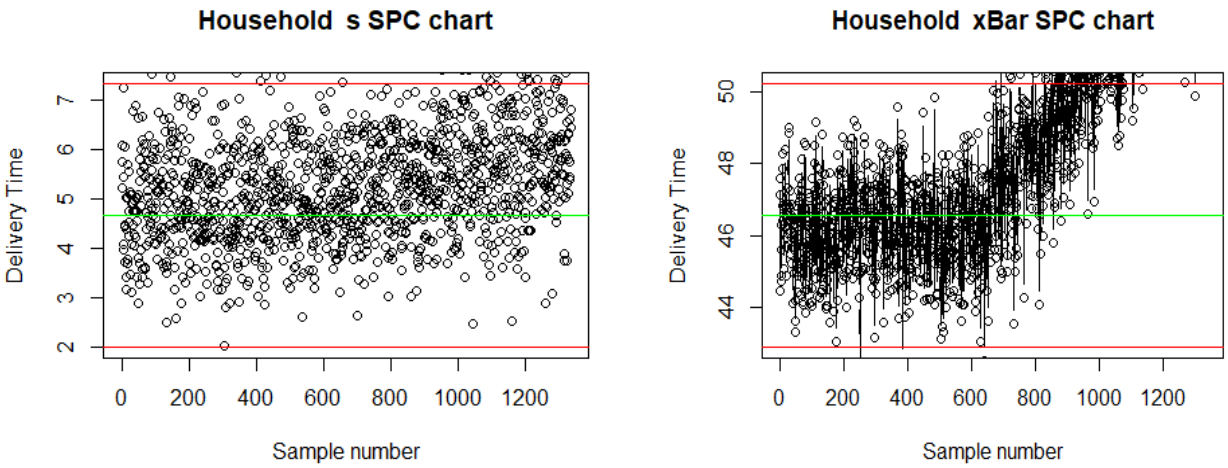
Gifts x&s SPC-charts:





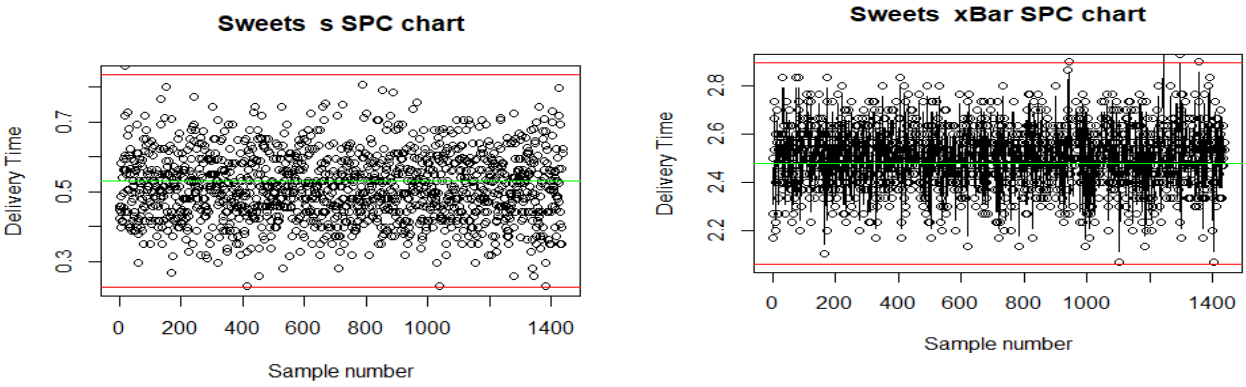
The s SPC chart shows that the delivery times are very consistent with very few outliers. On the other hand, the x SPC chart shows that the delivery times increased significantly after sample 250 resulting in many outliers. This should be analysed.

Household x&s SPC-chart:



The s SPC charts shows the delivery times are in control, with some outliers above the upper control limit. However, the x SPC chart shows that the delivery times increases significantly after sample 600. This should be analysed.

