



UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvenoot • your knowledge partner

ECSA PROJECT

Quality Assurance 344 Final Report

Authors
23841133- PS DU PREEZ

17 October 2022

saam vorentoe • masiye phambili • forward together

Department of Mechanical and Mechatronic Engineering
Departement Meganiese en Megatroniese Ingenieurswese
Privaat Sak X1, Private Bag X1, Matieland, 7602
Tel: +27 21 808 4204 | www.eng.sun.ac.za



ENGINEERING
EZOBUNJINELI
INGENIEURSWESE

Table of Contents

List of figures	iv
List of tables	vi
List of symbols	vii
1 Introduction	1
1.1 Background.....	1
1.2 Objectives.....	1
1.3 Motivation	1
2 Valid and Invalid Data.....	2
2.1 Isolating valid data	2
2.2 Isolating invalid data	3
3 Descriptive statistics.....	5
3.1 Categorical features	5
3.1.1 ID	5
3.1.2 Class	6
3.1.3 Year	6
3.1.4 Month.....	7
3.1.5 Days	7
3.1.6 Why bought?	7
3.2 Continuous features	7
3.2.1 Age	8
3.2.2 Delivery Time	10
3.2.3 Price	12
3.3 Scatter Plot Matrix (SPLOM)	13
3.4 Process Capability Indices	15
3.4.1 Potential Capability	16
4 Statistical Process Control.....	17
4.1 Tables	17
4.1.1 Values for S-chart	17
4.1.2 Values for X-chart	17
4.2 First 30 samples.....	17
4.2.1 Luxury	18
4.2.2 Gifts	19
4.2.3 Sweets	20

4.2.4	Technology	21
4.2.5	Food	22
4.2.6	Household.....	23
4.2.7	Clothing.....	24
4.3	All Samples	25
4.3.1	Luxury	25
4.3.2	Gifts	26
4.3.3	Sweets	27
4.3.4	Technology	28
4.3.5	Food	29
4.3.6	Household.....	30
4.3.7	Clothing.....	31
5	Optimizing Delivery Process.....	32
5.1	Samples beyond control limits	32
5.2	First 3 and last 3 samples outside control limits.....	32
5.3	Most consecutive samples within -0.3 and +0.4 Sigma	36
5.4	Estimate the likelihood of making type 1 error for A & B.....	37
5.4.1	For A	37
5.4.2	For B	37
5.5	Minimizing delivery cost.....	38
5.6	Estimate the likelihood of making type II error for A	38
6	MANOVA testing	40
6.1	Hypothesis 1 (effect of class)	40
6.1.1	Hypothesis statements	40
6.1.2	P-Values	40
6.1.3	Visualisation.....	40
6.1.4	Conclusion	41
6.2	Hypothesis 2 (effect of why bought)	42
6.2.1	Hypothesis statements	42
6.2.2	P-values	42
6.2.3	Visualisation.....	42
6.2.4	Conclusion	43
7	Reliability of the service and products	45
7.1	Taguchi loss	45
7.1.1	At scrap cost of \$45	45
7.1.2	At scrap cost of \$35	46
7.1.3	Process deviation from target is reduced to 0.027	46
7.2	System reliability	47
7.2.1	Reliability if only one machine at A, B and C is used	48

7.2.2	Reliability if two machines at A, B and C is used	48
7.3	Binomial distribution	48
7.3.1	Case 1: 20 vehicles available	48
7.3.2	Case 2: 21 vehicles available	49
7.3.3	Conclusion:	49
8	Conclusion.....	50
9	References	Error! Bookmark not defined.

List of figures

Figure 1: Example of isolated valid data	2
Figure 2: Example of removed missing values.....	3
Figure 3: Example of removed negative values	3
Figure 4: Invalid dataset.....	4
Figure 5: Enlarged distribution of Age	8
Figure 6: Age distribution per class	9
Figure 7: Enlarged distribution of delivery time	10
Figure 8: Delivery time per class.....	11
Figure 9: Delivery time per class y-axis shortened	11
Figure 10: Class vs Price	12
Figure 11: Class vs Price zoomed	13
Figure 12: Scatterplot Matrix for continuous features.....	13
Figure 13: Delivery Time vs Price scatterplot	14
Figure 14: Delivery time vs Price + Regression line	15
Figure 15: Delivery time for technology	16
Figure 16: S- and X charts for first 30 Luxury samples.....	18
Figure 17: S- and X charts for first 30 Gifts samples	19
Figure 18: S- and X charts for first 30 Sweet samples	20
Figure 19: S- and X charts for first 30 Technology samples.....	21
Figure 20: S- and X charts for first 30 Food samples	22
Figure 21: S- and X charts for first 30 Household samples.....	23
Figure 22: S- and X charts for first 30 Clothing samples.....	24
Figure 23: S- and X charts for all Luxury samples	25
Figure 24: S- and X charts for all Gift samples	26
Figure 25: S- and X charts for all Sweet samples	27
Figure 26: : S- and X charts for all Technology samples	28
Figure 27: S- and X charts for all Food samples.....	29
Figure 28: S- and X charts for all Household samples.....	30
Figure 29: S- and X charts for all Clothing samples.....	31

Figure 30: The first and last 3 outliers for Gifts	33
Figure 31: The first and last 3 outliers for Household	34
Figure 32: The first and last 3 outliers for Luxury	35
Figure 33: Technology S-chart overlaid with -0.3 and +0.4 Sigma Control lines...	36
Figure 34: Most consecutive technology S-samples within limits.....	37
Figure 35: Graph of total cost for various delivery times	38
Figure 36: Age vs Class of product	40
Figure 37: Delivery time vs Class of product	41
Figure 38: Class vs Price of product	41
Figure 39: Year vs Why Bought	42
Figure 40: Month vs Why bought	43
Figure 41: Day vs Why bought	43
Figure 42: Taguchi loss function at \$45 scrap cost	45
Figure 43: Taguchi loss function at \$35 scrap cost	46
Figure 44: Taguchi loss at process deviation of 0.027 cm	47
Figure 45: Reliability of machines	47

List of tables

Table 1: Categorical features part 1	5
Table 2: Categorical features part 2	6
Table 3: Continuous features.....	7
Table 4: Table of process capability indices.....	15
Table 5: Values for S-chart.....	17
Table 6: Values for X-chart.....	17
Table 7: Number of outliers per class	32
Table 8: Maximum consecutive samples.....	36
Table 9: MANOVA table for hypothesis 1	40
Table 10: MANOVA values for hypothesis test 2.....	42

List of symbols

$H0$	Null hypothesis
$H1$	Alternative hypothesis
L	Loss (\$)
k	Taguchi loss constant
P	Probability

1 Introduction

1.1 Background

An online company at the start of 2030 has been collecting data on all sales over the last 9 years (since 2021). They sell a variety of different consumer products to a diverse customer group. A data analyst has been employed to identify meaningful correlations and comment on the quality of service they have been giving their customers.

1.2 Objectives

The objectives of this report is to wrangle the data in to a suitable set for analysis. Thereafter to obtain meaningful descriptive statistics followed by statistical process control. The delivery process can then be optimized and MONOVA testing of various relevant hypothesis is executed. The reliability of the service and product delivery will then be analysed.

1.3 Motivation

Quality Control and analysing past performance of business is crucial for sustainable success. Identifying recurring problems that are not evident before analysis can save the business a lot of money in the long run. Useful correlations can also be identified to help the business focus on the correct products to maximise profits.

2 Valid and Invalid Data

Before analysing the client data for the online business, invalid data must be cleaned or removed from the dataset. The dataset contains 180 000 instances with 10 features each. Missing values in any feature of an instance classify the instance as compromised or incomplete, and any instance that contain them must be eliminated from the dataset. Instances that contain negative values for non-negative data (such as time and price) also need to be removed.

2.1 Isolating valid data

The valid data set created is all data that does not contain negative values for non-negative features and has no missing values. According to these constraints 179 978 data entries were regarded as valid data. These entries were isolated and placed in a new Excel sheet.

rowValid	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	1	19966	54	Sweets	246.21	2021	7	3	1.5	Recommended
2	2	34006	36	Household	1708.21	2026	4	1	58.5	Website
3	3	62566	41	Gifts	4050.53	2027	8	10	15.5	Recommended
4	4	70731	48	Technolog	41843.21	2029	10	22	27	Recommended
5	5	92178	76	Household	19215.01	2027	11	26	61.5	Recommended
6	6	50586	78	Gifts	4929.82	2027	4	24	14.5	Random
7	7	73419	35	Luxury	108953.5	2029	11	13	4	Recommended
8	8	32624	58	Sweets	389.62	2025	7	2	2	Recommended
9	9	51401	82	Gifts	3312.11	2025	12	18	12	Recommended
10	10	96430	24	Sweets	176.52	2027	11	4	3	Recommended
11	11	87530	33	Technolog	8515.63	2026	7	15	21	Browsing

Figure 1: Example of isolated valid data

The first 11 instances of the 179 978 are shown as a demonstration of how the new dataset looks.

The new dataset's primary key is labelled 'row Valid' and only starts to differ from the original primary key ('X') after instance 12 344 where the first missing value was in the original data set.

rowValid	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
12339	12338	12338	38677	64 Food	224.82	2025	5	2	2.5	Recommended
12340	12339	12339	69019	44 Gifts	1795.49	2023	6	1	10.5	Recommended
12341	12340	12340	16593	30 Clothing	633.99	2021	5	10	9	Website
12342	12341	12341	78109	71 Gifts	3424.01	2025	7	29	12.5	Random
12343	12342	12342	88576	54 Gifts	752.79	2028	7	21	13.5	Recommended
12344	12343	12343	27986	37 Clothing	712.19	2021	10	10	9	Recommended
12345	12344	12344	90260	34 Luxury	42891.66	2025	8	4	4	Recommended
12346	12345	12346	92286	32 Technolog	38167.24	2028	7	6	19.5	Website
12347	12346	12347	89263	44 Clothing	891.71	2021	7	2	8.5	Recommended
12348	12347	12348	71191	49 Househol	14936.31	2025	10	11	43.5	Recommended
12349	12348	12349	24801	28 Food	425.96	2022	1	29	2.5	Recommended
12350	12349	12350	85475	57 Luxury	78817.55	2026	3	21	5	Browsing
12351	12350	12351	61842	24 Clothing	1008.78	2025	7	16	8	Recommended
12352	12351	12352	49373	34 Technolog	17277.26	2024	10	11	14.5	Browsing
12353	12352	12353	40283	45 Technolog	16930.76	2025	3	9	27.5	EMail
12354	12353	12354	19084	56 Sweets	171.81	2026	10	8	1.5	Random
12355	12354	12355	53251	30 Clothing	322.12	2021	8	26	10	Recommended
12356	12355	12356	21484	63 Gifts	2099.09	2027	1	3	12	Browsing

Figure 2: Example of removed missing values

The highlighted entries are to indicate where the original primary key and the new primary key begins to differ.

Negative values were also removed, and the first negative value was at instance 16320. Highlighted in this image is where the original primary key again starts to differ from the new primary key. The reason for it jumping with 2 is because instance 16321 was also removed for containing a missing value.

rowValid	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
16314	16313	16314	93909	52 Clothing	584.54	2021	4	12	8.5	Recommended
16315	16314	16315	55073	74 Gifts	5495.16	2021	3	18	9.5	Recommended
16316	16315	16316	62208	22 Clothing	597.77	2023	8	11	9	Recommended
16317	16316	16317	29089	53 Food	270.36	2021	5	8	2.5	Website
16318	16317	16318	17238	50 Gifts	5414.89	2028	1	11	14	Recommended
16319	16318	16319	94129	46 Food	654.58	2028	11	29	2.5	Recommended
16320	16319	16322	84530	31 Food	359.03	2021	6	16	2.5	Recommended
16321	16320	16323	96374	43 Househol	19177.63	2025	1	15	42.5	Website
16322	16321	16324	82990	92 Food	497.14	2026	8	23	2.5	EMail
16323	16322	16325	42833	32 Househol	15031.99	2025	9	24	47.5	Recommended
16324	16323	16326	90408	52 Technolog	7745.94	2025	4	16	25	Browsing
16325	16324	16327	56475	49 Food	214.25	2023	10	1	2	Random

Figure 3: Example of removed negative values

2.2 Isolating invalid data

To still have access to invalid data in case it is needed in the future it was isolated and stored in its own Excel file.

The invalid data contains 17 instances with missing values and 5 instances with negative values in the price column. Negative values in the price column can either be for user input

errors or were mistakenly used to indicate refunds. However, placing refunds in the sales data could have a noticeable effect on categorical features such as means calculated for sales.

	A	B	C	D	E	F	G	H	I	J	K
1	rowN	X	ID	AGE	Class	Price	Year	Mont	Day	Delivery.tim	Why.Bough
2	1	16320	44142	82	Household	-588.8	2023	10	2	48	Email
3	2	19540	65689	96	Sweets	-588.8	2028	4	7	3	Random
4	3	19998	68743	45	Household	-588.8	2024	7	16	45.5	Recommended
5	4	144443	37737	81	Food	-588.8	2022	12	10	2.5	Recommended
6	5	155554	36599	29	Luxury	-588.8	2026	4	14	3.5	Recommended
7	6	12345	18973	93	Gifts		2026	6	11	15.5	Website
8	7	16321	81959	43	Technology		2029	9	6	22	Recommended
9	8	19541	71169	42	Technology		2025	1	19	20.5	Recommended
10	9	19999	67228	89	Gifts		2026	2	4	15	Recommended
11	10	23456	88622	71	Food		2027	4	18	2.5	Random
12	11	34567	18748	48	Clothing		2021	4	9	8	Recommended
13	12	45678	89095	65	Sweets		2029	11	6	2	Recommended
14	13	54321	62209	34	Clothing		2021	3	24	9.5	Recommended
15	14	56789	63849	51	Gifts		2024	5	3	10.5	Website
16	15	65432	51904	31	Gifts		2027	7	24	14.5	Recommended
17	16	76543	79732	71	Food		2028	9	24	2.5	Recommended
18	17	87654	40983	33	Food		2024	8	27	2	Recommended
19	18	98765	64288	25	Clothing		2021	1	24	8.5	Browsing
20	19	144444	70761	70	Food		2027	9	28	2.5	Recommended
21	20	155555	33583	56	Gifts		2022	12	9	10	Recommended
22	21	166666	60188	37	Technology		2024	10	9	21.5	Website
23	22	177777	68698	30	Food		2023	8	14	2.5	Recommended

Figure 4: Invalid dataset

The 'X' column indicates the original primary keys from where features were removed. The 'rowNew' column is used as the new primary key.

3 Descriptive statistics

Descriptive statistics are a set of standard recognized calculations and graphs that allow for the identification of trends, correlations, and other useful insights in data. The 2 main types of data analysed are categorical and continuous features.

3.1 Categorical features

Categorical features have a set number of possible answers (Philips, 2013) and are often integers or strings/characters. The categorical features in this data set are:

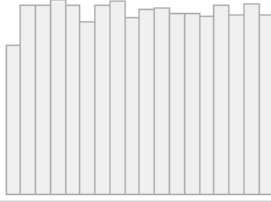


Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
ID [integer]	Mean (sd) : 55234.7 (25740.3) min ≤ med ≤ max: 11126 ≤ 55081 ≤ 99992 IQR (CV) : 44937 (0.5)	15000 distinct values		0 (0.0%)
Class [character]	1. Clothing 2. Food 3. Gifts 4. Household 5. Luxury 6. Sweets 7. Technology	26403 (14.7%) 24582 (13.7%) 39149 (21.8%) 20065 (11.1%) 11868 (6.6%) 21564 (12.0%) 36347 (20.2%)		0 (0.0%)
Year [integer]	Mean (sd) : 2024.9 (2.8) min ≤ med ≤ max: 2021 ≤ 2025 ≤ 2029 IQR (CV) : 5 (0)	2021 : 33443 (18.6%) 2022 : 15546 (8.6%) 2023 : 17128 (9.5%) 2024 : 17698 (9.8%) 2025 : 17267 (9.6%) 2026 : 17152 (9.5%) 2027 : 18656 (10.4%) 2028 : 20613 (11.5%) 2029 : 22475 (12.5%)		0 (0.0%)

Table 1: Categorical features part 1

3.1.1 ID

ID is the client ID that made every purchase. The distribution of ID's is fairly uniform in nature and there are 15000 distinct values. Based on this information an approximate can be made that the average client made 12 purchases.

$$\frac{17\,978 \text{ purchases}}{15000 \text{ unique clients}} = 11.99 \approx 12 \text{ purchases per client}$$

3.1.2 Class

There are 7 distinct classes in which all sales are contained. The mode is Gifts at 39 149 accounting for 21.8% of all data. Technology is shortly behind it. Luxury items account for only 6.6% of the sales most probably due to the high prices. Further analysis of the effect that classes have on various continuous features will be discussed later in the report.

Emphasis should be placed on the quality of Gifts and Technology and enough safety stock should be kept ensuring no stockouts. These 2 classes account for almost half of all sales and any problems will thus affect many customers.

3.1.3 Year

The sales data span across 9 years, from 2021 to 2029. A large spike in sales can be seen in 2021. This can most probably be attributed to all sales before 2021 also being grouped to 2021. The rest of the data is trending upward with a gradual rise from 2022 (15 546) to 2029 (22 475) – an average yearly increase of 5.57%. This trend appears to be sustainable across multiple years and can be used to forecast sales in 2030. This forecast can be used in capacity calculations and for ordering raw material.