Sweets x&s SPC-chart:



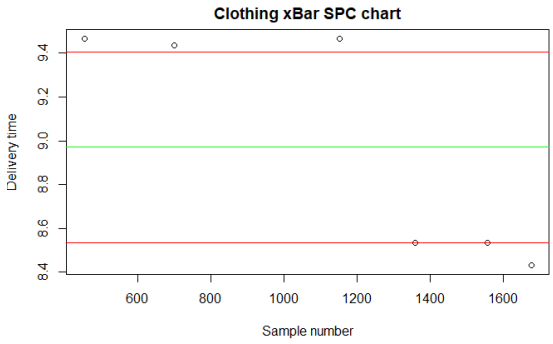
Lastly for sweets, both the x&s SPC charts shows that the delivery times are in control with very few outliers.

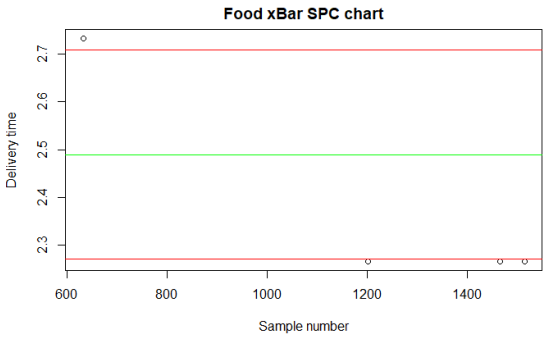
Part 4: Optimisation of the delivery process

Indications of out-of-control sample numbers

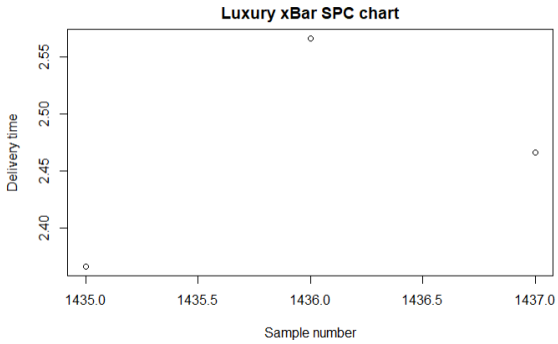
All the valid data is available since it is the end of the period. The following plots are the sample numbers for every class that gave indication of out-of-control samples of the delivery times. The plot shows only the first three samples above or below the control limits.

Figure 7: out of control charts

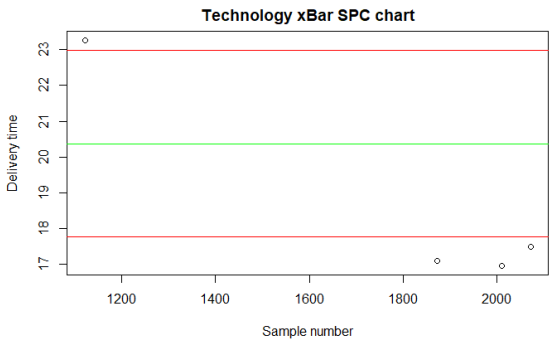


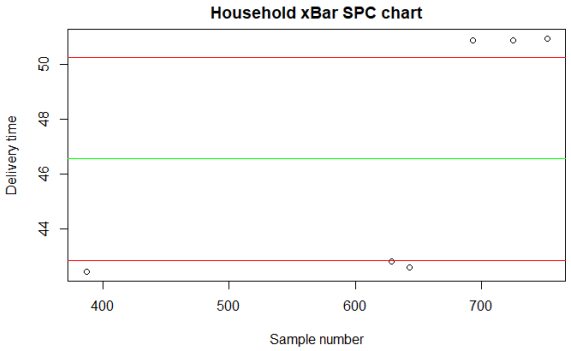


There are some samples below and above the LCL and UCL respectively, for both food and clothing.

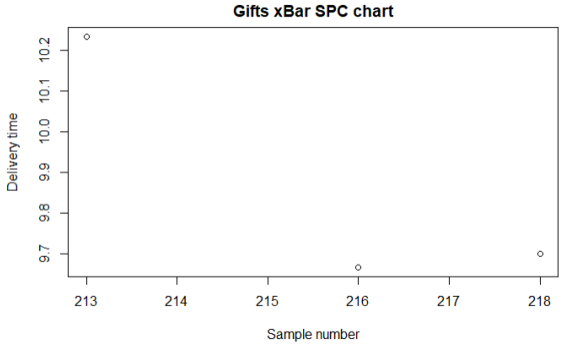


There are only samples that lays below the LCL of 3.977.