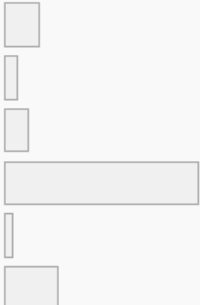
Month [integer]	Mean (sd) : 6.5 (3.5) min ≤ med ≤ max: 1 ≤ 7 ≤ 12 IQR (CV) : 6 (0.5)	12 distinct values		0 (0.0%)
Day [integer]	Mean (sd) : 15.5 (8.6) min ≤ med ≤ max: 1 ≤ 16 ≤ 30 IQR (CV) : 15 (0.6)	30 distinct values		0 (0.0%)
Why.Bought [character]	1. Browsing 2. EMail 3. Random 4. Recommended 5. Spam 6. Website	18994 (10.6%) 7224 (4.0%) 13120 (7.3%) 106985 (59.4%) 4208 (2.3%) 29447 (16.4%)		0 (0.0%)

Table 2: Categorical features part 2

3.1.4 Month

Month has, as expected, 12 distinct values. Sales are very uniformly distributed – but as visible on the graph there are slight rises in sales in the two large school holiday periods, June and December. This is something the marketing team needs to be aware of to ensure they have enough stock during these times. They can also potentially run promotions to maximise these times for profit.

3.1.5 Days

Days are uniformly distributed and have no clear trends. The only small rise in sales is right at the end of the month which could be due to pay day traditionally being the last Friday of the month in South Africa.

3.1.6 Why bought?

Recommendation is by far the most common reason for sales. The quality of products needs to be upheld to ensure that clients keep on recommending their products to others. If the quality drops and people stop recommending it to each other it will have a drastic effect on sales. For this reason, the company needs to improve their website (which is the second largest reason) to get more sales from non-variable reasons in the case a bad batch is produced or raw materials are of lower quality.

Staff currently allocated to emails and spam can be reallocated to the website as emails and spam together do not even contribute as much as the third lowest reason for purchase.

3.2 Continuous features

Information that is continuous in nature can have almost any value (Philips, 2013). They are numerical and often decimal. The continuous features in this dataset are summarised as followed:

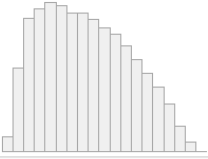
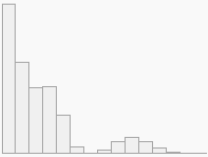

Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
AGE [integer]	Mean (sd) : 54.6 (20.4) min ≤ med ≤ max: 18 ≤ 53 ≤ 108 Q1 - Q3 : 38 - 70	91 distinct values		0 (0.0%)
Delivery.time [numeric]	Mean (sd) : 14.5 (14) min ≤ med ≤ max: 0.5 ≤ 10 ≤ 75 Q1 - Q3 : 3 - 18.5	148 distinct values		0 (0.0%)
Price [numeric]	Mean (sd) : 12294.1 (20889.2) min ≤ med ≤ max: 35.6 ≤ 2259.6 ≤ 116619 Q1 - Q3 : 482.3 - 15271	78832 distinct values		0 (0.0%)

Table 3: Continuous features

3.2.1 Age

Age has a normal distribution skewed right (Yi, 2021). This indicates that there is a larger collection of young clients. There are 91 different ages between the values of 18 and 108 with the average age of clients being 54.6 with a standard deviation of 20.4. The first quartile is at 38 which provides a good idea of the age market to target as the distribution is skewed right. There are no missing values in the dataset for age.

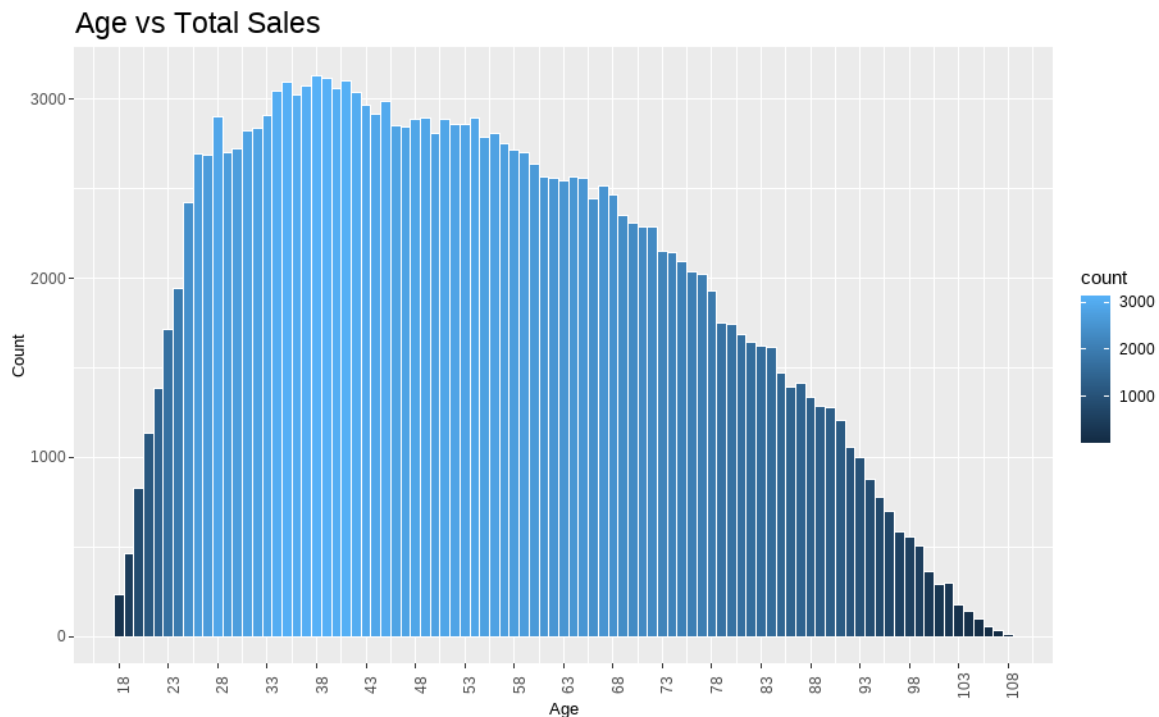


Figure 5: Enlarged distribution of Age

Looking at the enlarged graph ages between 33 and 43 are the best age group to target. . The sales team can either aim to target this age group in the future to increase the number of purchases they make, or they can focus their new marketing on the senior population to strive for the same number of sales in other age groups and thus create a more uniform distribution.

3.2.1.1 Age in each class

To better understand the distribution of sales across ages the distribution per class can be analysed:

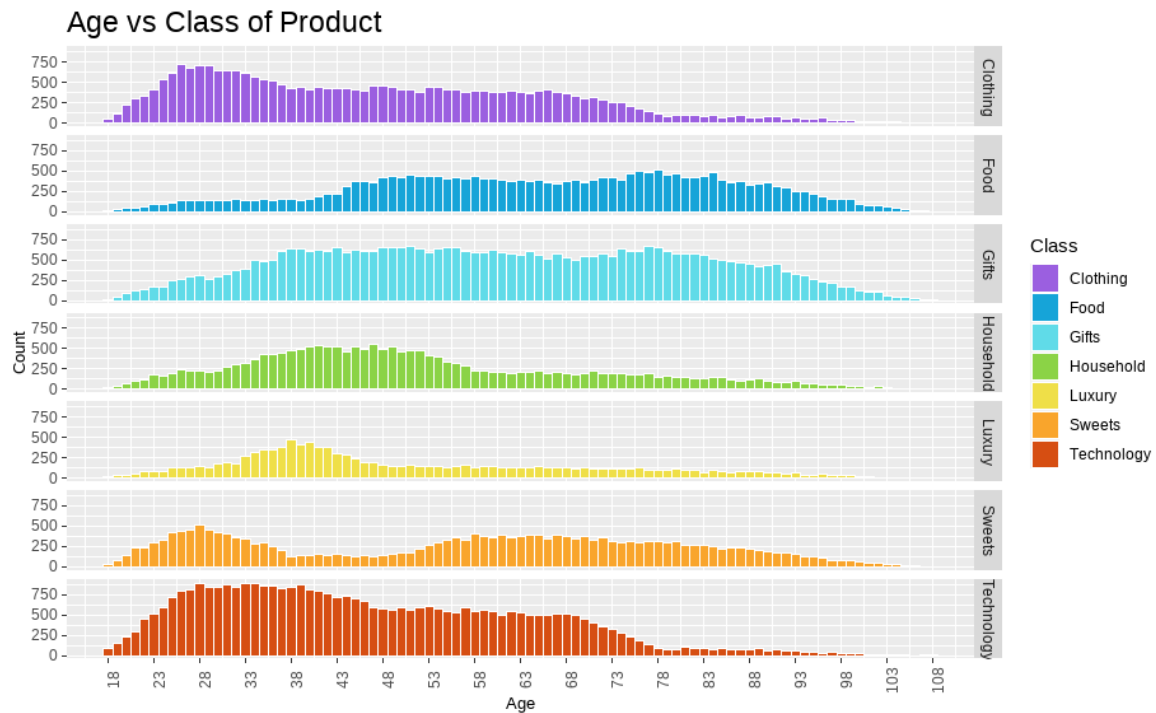


Figure 6: Age distribution per class

Clothing is mostly uniformly distributed due to clothes being an essential product for all age groups. There is, however, an increase of clothing sales between 25 and 35. This could be due to young adults who have a desire to wear the latest fashion starting to earn their first salaries. As clients age increase, they wear a piece of clothing longer and thus buy new clothing less frequently.

Food is uniformly distributed from 18 to 38. It then nearly doubles and remains the same until very old age. This is most probably due to couples having kids that need to eat increasingly more as they age. These couples then need to buy food for the whole family.

Household items are mostly bought between 28 and 58 and normally distributed in this time. This is most probably due to couples with families moving into larger houses to accommodate kids, pets etc. when they have saved up money and then furnishing it.

Luxury items are normally distributed between 28 and 48 and are almost exclusively bought in this range. This could be due to the purchase of engagement rings once couples have enough money to get married.

Sweets are normally distributed between 18 and 38 as young people with fast metabolisms enjoy sweets. Moving with the assumption made at Luxury purchases these clients then stop buying sweets around the time they get married. However, as they start having kids and then later grandkids, they start buying sweets for them again around the age of 55.

Technology is bought mostly by young age groups, peaking at 28 and then slowly tapering down as generations have less and less knowledge about technology. From 78 to 108 very few technology sales are made.

3.2.2 Delivery Time

Delivery time has a large range from half a day up to 75 days. However, both the median and the mean are much lower than the max at 10 and 14.5 respectively. According to the simplified graph there is a distinct initial downward trend in deliveries time initially before rising later in to something that resembles a normal distribution. To get a better understanding of the multimodal distribution an enlarged graph is required.

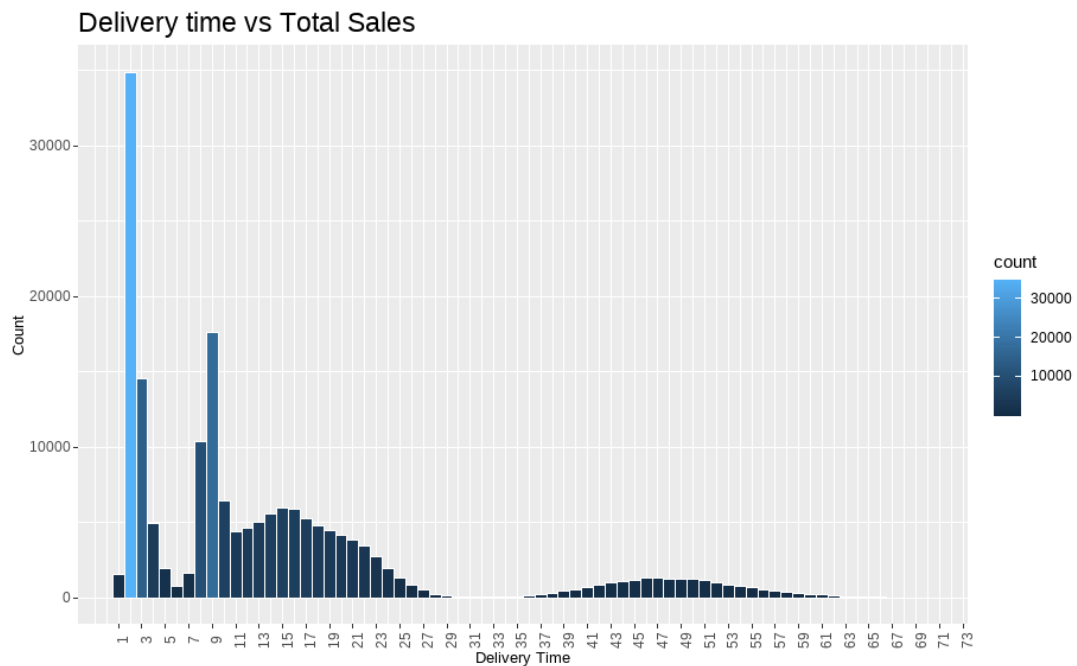


Figure 7: Enlarged distribution of delivery time

Delivery time has a normal distribution between 35 and 61. It has a large spike at 2 days with this being a popular delivery option for clients. It has a local minimum at 6. There is a negative trend in deliveries from 15 days to 35.

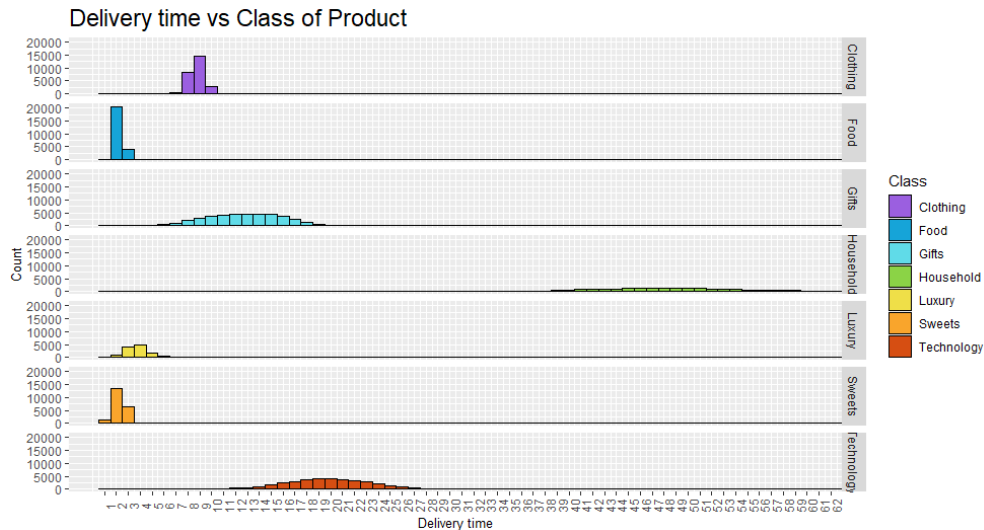


Figure 8: Delivery time per class

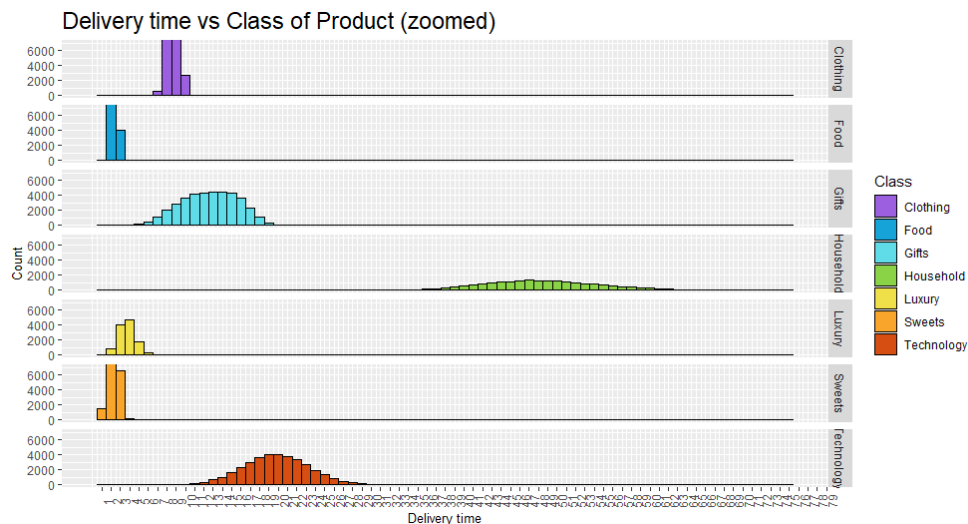


Figure 9: Delivery time per class y-axis shortened

Looking at the delivery time per class graph however it is clear that most delivery times are normally distributed per class and that the various classes are the cause of the irregular distribution for all deliveries.

Delivery times of household items take much longer than most other classes and is not distributed well for control. A new delivery method or courier should be considered for faster and more reliable delivery as Household items are the singular culprit for the second wave of deliveries in the total graph.

Technology has a near perfect normal distribution which makes control very easy but there should be an effort made to reduce the mean currently at 18 days, which is a long time for consumers to wait and might lead to customer loyalty problems.

3.2.3 Price

Price has an exponential distribution with most products being low priced. This is natural due to food, sweet and clothes being represented on the same graph as luxury and technology items. Due to the extremely high cost of luxury items reaching all the way to R116 619 the mean price of R12 294.1 is not an accurate metric to use to for expected price per sale. The median of R2259 is a better indication of usual sales.

3.2.3.1 Price vs Class

Price is thus better analysed by looking at it per class:

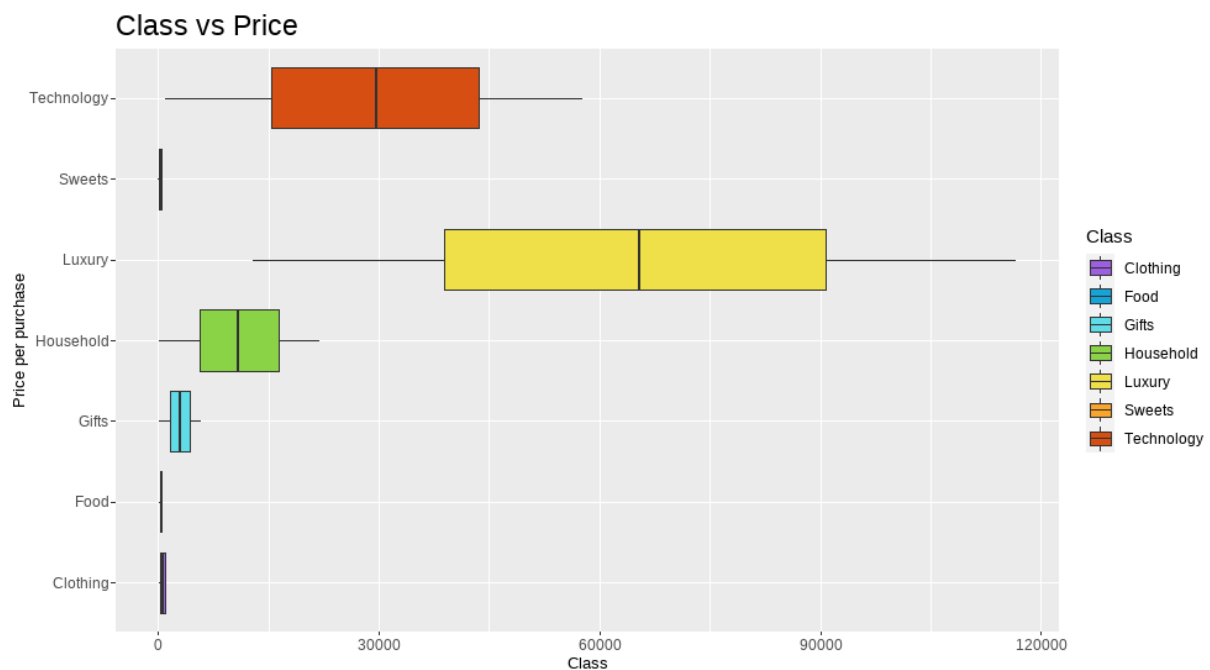


Figure 10: Class vs Price

As expected, luxury items are by far the most expensive products. It has a much larger range than any other class and stretches over a span of almost R100 000 with a mean at around R65000. Technology is the second most expensive class and has a high mean of R30 000. Low cost phone options should be added that has significant market in South Africa although it would drop the mean price it should make up for it in volume sold. Household items all range between very cheap and R22 500 and is comfortably the 3rd largest Class.

Sweets, Food, Clothing and Gifts are the 4 least expensive classes and to better visualize them a zoomed in version of the graph was created:

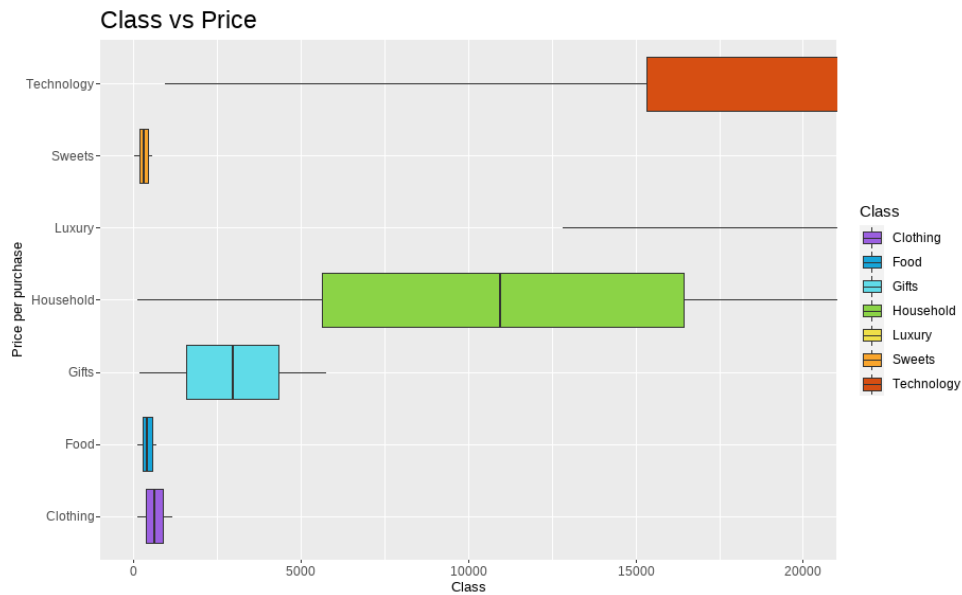


Figure 11: Class vs Price zoomed

Food and sweets are the 2 least expensive items with very predictable prices and a small range. Clothing has a slightly larger range due to brand name clothing being much more expensive than basic clothing. Gifts are on average more expensive than the above mentioned 3 classes but has a large range. This is expected because a multitude of items can be classified as gifts.

3.3 Scatter Plot Matrix (SPLOM)

A scatter plot matrix is useful to identify potential correlations between continuous features on a large scale to ensure the correct correlations are investigated further.

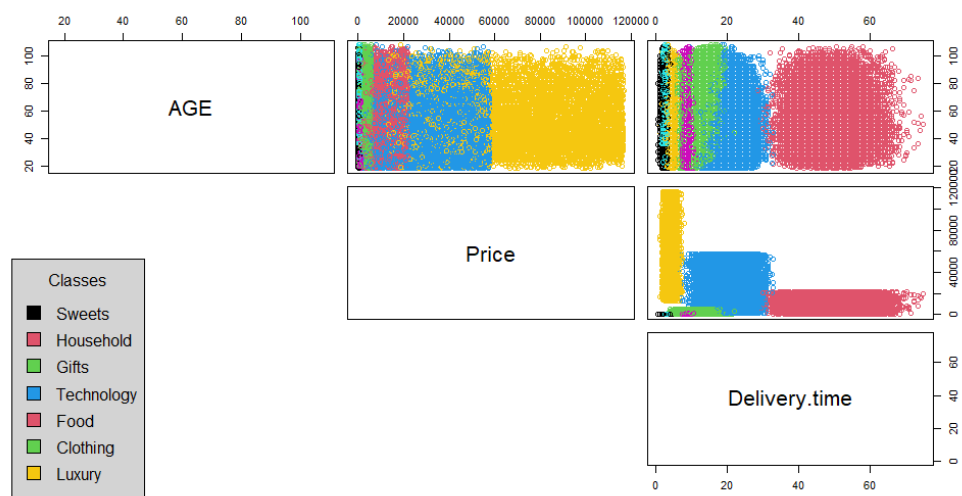


Figure 12: Scatterplot Matrix for continuous features

The only compartment of the Scatter Plot Matrix that shows meaningful separation is the Price vs Delivery time scatterplot. It could be beneficial to analyse this further:



Figure 13: Delivery Time vs Price scatterplot

The scatterplot confirms that there is good separation between the classes and clearly shows that between the classes Luxury, Technology and Household the classes with a more expensive mean are faster to deliver. However, within an isolated class there is no meaningful correlation between the price of a product and its delivery time as they are all rectangular shaped.

When the classes are ignored and an attempt is made to identify a trend between the price of any given product and the delivery time of any given product, there is none. To test this a regression line was fitted:



Figure 14: Delivery time vs Price + Regression line

The regression line is nearly flat and confirms that there is no meaningful relationship between the price of any given product and the delivery time of any given product. If a regression line is drawn per class the results would be an almost completely flat line due to the square shape of the scatterplot distribution.

3.4 Process Capability Indices

LSL	USL	Standard Deviation	Mean	CP	CPU	CPL	CPK
0	24	3.502	20.011	1.142	0.38	1.905	0.38

Table 4: Table of process capability indices

An LSL of 0 is logical because no delivery can occur faster than 0 days/hours. The mean of technology deliveries is already close to the USL, although it has a low standard deviation. To better understand the performance of the delivery, process the indices need to be reviewed.

3.4.1 Potential Capability

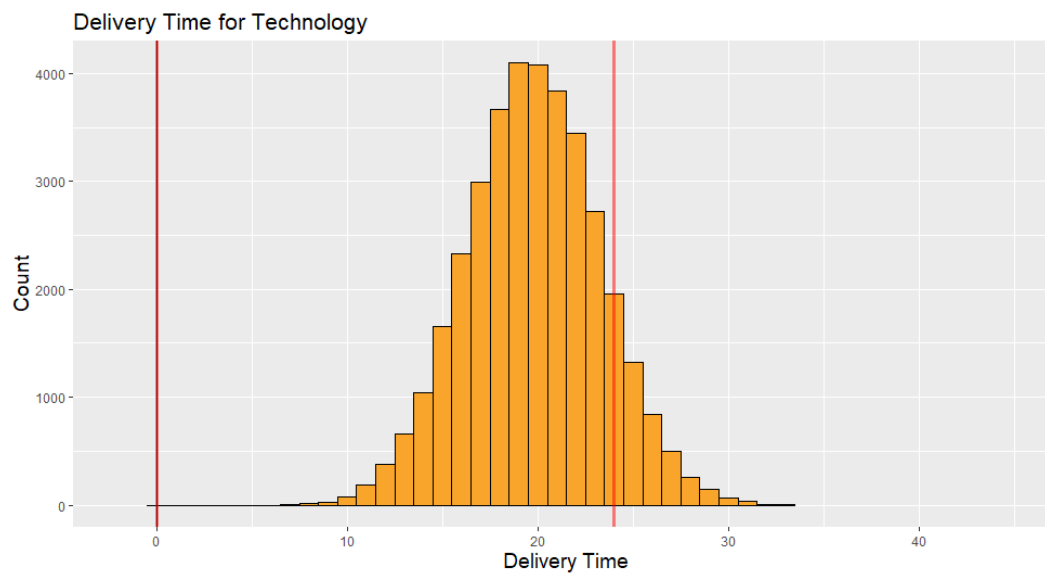


Figure 15: Delivery time for technology

An indication of how the distribution stacks up against the standard width is the Cp (Process Capability Ratio). Since there are multiple instances, the CP demonstrates the process's capacity. Technology can be delivered within the required limits.

If there is conformity to the specifications, it is shown by the Cpk value (process capability index). A procedure with a low Cpk number may need improvement, whereas one with a high Cpk value is certain to be more complete. The process is not centered between the prescribed limits as seen by the CPK being less than the CP. This demonstrates how moving the mean to the left could enhance the process.

Many sectors use the benchmark Cpk of 1.33 to evaluate the process capability. As the company's Cpk value is significantly lower than this benchmark value. There is thus potential for improvement in the company's processes.

4 Statistical Process Control

The data is ordered according to date and an X- and S chart for delivery times is constructed using the first 30 samples that contain 15 Sales each. The control levels of the X-charts are monitored and the validity of the conclusions are discussed with reference to the S-charts. Thereafter the process is repeated but for all samples.

4.1 Tables

4.1.1 Values for S-chart

Class	UCL	UCL2	UCL1	CL	LCL1	LCL2	LCL
Clothing	0.866559568463719	0.761458902227527	0.656352723525888	0.551246544824249	0.446145878588057	0.341039699886418	0.235933521184778
Household	7.34418006586244	6.45344128001172	5.56265577545829	4.67187027090486	3.78113148505414	2.89034598050071	1.99956047594728
Food	0.437246583672721	0.384215137334908	0.331180909530279	0.278146681725649	0.225115235387837	0.172081007583207	0.119046779778578
Technology	5.18056970372824	4.55224437312212	3.92388608723798	3.29552780135385	2.66720247074772	2.03884418486358	1.41048589897945
Sweets	0.835339146409308	0.73402504866478	0.632705637057997	0.531386225451214	0.430072127706685	0.328752716099902	0.22743330449312
Gifts	2.24633333311156	1.97388682338443	1.70142602400505	1.42896522462567	1.15651871489854	0.884057915519165	0.611597116139788
Luxury	1.51105176847233	1.32778387395774	1.14450636715455	0.961228860351357	0.777960965836767	0.594683459033574	0.411405952230381

Table 5: Values for S-chart

4.1.2 Values for X-chart

Class	UCL	UCL2	UCL1	CL	LCL1	LCL2	LCL
Clothing	9.40493352386633	0.761458902227527	0.656352723525888	8.97	8.82502215871122	8.68004431742245	8.53506647613367
Household	50.2483278659662	6.45344128001172	5.56265577545829	46.5622222222222	45.3335203409742	44.1048184597263	42.8761165784783
Food	2.70945773188154	0.384215137334908	0.331180909530279	2.49	2.41684742270615	2.34369484541231	2.27054226811846
Technology	22.9746158797126	4.55224437312212	3.92388608723798	20.3744444444444	19.5077206326884	18.6409968209323	17.7742730091763
Sweets	2.89704150965879	0.73402504866478	0.632705637057997	2.47777777777778	2.33802320048411	2.19826862319044	2.05851404589677
Gifts	9.48856467334077	1.97388682338443	1.70142602400505	8.36111111111111	7.98529325703456	7.60947540295801	7.23365754888145
Luxury	5.49396512637278	1.32778387395774	1.14450636715455	4.73555555555556	4.48275236528315	4.22994917501074	3.97714598473833

Table 6: Values for X-chart

4.2 First 30 samples

To get an initial understanding of the data and to see if it is under control only the first 30 are plotted for each class.

4.2.1 Luxury

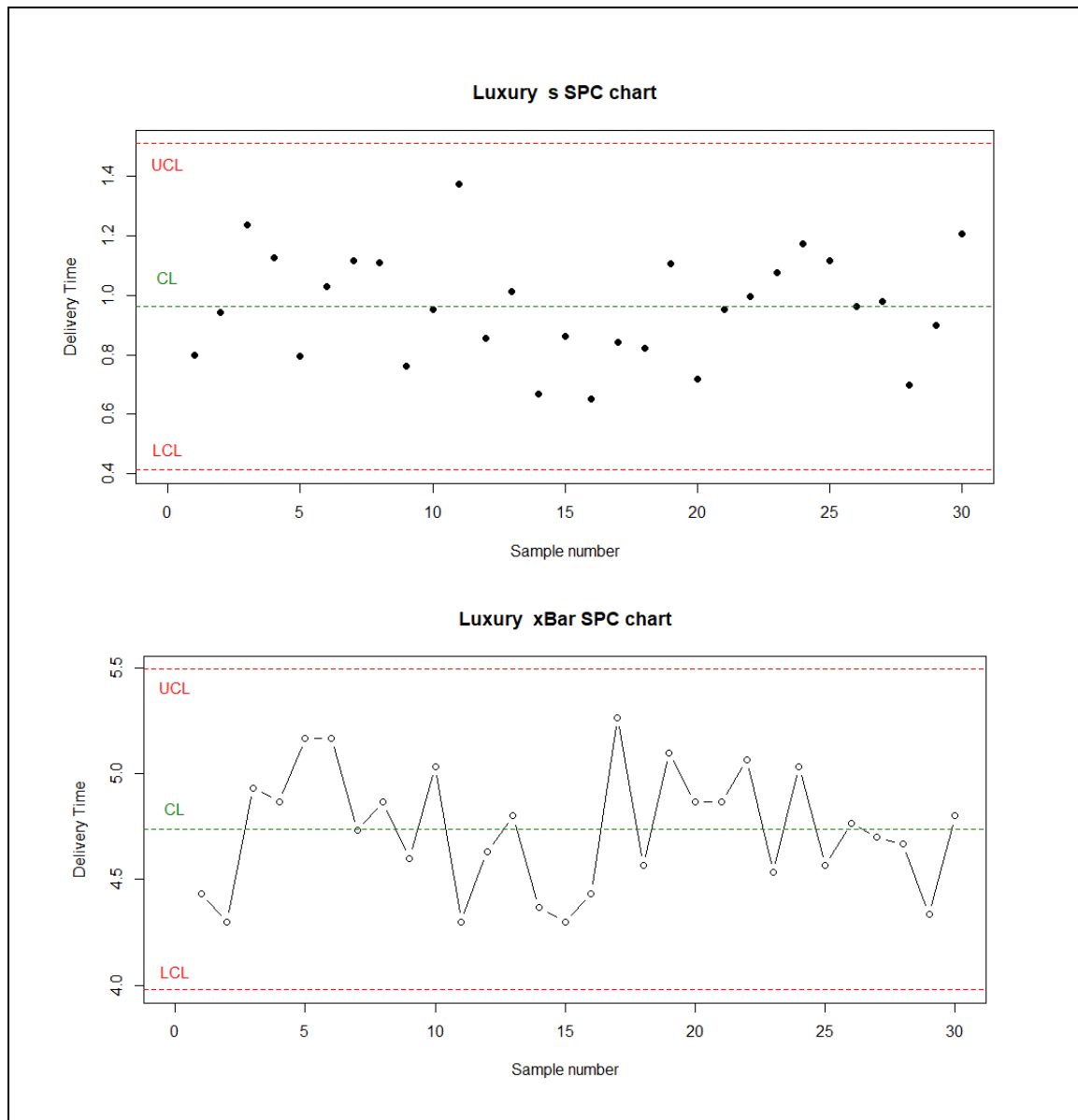


Figure 16: S- and X charts for first 30 Luxury samples

The xBar chart does not spike outside of the top and lower control boundaries for the first 30 samples. The Luxury class can thus be identified as under control. This indicates that there is no cause of variation in the Luxury ordering procedure. The S-bar chart is satisfactory and thus the X-bar chart can be assessed.