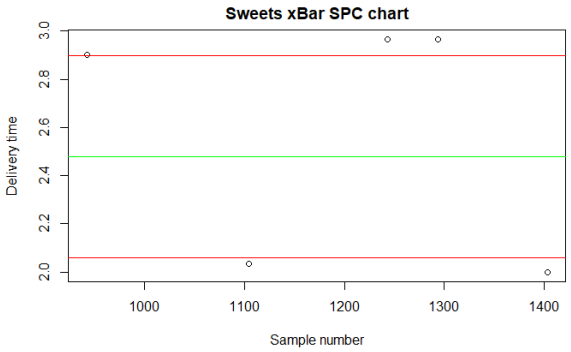




For both Technology and Household, there are very few samples below and above the LCL and UCL respectively. For technology the samples below the LCL are after sample 1800 which can be analysed as to why this is happening at the time period, as well as with household, the samples above the upper control limit are after sample 650, which also can be analysed.



All the samples for gifts are above the UCL of 9.488.



There are very few samples below and above the LCL and UCL respectively.

Most consecutive sample standard deviations

The most consecutive samples standard deviations for the delivery time between -0.3 and +0.4 sigma control limits and the ending sample number are given.

Table 8: most consecutive samples vs the last sample

Class	Most consecutive samples	Index of last sample
Clothing	27	451
Household	14	195
Food	26	176
Technology	17	2108
Sweets	24	1047
Gifts	27	37
Luxury	0	0

By looking at table 8, it can be seen which class's problem has been resolved, and which has not. Clothing, household, food, and gifts have a few consecutive samples, but the index of the last sample is not large, indicating that the issue has been resolved. Technology and sweets also have a few consecutive samples, and their index of last sample is quite large so it can mean that the samples is still recent, meaning problem was not resolved.

Likelihood of making a type 1 error

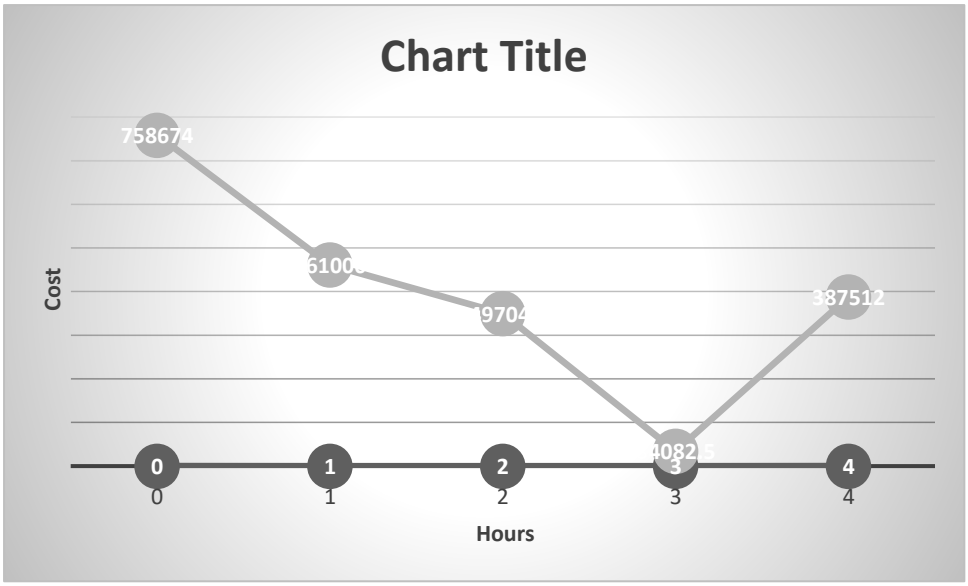
The probability of making a manufacturers (type 1) error is the probability of rejecting Ho (null hypothesis) when Ho is in fact true. Assumption is made of Ho that the process is in control and centred on the centre line using the first 30 samples, so it is assumed that the process is fine, until there is an indication that something might be wrong.

For the "Indication of out-of-control sample numbers" it would be the probability that the means fall outside of the control limits when they are in fact inside the control limits. For "Most consecutive sample standard deviations" it would be that there are many consecutive delivery times between -0.3 and 0.4, when they are not within that bound.

The probability of making this type 1 error for "Indication of out-of-control sample numbers" is 0.002699796. The probability of making this type 1 error for "Most consecutive sample standard deviations" is 0.7266668.

How many hours to centre the delivery process for the best profit

Figure 8: Delivery time vs hours



There was brute force implemented to evaluate how to centre the technology delivery process for the best profit. From the graph it shows that -3 hours shows the minimum cost.

Likelihood of making a type 2 error

The probability of a Consumer's (type 2) error is not to reject Ho when Ho is false. This means thinking the process is in control when it is not. In the type II error, the Ha is true, but we fail to identify this, due to the sample Xbar value being between LCL and UCL.

For "Indication of out-of-control sample numbers" it would be the probability that the sample means are within the outer control limits, when they are in fact not. For "Most consecutive sample standard deviations" it would be the probability that there are no large consecutive delivery times within the -0.3 and 0.4 bounds, when in fact there are large consecutive delivery times within that bound. It can only be made when the process is not in control.

The probability of making this type 1 error for "Indication of out-of-control sample numbers" is 0.784944.

Part 5: MANOVA Test

Test 1

Manova tests are done to compare the difference in price between the two classes "Household" and "Luxury".

Figure 9: Manova test 1

```
 Response 1 :
              Df  Sum Sq Mean Sq F value    Pr(>F)
Price          1     8.5  8.4756  85.597 < 2.2e-16 ***
Residuals 179980 17821.2  0.0990
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response 2 :
              Df Sum Sq Mean Sq F value    Pr(>F)
Price          1 4956.0    4956  145504 < 2.2e-16 ***
Residuals 179980 6130.3       0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

17 observations deleted due to missingness
```

The Probability value of 2.2e-16 is very small. This means that Ho is rejected, and the process is in control.