4.2.2 Gifts

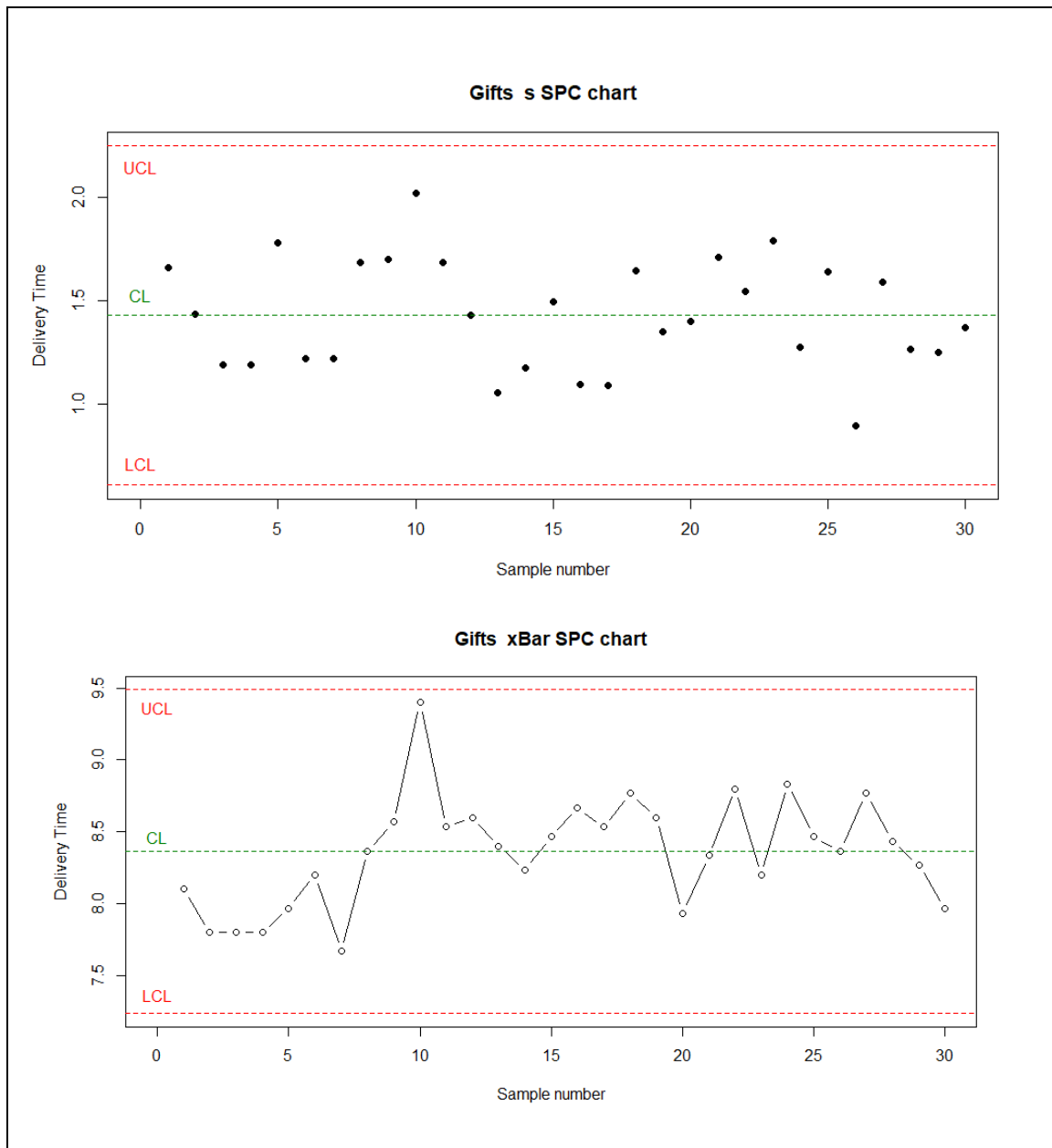


Figure 17: S- and X charts for first 30 Gifts samples

The xBar chart does not spike outside of the top and lower control boundaries for the first 30 samples. The Gifts class can thus be identified as under control. This indicates that there is no cause of variation in the Gifts ordering procedure. The S-bar chart is satisfactory and thus the X-bar chart can be assessed.

4.2.3 Sweets

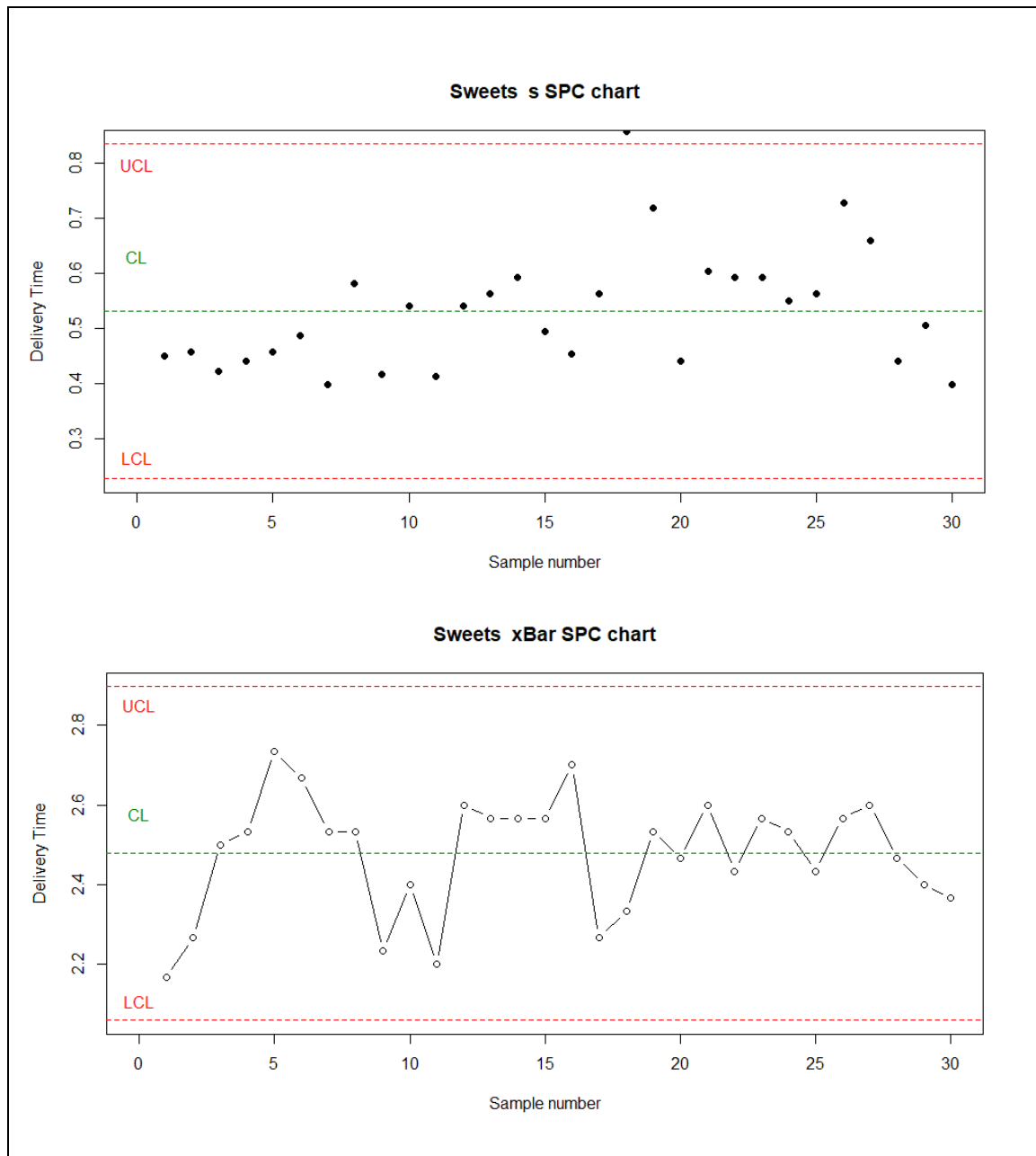


Figure 18: S- and X charts for first 30 Sweet samples

The first 30 samples show that the sweets class is under control, with the exception of sample 18 – which standard deviation is out of control limits. There is, however, still little cause of variation in the Sweets ordering procedure.

4.2.4 Technology

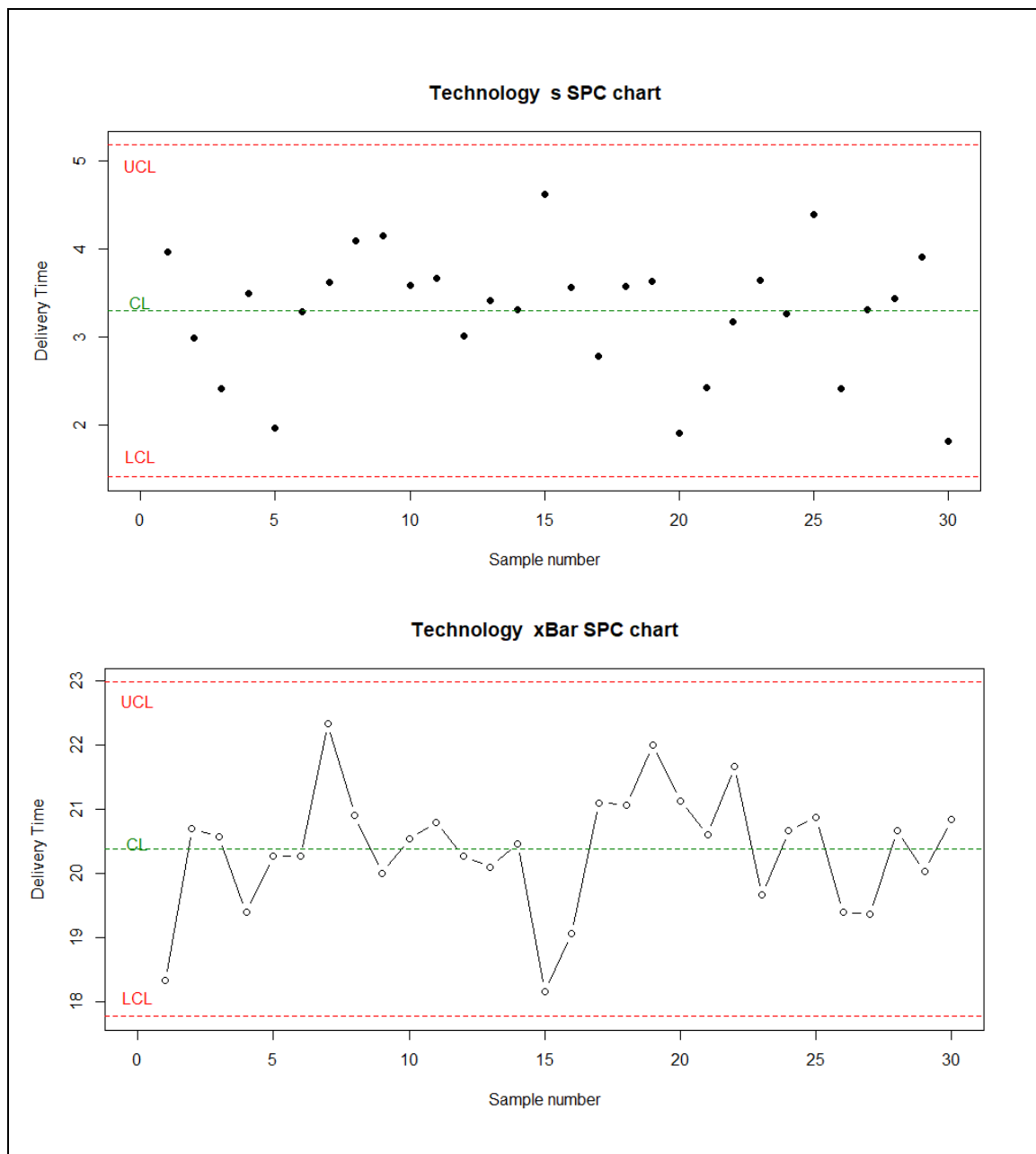


Figure 19: S- and X charts for first 30 Technology samples

The xBar chart does not spike outside of the top and lower control boundaries for the first 30 samples. The Technology class can thus be identified as under control. This indicates that there is no cause of variation in the Technology ordering procedure. The S-bar chart is satisfactory and thus the X-bar chart can be assessed.

4.2.5 Food

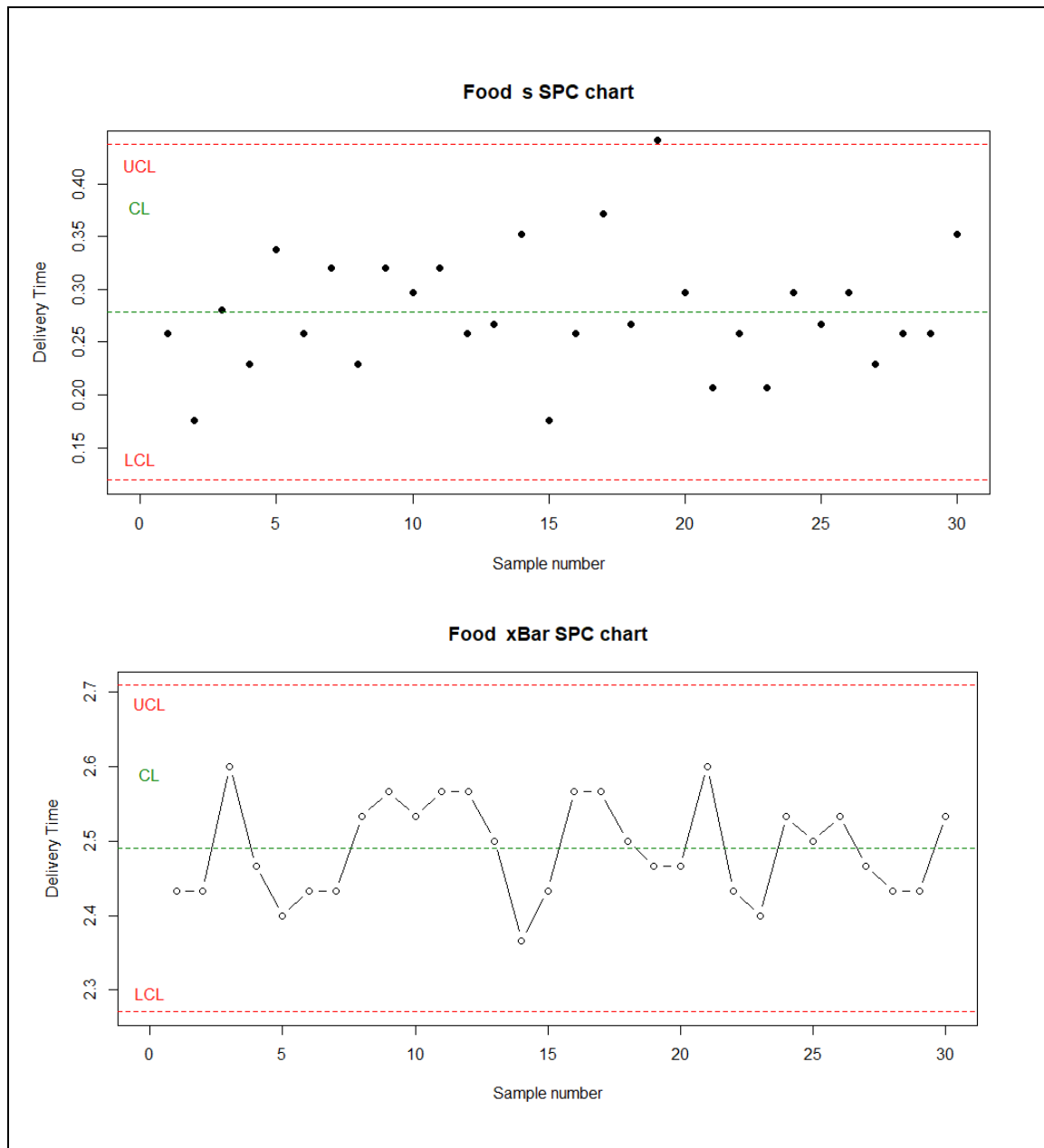


Figure 20: S- and X charts for first 30 Food samples

The first 30 samples show that the sweets class is under control, with the exception of sample 19 – which standard deviation is out of control limits. There is, however, still little cause of variation in the Sweets ordering procedure.

4.2.6 Household

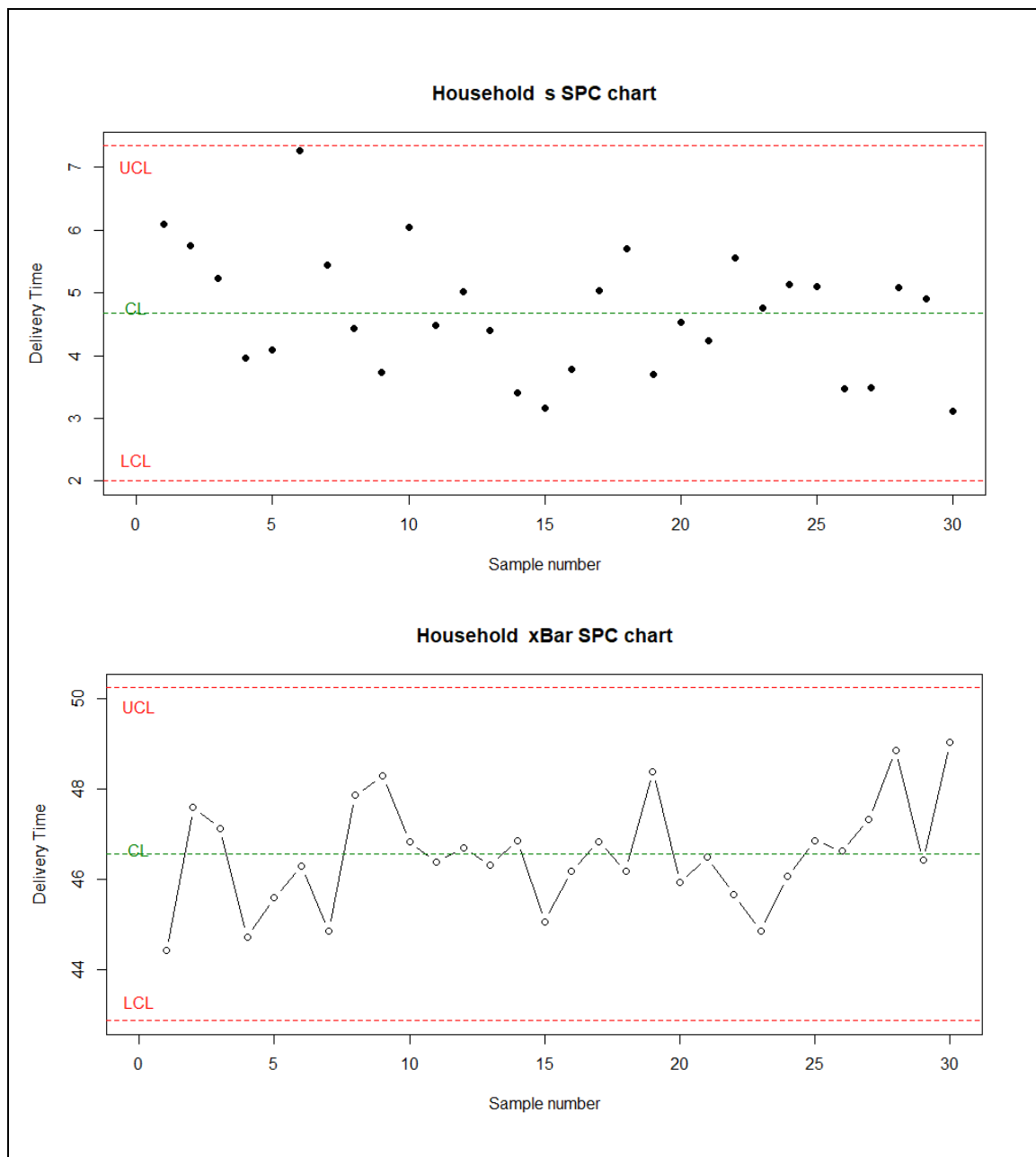


Figure 21: S- and X charts for first 30 Household samples

The xBar chart does not spike outside of the top and lower control boundaries for the first 30 samples. The Household class can thus be identified as under control. This indicates that there is no cause of variation in the Household ordering procedure. The S-bar chart is satisfactory and thus the X-bar chart can be assessed.

4.2.7 Clothing

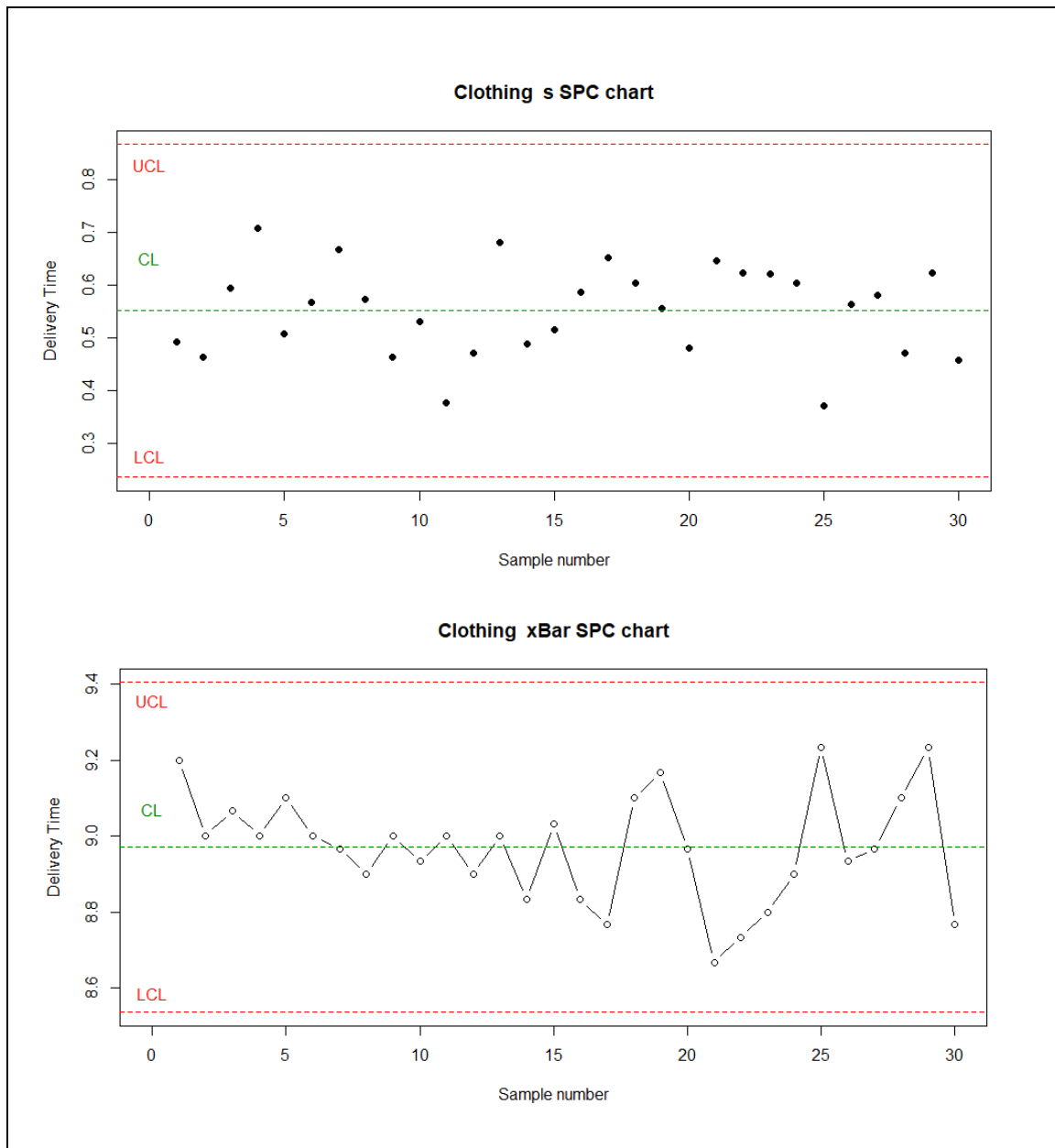


Figure 22: S- and X charts for first 30 Clothing samples

The xBar chart does not spike outside of the top and lower control boundaries for the first 30 samples. The Clothing class can thus be identified as under control. This indicates that there is no cause of variation in the Clothing ordering procedure. The S-bar chart is satisfactory and thus the X-bar chart can be assessed.

4.3 All Samples

All samples are then charted to identify if the level of control has changed with time.

4.3.1 Luxury

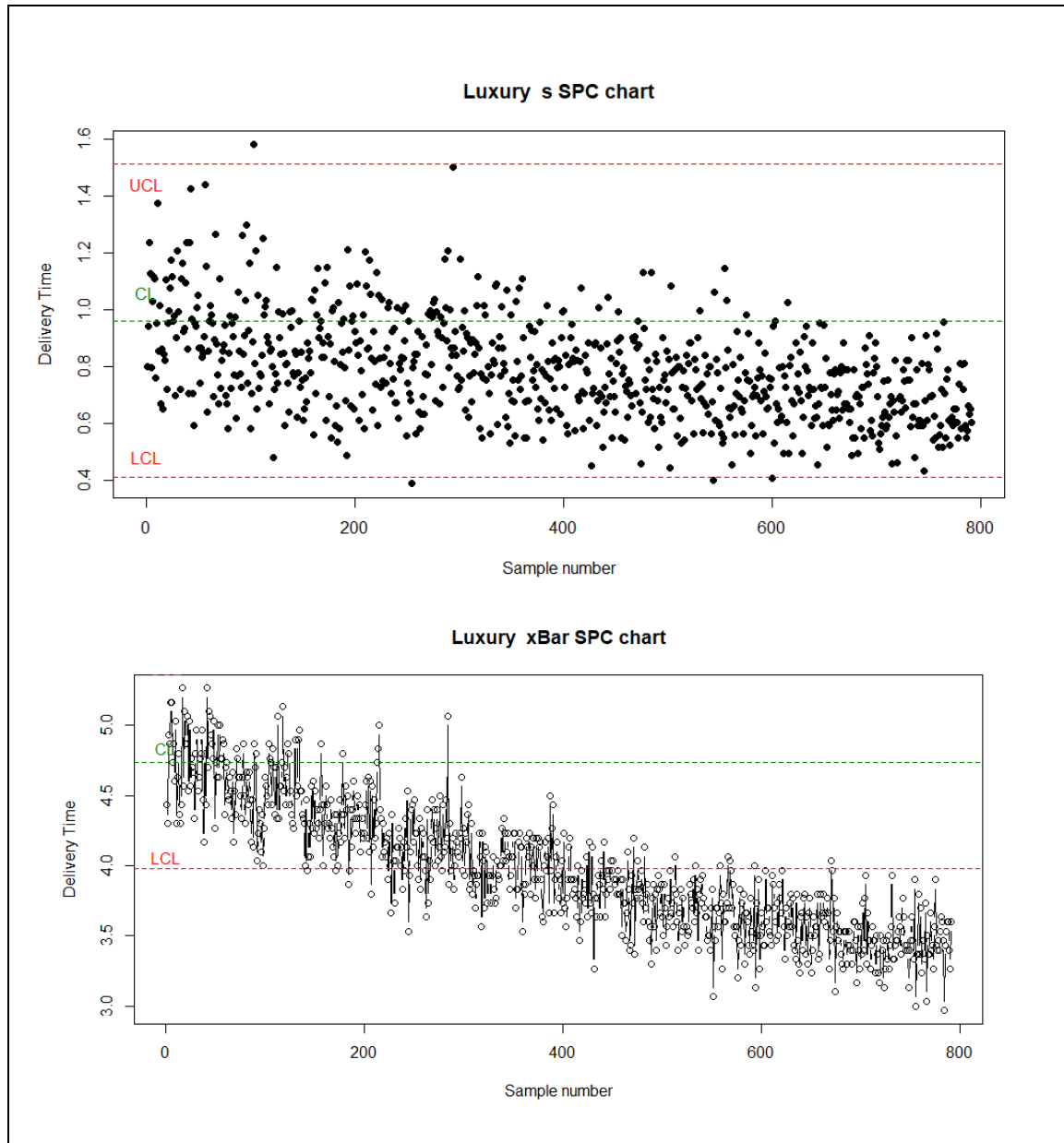


Figure 23: S- and X charts for all Luxury samples

Luxury products' delivery time is consistently decreasing. Since around the 200 sample mark Luxury items have consistently been outside of the control limits. The type of clients that buy luxury items may be insisting on fast delivery as they are typically higher class individuals who are used to fast service. The company might be going out of their way to make these customers happy to ensure they remain loyal customers. The sales department should investigate if this is the reason or if there is another. The S-bar chart has only 2 out of 800

samples outside the control limits and thus the conclusions made regarding the X-Chart is valid.

4.3.2 Gifts

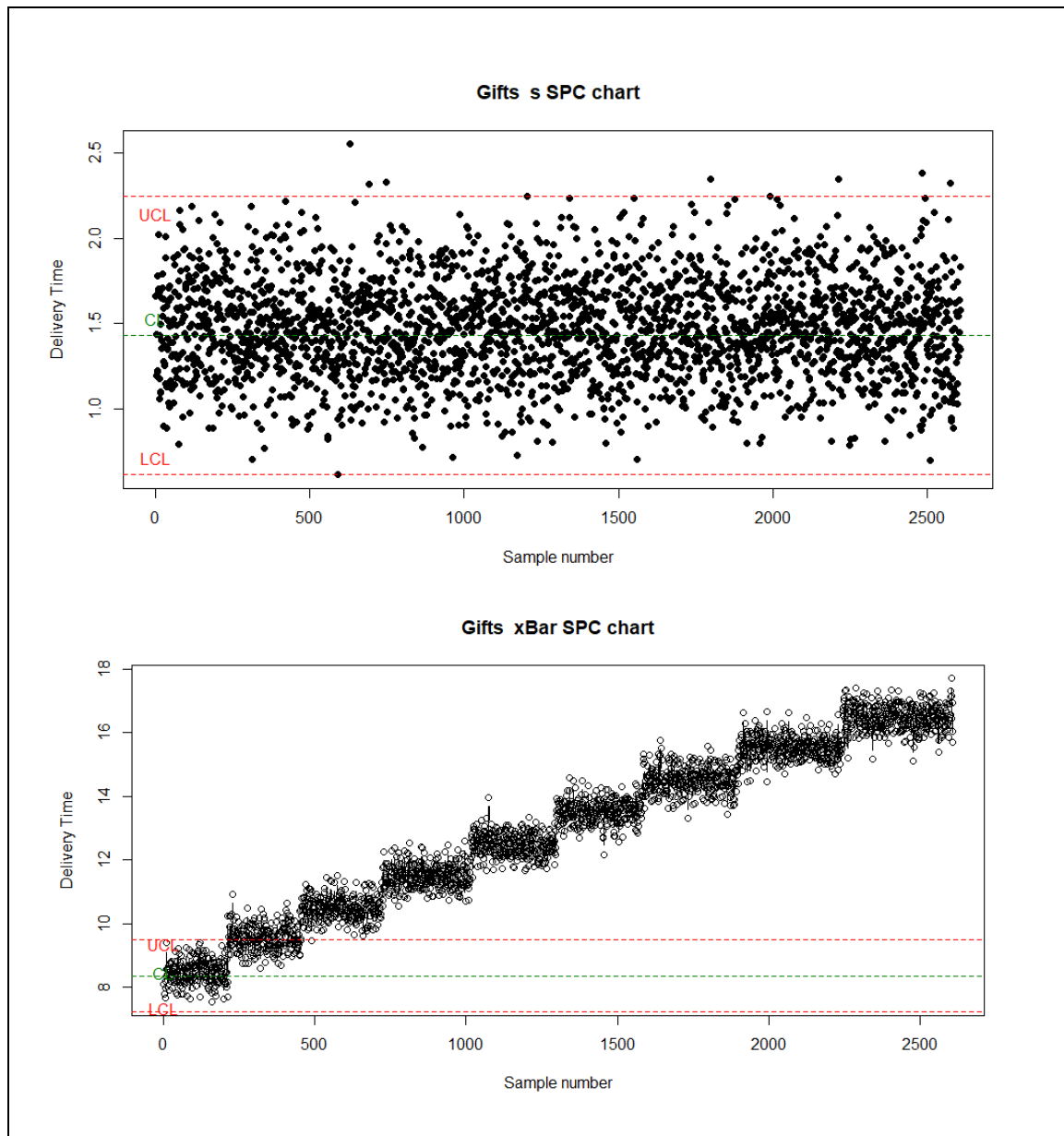


Figure 24: S- and X charts for all Gift samples

The delivery time of gifts is rapidly increasing and is an urgent cause of concern. The reason for this should be investigated as soon as possible. It has reached almost as far as double the upper control limit. The delivery time for gifts is out of control and unstable. The demand for gifts is clearly outweighing the logistic capabilities of the company and an urgent plan needs to be made to correct this by looking at new couriers or another DC close to their most frequent customers. The S-chart only has 7 samples out of control limits which indicates the conclusions from the X-chart is valid.