Test 2

The second test is to compare delivery times over different years.

```
 Response 1 :
                Df  Sum Sq Mean Sq F value    Pr(>F)
Delivery.time    1   934.1  934.14  6394.1 < 2.2e-16 ***
Residuals   179997 26296.6    0.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 2 :
                Df  Sum Sq Mean Sq F value    Pr(>F)
Delivery.time    1    14.8 14.8348  188.17 < 2.2e-16 ***
Residuals   179997 14190.2  0.0788
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 3 :
                Df  Sum Sq Mean Sq F value    Pr(>F)
Delivery.time    1    14.9  14.895  173.14 < 2.2e-16 ***
Residuals   179997 15484.9   0.086
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 4 :
                Df  Sum Sq Mean Sq F value    Pr(>F)
Delivery.time    1    28.3 28.2825  319.51 < 2.2e-16 ***
Residuals   179997 15932.8  0.0885
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 5 :
                Df  Sum Sq Mean Sq F value    Pr(>F)
Delivery.time    1    33.4  33.422  386.17 < 2.2e-16 ***
Residuals   179997 15578.0   0.087
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 6 :
                Df  Sum Sq Mean Sq F value    Pr(>F)
Delivery.time    1    12.6 12.5771  145.98 < 2.2e-16 ***
Residuals   179997 15507.4  0.0862
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 7 :
                Df  Sum Sq Mean Sq F value    Pr(>F)
Delivery.time    1    18.3 18.2652  196.79 < 2.2e-16 ***
Residuals   179997 16706.5  0.0928
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
 Response 8 :
                Df  Sum Sq Mean Sq F value    Pr(>F)
Delivery.time    1     6.9  6.9483  68.541 < 2.2e-16 ***
Residuals   179997 18247.0  0.1014
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 9 :
                Df  Sum Sq Mean Sq F value    Pr(>F)
Delivery.time    1     1.7 1.68284  15.401 8.699e-05 ***
Residuals   179997 19668.5 0.10927
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The probability value is also very small. The Ho is rejected, and the process is not in control.

Part 6: Reliability of the service and products

Taguchi loss function

Problem 6

Scrap cost = $45 with specification 0.04 cm

k = 45/(0.04)^2 = 28125

Taguchi loss function = L(x) = $k(x - T)^2$

=28125 $k(x - T)^2$

Problem 7

a) Scrap cost = $35 with specification 0.04 cm

k = 35/(0.04)^2 = 21875

Taguchi loss function = L(x) = $k(x - T)^2$

=21875 $k(x - T)^2$

b) $L(x) = 21875(0.027 - 0.04)^2$
$= \$3.70$

System reliability

Problem 27

a) The machines are in series:
$R_s = 0.85 \times 0.92 \times 0.9$
$= 70.38\%$ Reliability

b) The machines are in parallel:
$R_p = (1 - (1 - 0.85)) \times (1 - (1 - 0.92)) \times (1 - (1 - 0.9))$
$= 99.88\%$ Reliability

Binomial Probabilities

Vehicles:

Probability that vehicles is available:

Weighted probability = 0.7052%

Probability that a number of vehicles fail in days:

Probability that 0 fails = 86.19% in 1344 days

Probability that 1 fails = 12.86% in 200 days

Probability that 2 fails = 0.9131% in 14 days

Probability that 3 fails = 0.0410% in 0.64 days

Probability that 4 fails = 0.0013% in 0.02 days

Expected percentage of reliable days: 99.96%

Expected number of days: 364.84

Drivers:

Probability that drivers are available:

Weighted probability = 0.3221%

Probability that drivers fail in days:

Probability that 0 fails = 93.45% in 1457 days

Probability that 1 fails = 6.34% in 98 days

Probability that 2 fails = 0.20% in 3 days

Probability that 3 fails = 0.004% in 0.065 days

Overall probability:

Percentage of reliable days = 99.95%

Expected number of days = 364 days

Expected percentage of reliable days: 99.99%

Expected number of days: 364.98

Vehicles and drivers:

P(total reliability) = P(vehicle reliability) x P(Drivers reliability)

= 99.99% x 99.96%

= 99.95%

Changing to 22 Vehicles and 21 drivers:

Probability that vehicles fail in days:

Probability that 0 fails = 85.558% in 1335 days

Probability that 1 fails = 13.37% in 208 days

Probability that 2 fails = 0.997% in 15 days

Probability that 3 fails = 0.047% in 0.7367 days

Probability that drivers fail in days:

Probability that 0 fails = 93.45% in 1457 days

Probability that 1 fails = 6.34% in 98 days

Probability that 2 fails = 0.20% in 3 days

Probability that 3 fails = 0.004% in 0.065 days

Conclusion:

This report analysed the data for an online business. There were six parts to this report: Data wrangling, descriptive statistics, statistical process control, optimisation, MANOVA tests, reliability of the service and products. It was seen that the biggest reason for why the products is bought is "Recommended". It also showed that more products fall under R5000, and it is mostly the classes Sweets, Clothing and Food. Gifts and Household delivery time increased significantly which should be investigated. Luxury and Gifts have out of control samples which also should be investigated. The system has a high probability of 99.95%

References

Banerjee, A. et al., 2009. *National Library of Medicine.* [Online]
Available at:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996198/#:~:text=A%20type%20I%20error%20(false,actually%20false%20in%20the%20population.
[Accessed 15 10 2022].

McNeese, D. B., 2004. *Process Capability Part 2.* [Online]
Available at: https://www.spcforexcel.com/knowledge/process-capability/process-capability-part-2#:~:text=The%20value%20of%20Cpk%20is,as%20shown%20in%20the%20figure.
[Accessed 12 10 2022].

R Documentation, 2022. *R Documentation.* [Online]
Available at: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/uniroot
[Accessed 13 October 2022].