4.3.3 Sweets

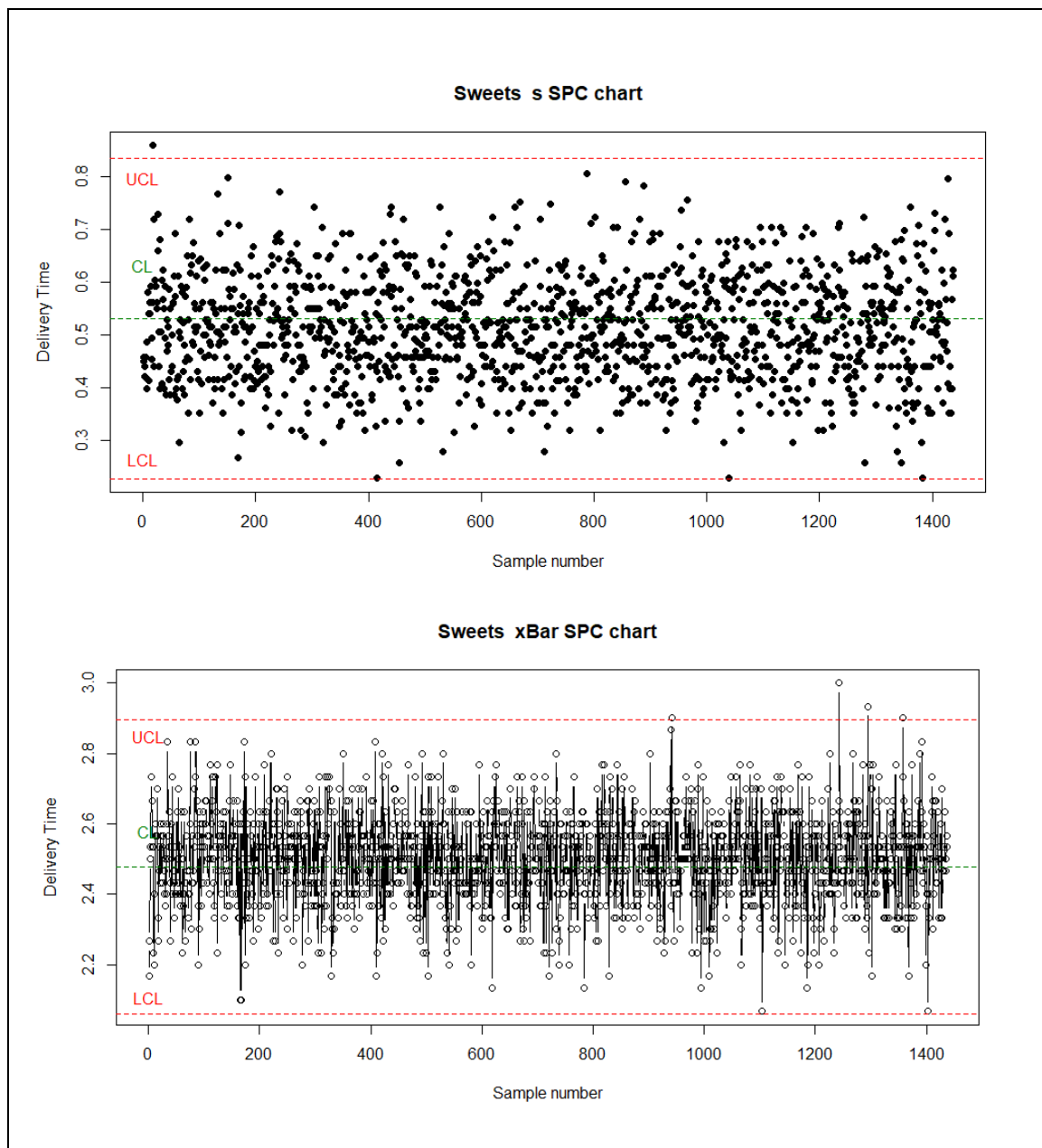


Figure 25: S- and X charts for all Sweet samples

Sweets seem to be under control with the exception of 2 or 3 samples that occurred recently. This should be investigated to ensure this trend does not continue, but is not a major area of concern as they also seem to be outliers. The S-chart only has one sample outside of control limits, and it occurred much earlier – so the conclusions from the X-chart is valid.

4.3.4 Technology

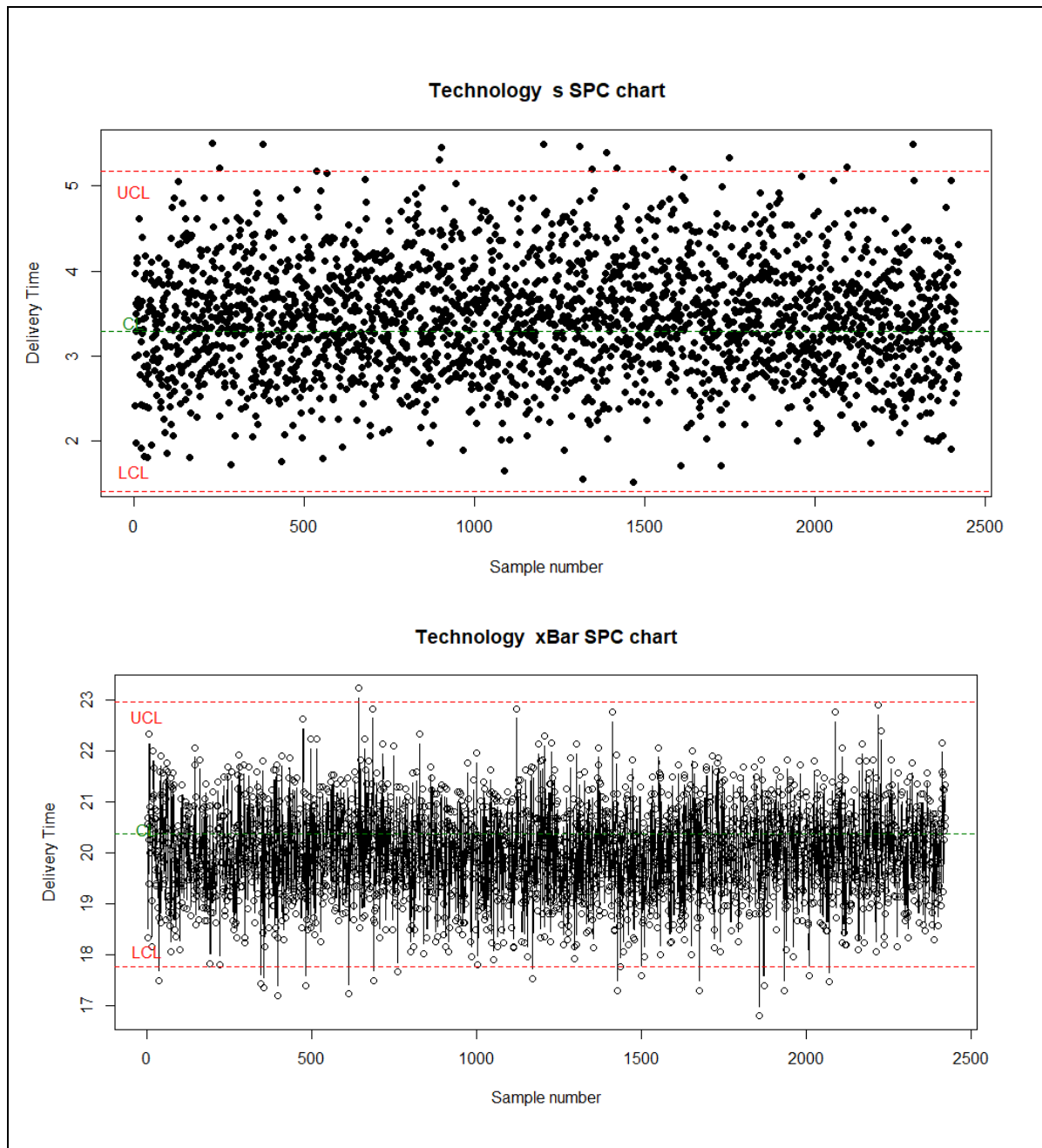


Figure 26: : S- and X charts for all Technology samples

Technology deliveries sometimes happen faster than the lower limit, but is not a major area of concern as the problem samples are fairly uniformly distributed across a large timeframe. There is no visible trend that the amount of problem samples will increase. The company should, however, still investigate why these samples occur. The S-chart has 18 samples outside of the control limits, which means the conclusions from the Xbar chart are valid, but are less reliable than some of the other X-charts.

4.3.5 Food

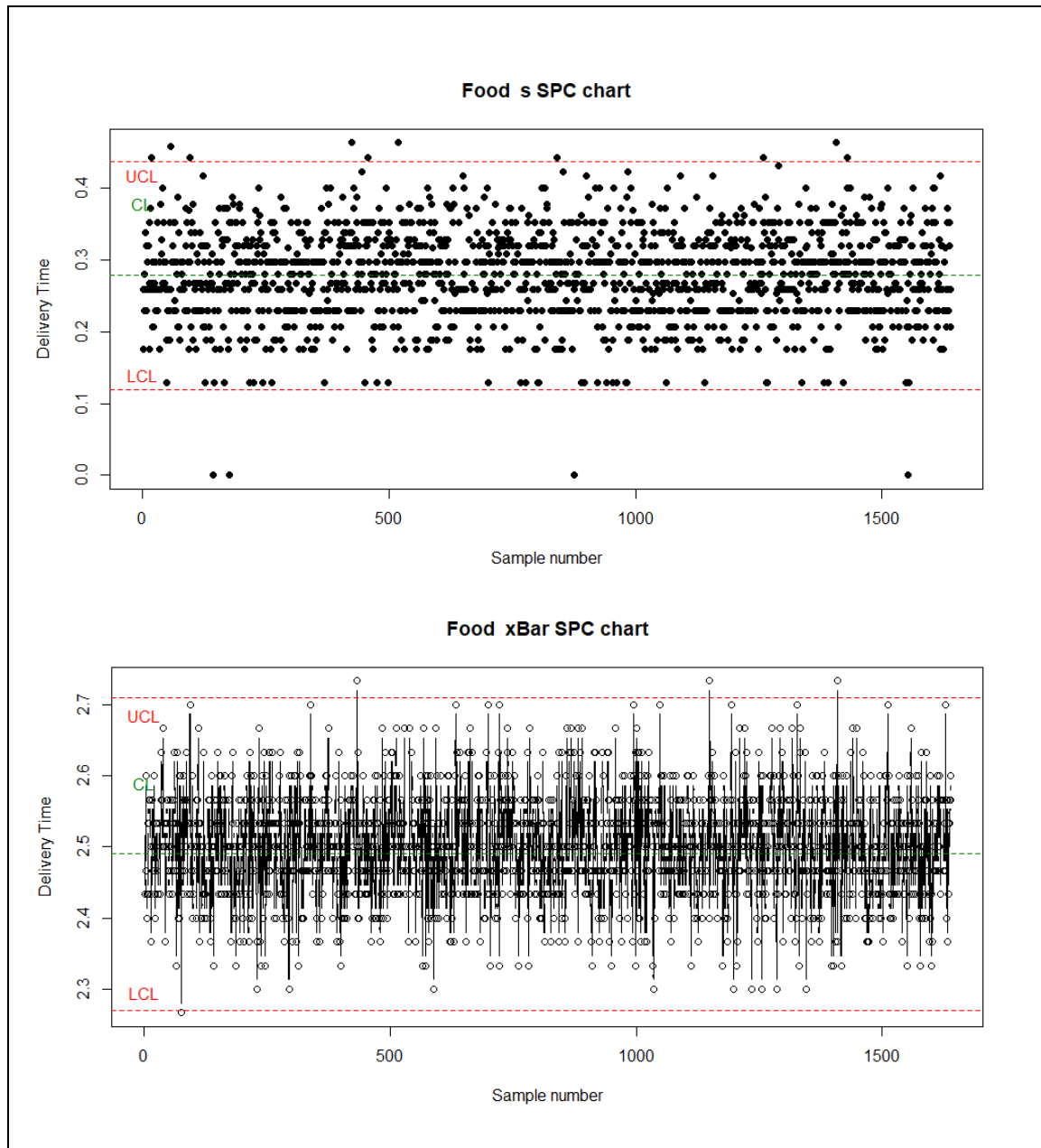


Figure 27: S- and X charts for all Food samples

Food is under good control, with only 3 samples on the X-bar chart occurring outside of the control limits. These samples are also randomly spread out across the timeframe of 9 years and has no visible trend of the problems occurring more often. The S-bar chart has only 14 samples out of the limits which indicates the conclusions made based on the Food X-bar chart are valid.

4.3.6 Household

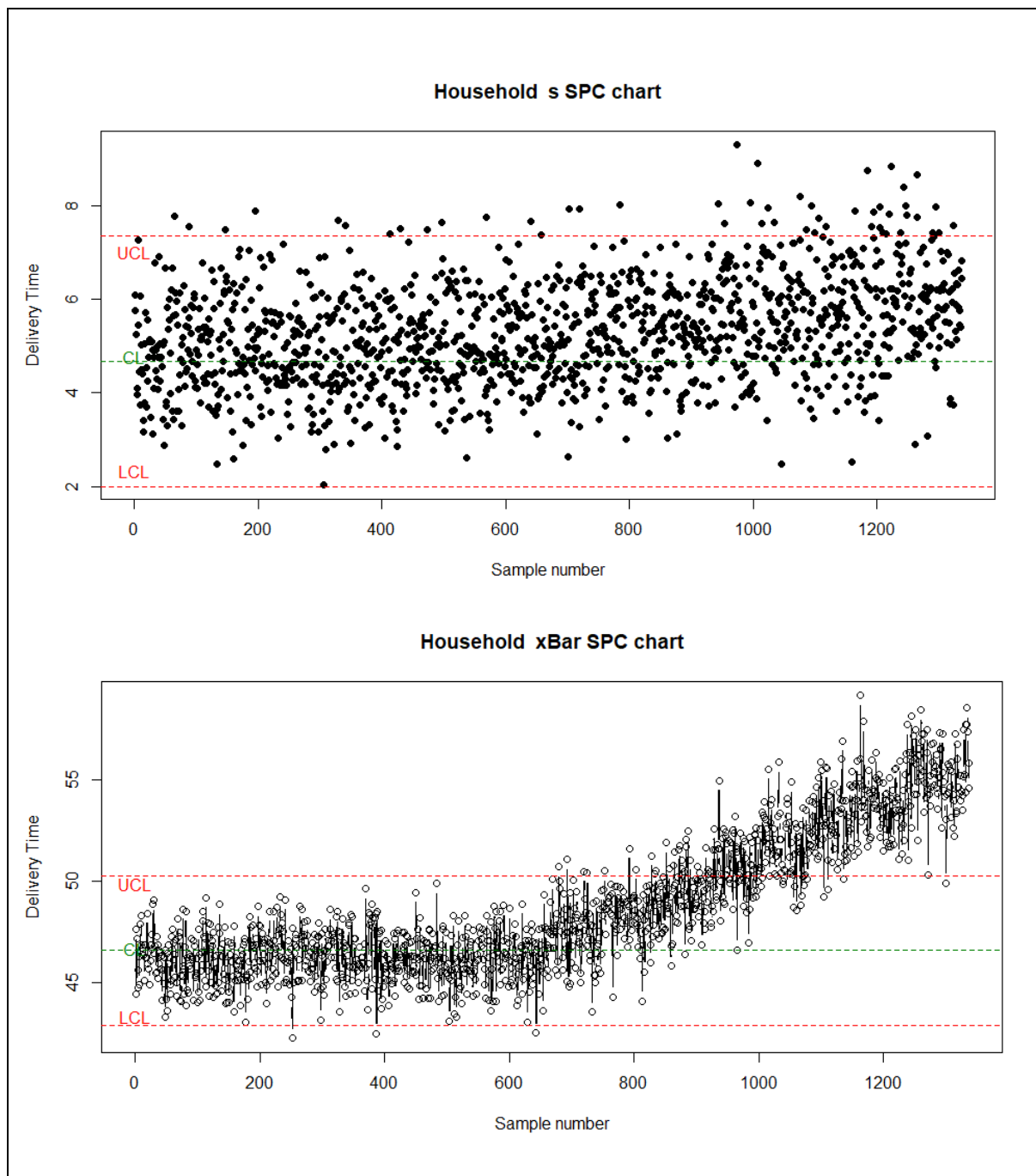


Figure 28: S- and X charts for all Household samples

The delivery times of household items are rising since the 650th sample and are starting to consistently be outside of the upper limit. This is a cause for concern and should be investigated. This could be due to large household items posing problems for couriers to handle or that household items are often out of stock, leading to a longer wait time before the courier can pick up the product. If not acted upon soon it could become much worse. Household items is already the class with the longest delivery time.

4.3.7 Clothing

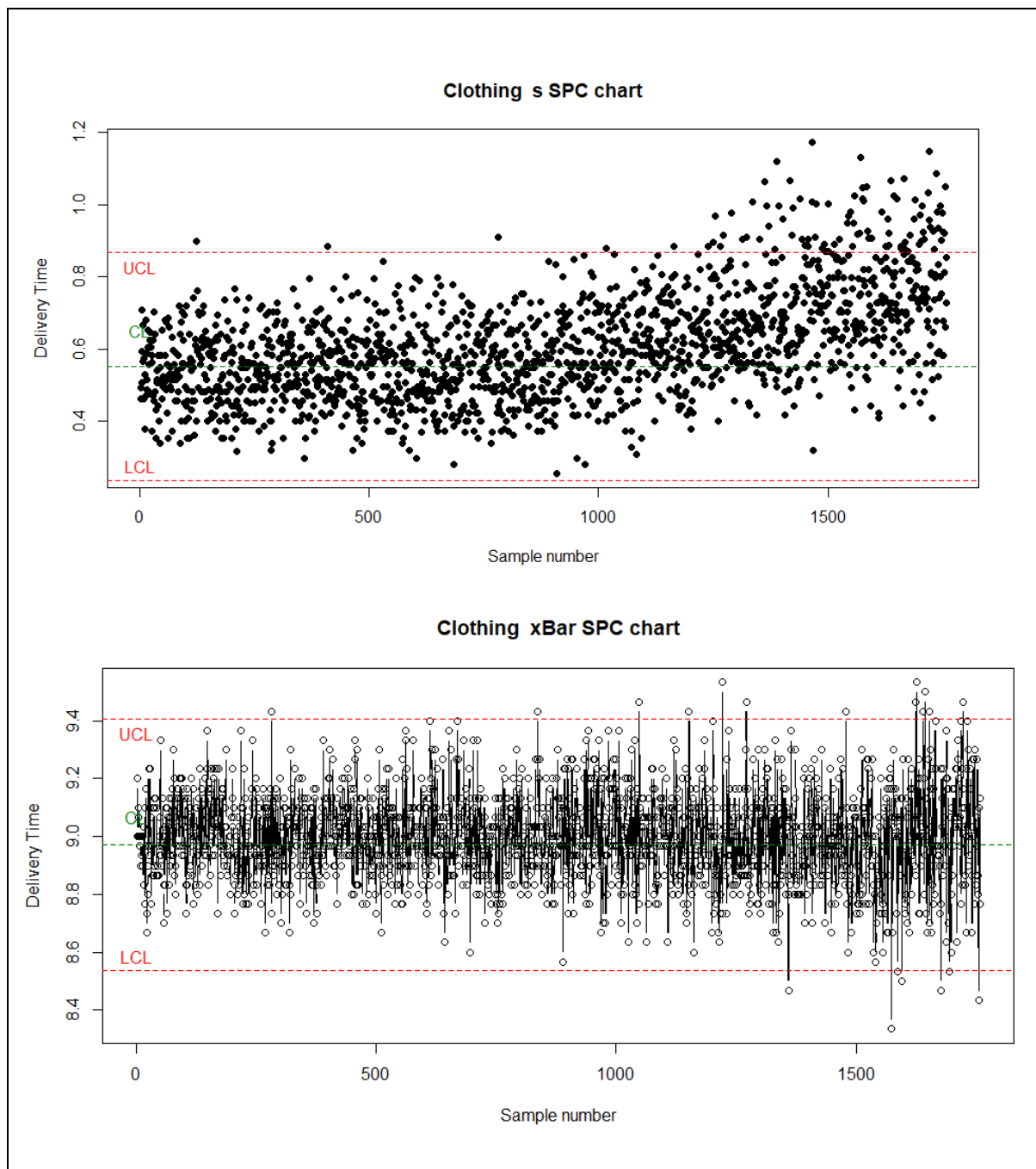


Figure 29: S- and X charts for all Clothing samples

The majority of X-bar chart samples are within the control limits, indicating that it is under control. However, the S-bar chart has recently started to have multiple samples outside of the control limits, which means the X-bar data could be unreliable. However, if the samples that are outside of the S-bar control limits is removed, the X-bar chart will still be within the limits. Thus the conclusions made from the X-bar chart is reliable.

5 Optimizing Delivery Process

5.1 Samples beyond control limits

	Clothing	Household	Food	Technology	Sweets	Gift	Luxury
# of Outliers	20	395	4	19	4	2287	440

Table 7: Number of outliers per class

Only a small number of samples in the Clothing, Food, Technology, and Sweets classes exceed the control limits. These classes are thus evidently being controlled well.

However, because there are so many samples that fall outside of the top and lower control limits in the Household, Luxury and Gift classes, it is clear that the delivery times of these classes are out of control. It is advised to investigate why the delivery times of these classes are not under control.

5.2 First 3 and last 3 samples outside control limits

The only groups plotted are Household, Gifts, and Luxury because they contain the highest percentages of samples outside the control boundaries. This suggests that nearly all deliveries will be anticipated to be on time for the classes of Technology, Clothing, Food, and Sweets.

The luxury, household, and gift classes cannot be expected to receive their deliveries on time. The first three and the final three samples that did not meet control standards are displayed in the ensuing figures.

The X-axis on each chart is manipulated to start at a sample close to the first important point. The sample at which the X-axis starts is indicated in the title of the plot.

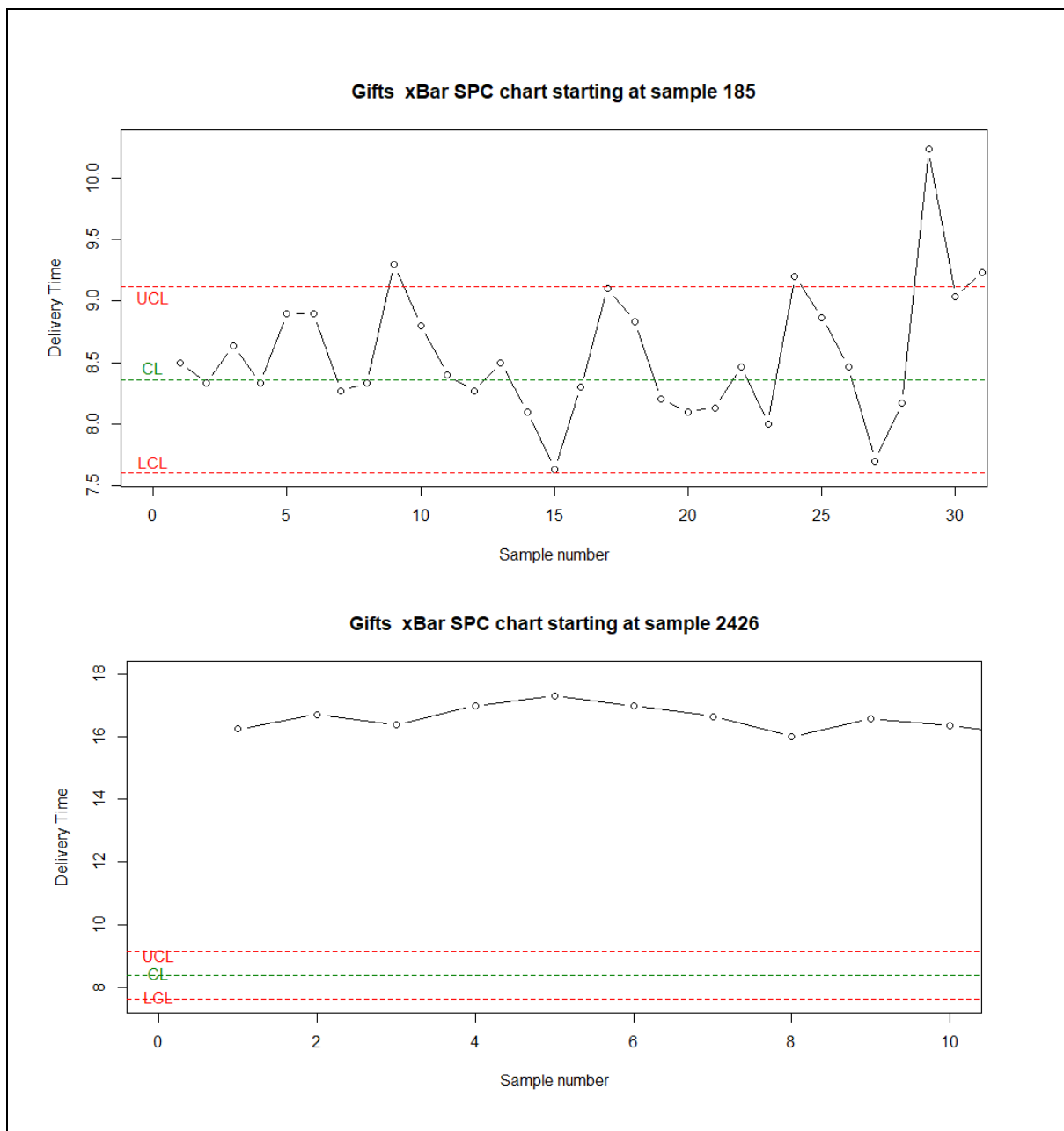


Figure 30: The first and last 3 outliers for Gifts

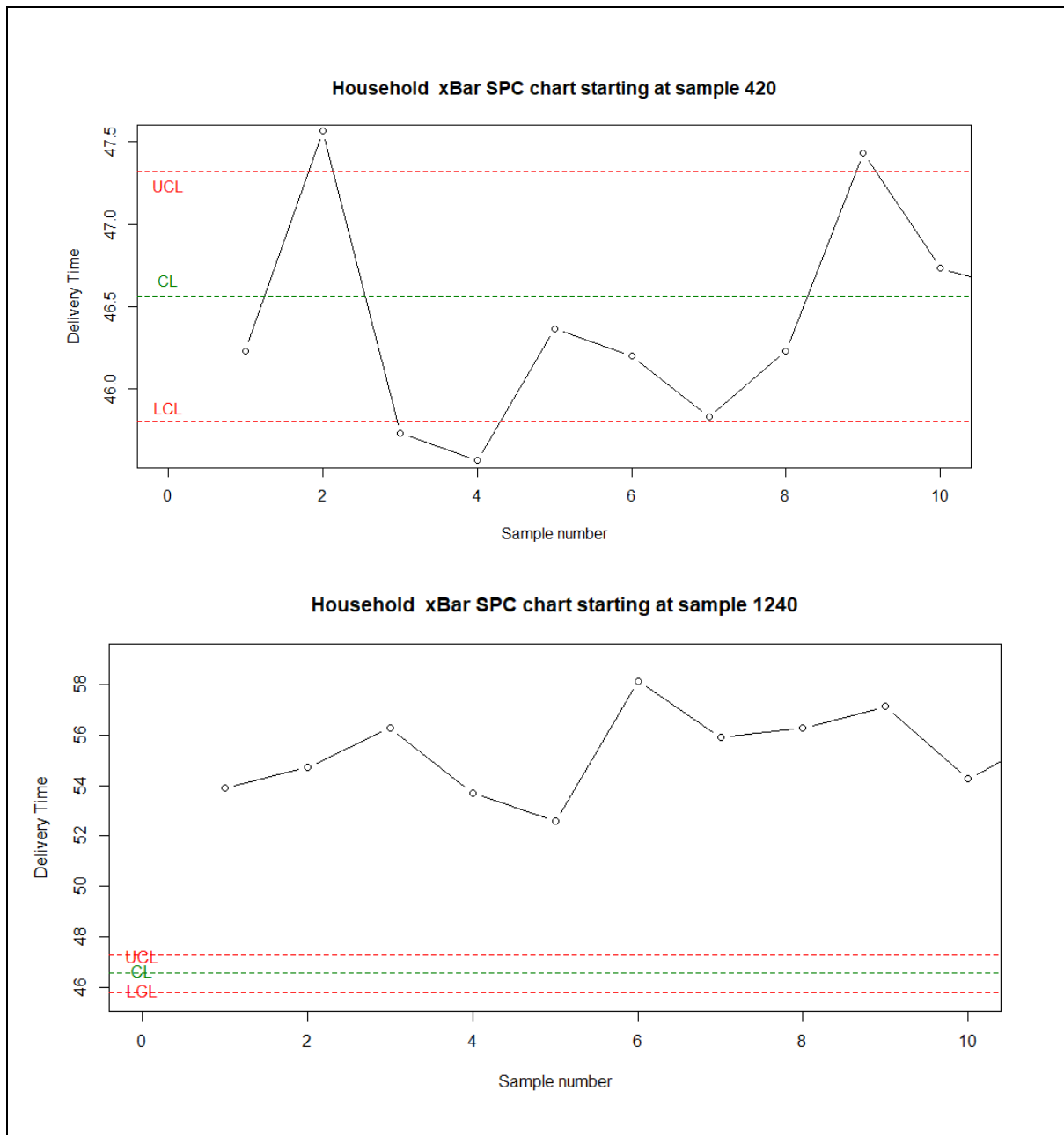


Figure 31: The first and last 3 outliers for Household

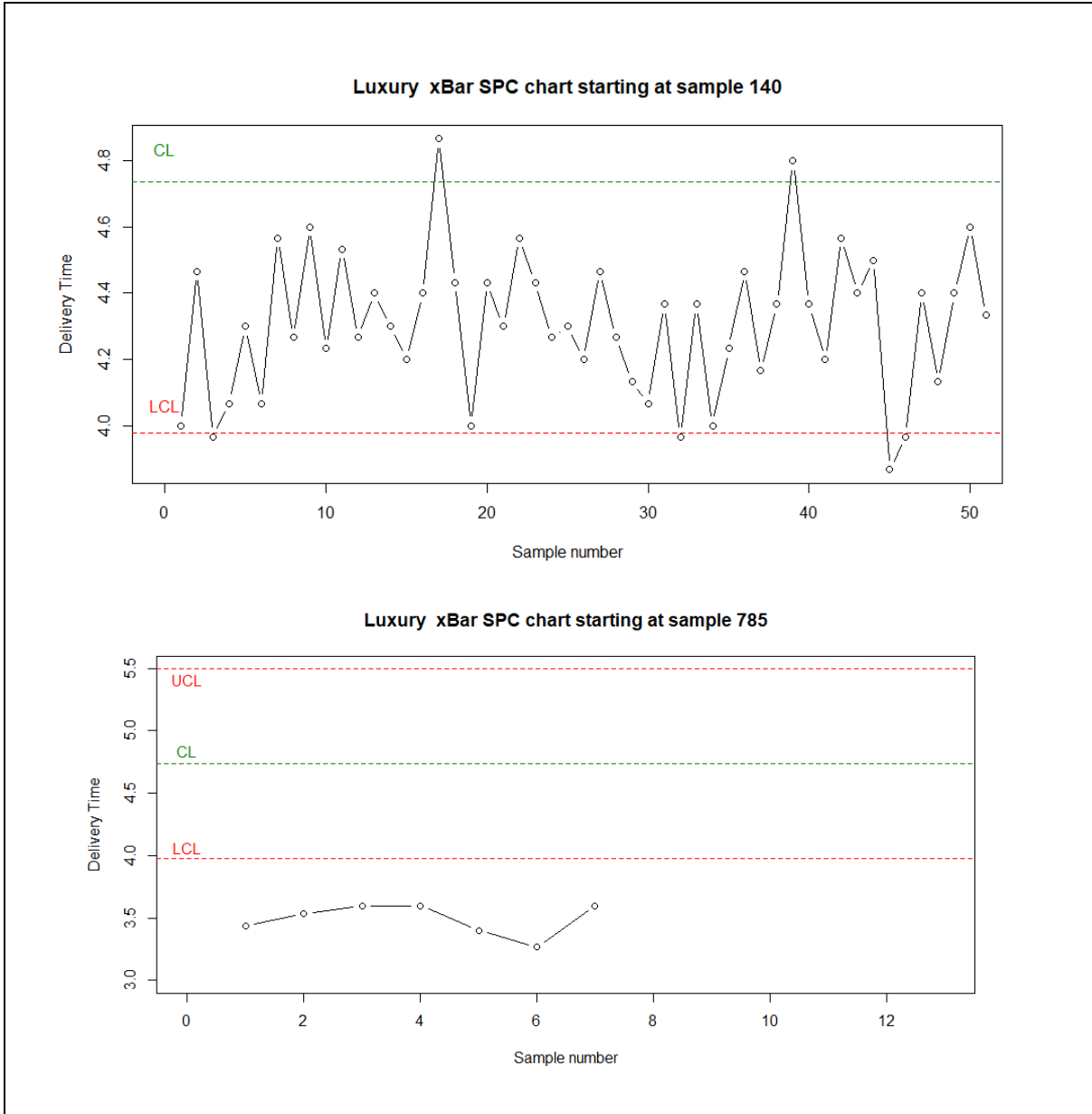


Figure 32: The first and last 3 outliers for Luxury

5.3 Most consecutive samples within -0.3 and +0.4 Sigma

	Clothing	Household	Food	Technology	Sweets	Gift	Luxury
Consecutive samples within limits	4	4	5	6	4	7	4
Final sample number	223	46	756	1598	94	2477	63

Table 8: Maximum consecutive samples

The Gifts and Technology classes has the most consecutive samples between the -0.3 and +0.4 sigma control limits. These classes are thus well controlled in the areas that these consecutive samples occur as they are operating well withing the larger UCL and LCL.

However, 7 consecutive samples are not a lot and could be a random occurrence. As seen in the previous chapter Gifts is not a well-controlled class.

To give an idea of how tight the margins are with -0.3 and +0.4 Sigma control limits the entire sweets class S-chart has been overlaid with purple lines indicating the -0.3 and +0.4 Sigma limits. Note how it is much more controlled than the UCL and LCL.

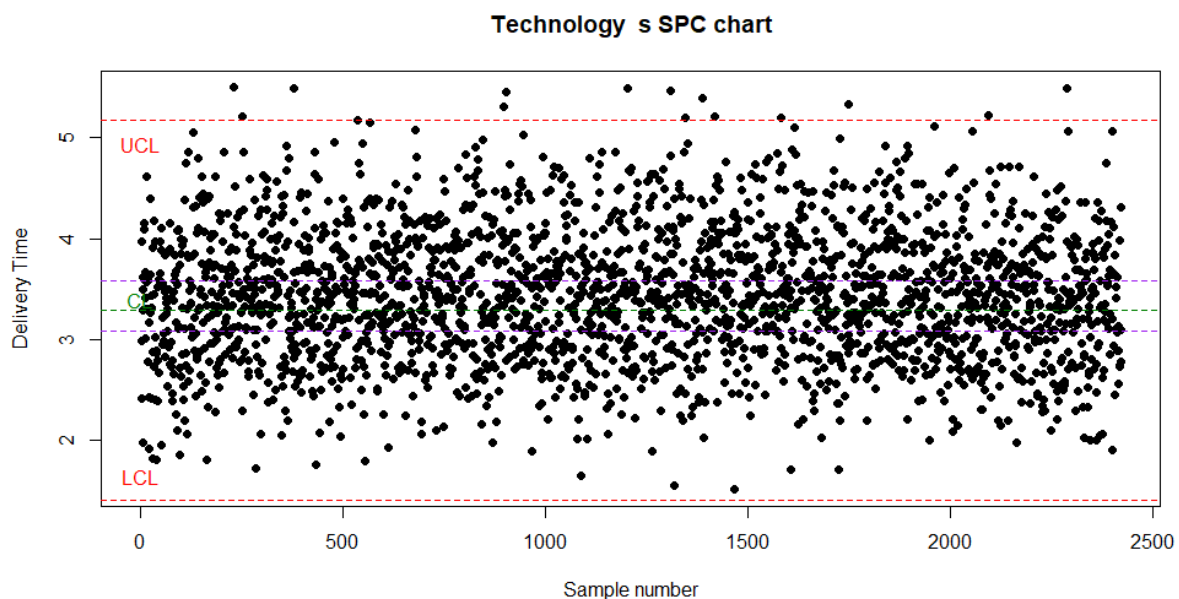


Figure 33: Technology S-chart overlaid with -0.3 and +0.4 Sigma Control lines

By zooming in on this chart and connecting consecutive samples the following chart is made:

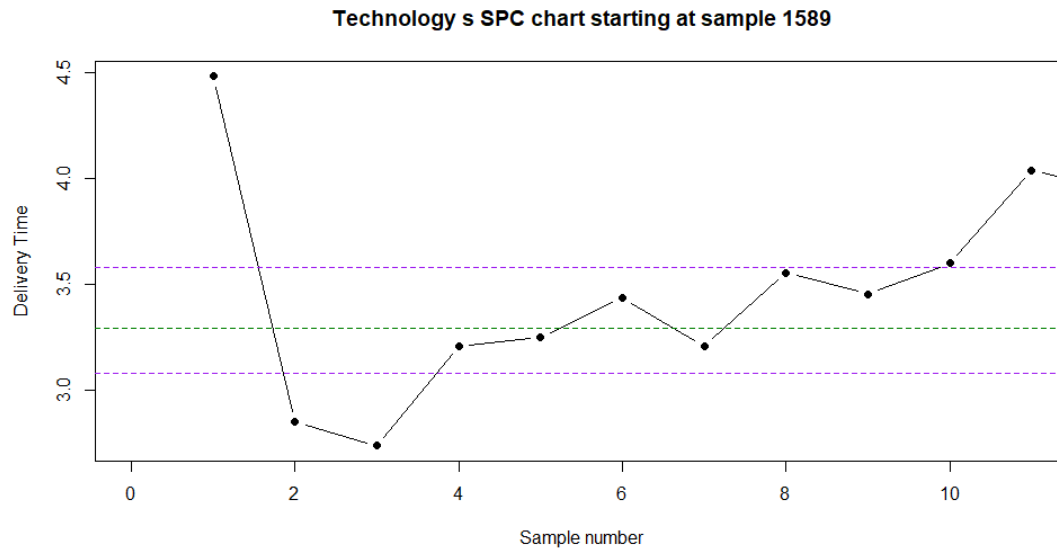


Figure 34: Most consecutive technology S-samples within limits

As indicated in the plot title the sample numbers on the X-axis start at sample 1590. This chart thus proves the table above to be legitimate and shows the 6 consecutive Technology samples between the limits, ending at sample 1598.

5.4 Estimate the likelihood of making type 1 error for A & B

A null hypothesis is made to calculate the type 1 error.

H_0 : the process is in control and centred on the centreline calculated using the first 30 samples.

H_1 : The process is not in control, nor centred on the centreline.

5.4.1 For A

$$P(\text{Type I error for A}) = \text{pnorm}(-3) \times 2 = 0.002699796$$

This means there is a 0.27% chance of mistakenly assuming products that were delivered on time were not.

5.4.2 For B

$$P(\text{Type I error for B}) = [1 - \text{pnorm}(0.4)] + \text{pnorm}(-0.3) = 0.5763928 = 57.639\%$$

This means there is a 57.64% chance of mistakenly assuming products that were delivered on time were not.

5.5 Minimizing delivery cost

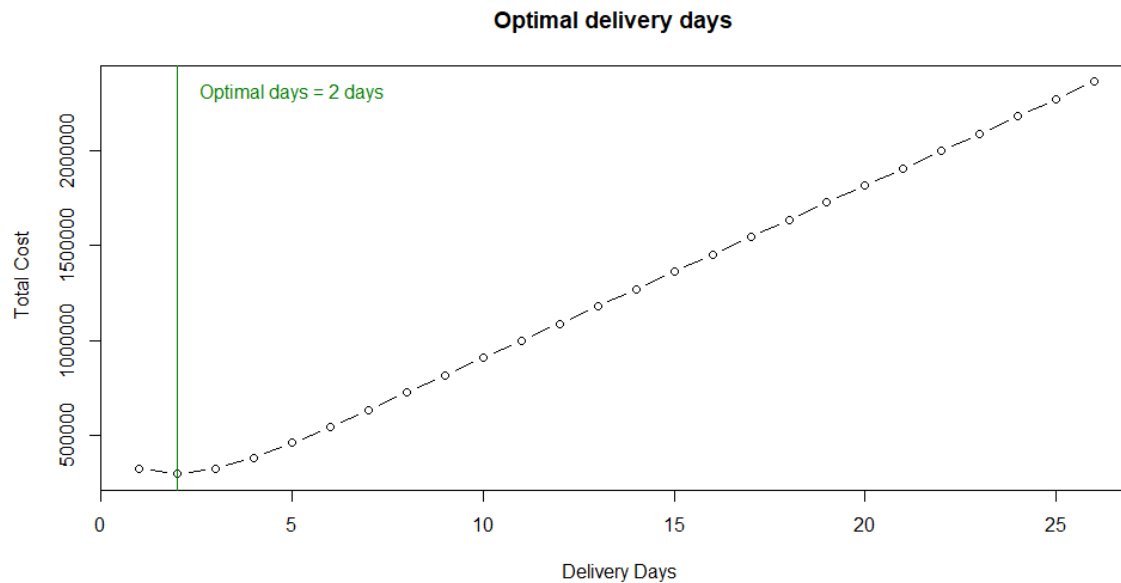


Figure 35: Graph of total cost for various delivery times

Currently, technology product deliveries take an average of around 20 days. 1356 sales have had delivery times over 26 days. If each failed sale costs R329, there is currently a loss of R446 124.

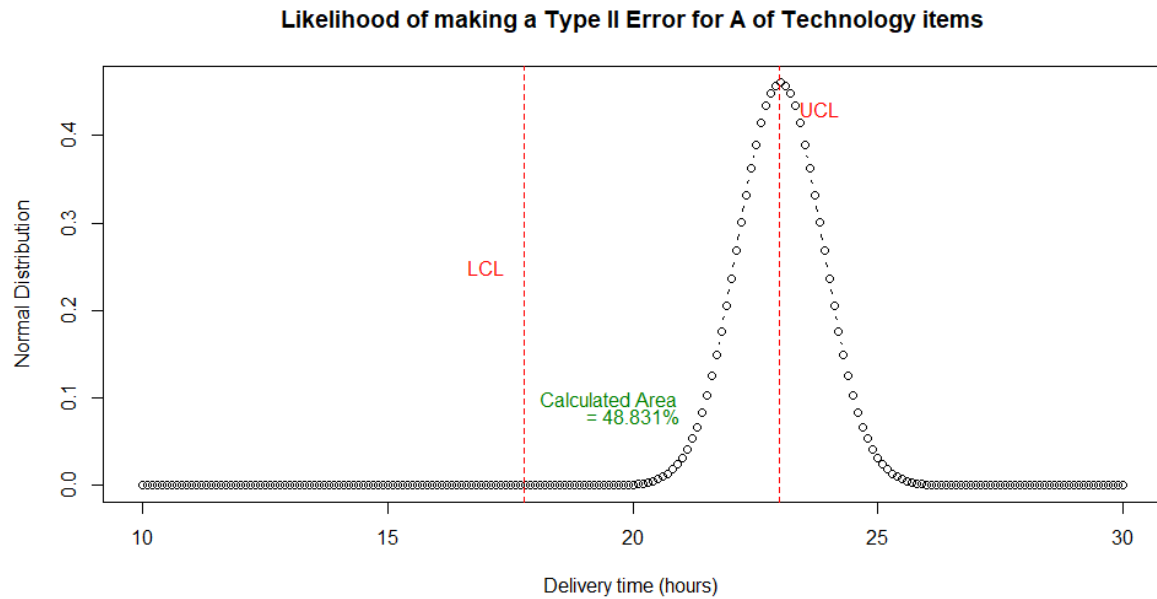
Any deviation from the target delivery time will result in an increasing loss because the consumer is happiest when the product is exactly on time. Shifting the distribution's mean to the left over over a 26-day period costs R1.5 per day. To move the whole distribution within 26 days will thus cost the company R381643.50

Therefore, the correct balance needs to be found between making as many as possible deliveries on target, but considering the cost of moving the mean.

By evaluating each day separately and plotting them on a chart this point (the local minimum) can be identified. The local minimum indicates the delivery time that will lead to the lowest total cost for the company. The optimal number of days is 2 days.

5.6 Estimate the likelihood of making type II error for A

The probability of incorrectly failing to reject the null hypothesis even though it does not apply to the entire population is known as a type II error. When it comes to delivery timings, it sometimes happens that while the goods are expected to be delivered on time, it isn't.



The red lines on the graph represent the outside control limits (UCL and LCL) that were re-used from the control charts in section 4. The area of the graph between the control limits represents the probability of a type II error. There is thus a 0.6372 probability for a type II error. Given that this likelihood is comparatively high, the business must take steps to ensure that the product is delivered on time and not just assume all deliveries expected to be on time automatically are.

6 MANOVA testing

Throughout this chapter a p-value of 0.05 is used as a threshold. This is a viable choice and indicates a 95% confidence level (Taylor, 2017).

6.1 Hypothesis 1 (effect of class)

6.1.1 Hypothesis statements

H₀: The class of product has no relationship to the Price, Delivery Time and Age of clients.

H₁: The class of product has a relationship to the Price, Delivery Time or Age of clients.

6.1.2 P-Values

```

Response Delivery.time :
      Df Sum Sq Mean Sq F value Pr(>F)
Class    6 33458565 5576427  629429 < 2.2e-16 ***
Residuals 179971 1594452      9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Price :
      Df Sum Sq Mean Sq F value Pr(>F)
Class    6 5.7168e+13 9.5281e+12  80258 < 2.2e-16 ***
Residuals 179971 2.1366e+13 1.1872e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response AGE :
      Df Sum Sq Mean Sq F value Pr(>F)
Class    6 8422401 1403733  3805 < 2.2e-16 ***
Residuals 179971 66394669      369
---

```

Table 9: MANOVA table for hypothesis 1

6.1.3 Visualisation

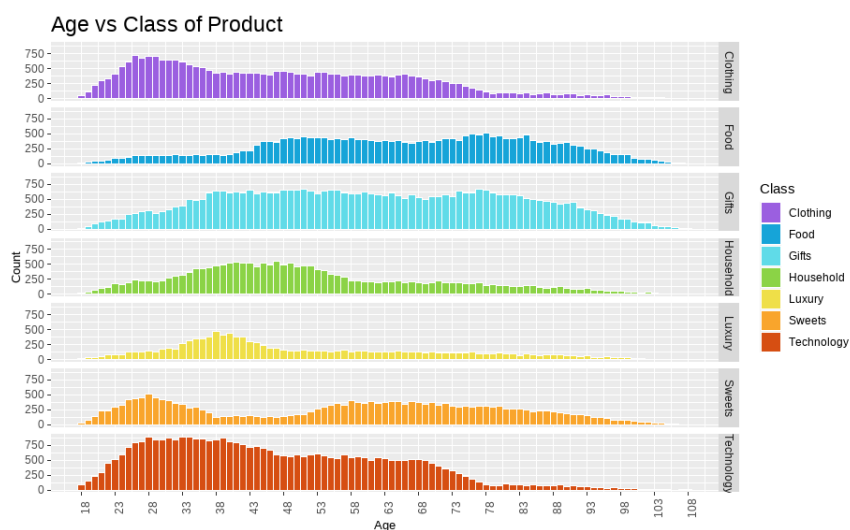


Figure 36: Age vs Class of product

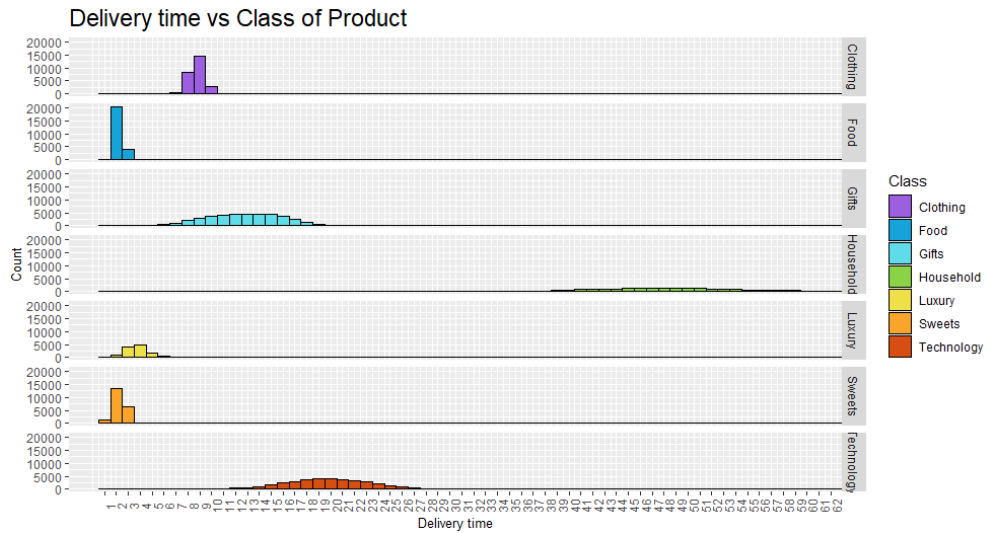


Figure 37: Delivery time vs Class of product

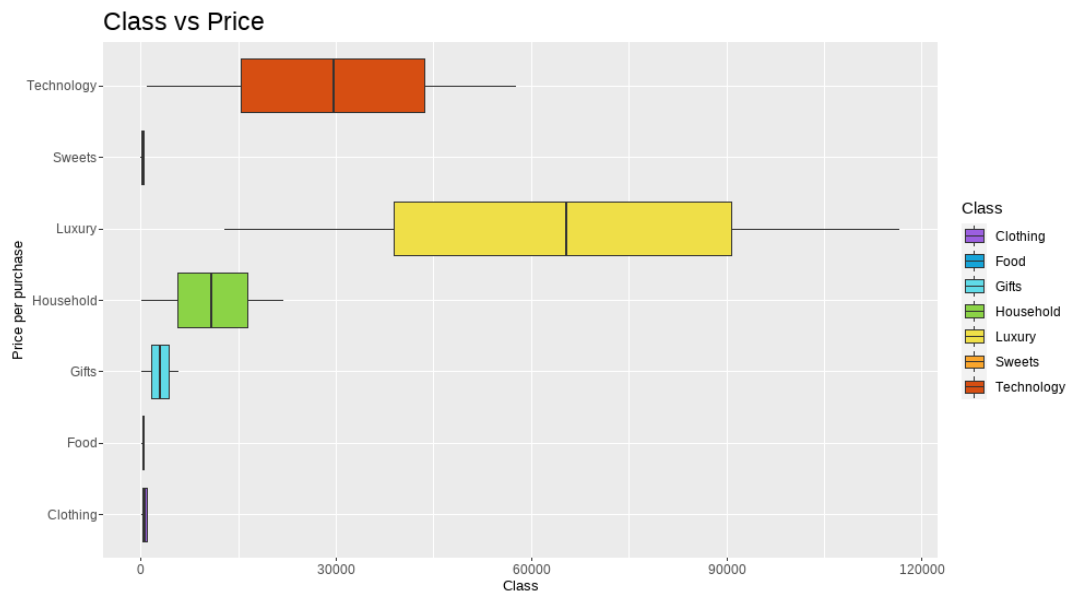


Figure 38: Class vs Price of product

6.1.4 Conclusion

As seen on the MANOVA table and supported by the graphs the P-values for the dependent variables Delivery Time, Price and Age are all **2.2e-16**. this value is lower than the threshold p-value of 0.5.

Thus, the null hypothesis is rejected. the class of the product bought influences the price, delivery time and the age of the clients buying the product.

6.2 Hypothesis 2 (effect of why bought)

6.2.1 Hypothesis statements

H₀: The reason for the purchase of product has no relationship to the Day, Month and Year in which the product is bought.

H₁: The reason for purchase of product has a relationship to at least one of the features Day, Month and/or Year.

6.2.2 P-values

Response Day :						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
why. Bought	5	299	59.829	0.7998	0.5495	
Residuals	179972	13462048	74.801			
Response Month :						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
why. Bought	5	29	5.7748	0.4841	0.7884	
Residuals	179972	2146930	11.9292			
Response Year :						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
why. Bought	5	3191	638.21	82.567	< 2.2e-16 ***	
Residuals	179972	1391111	7.73			

Table 10: MANOVA values for hypothesis test 2

6.2.3 Visualisation

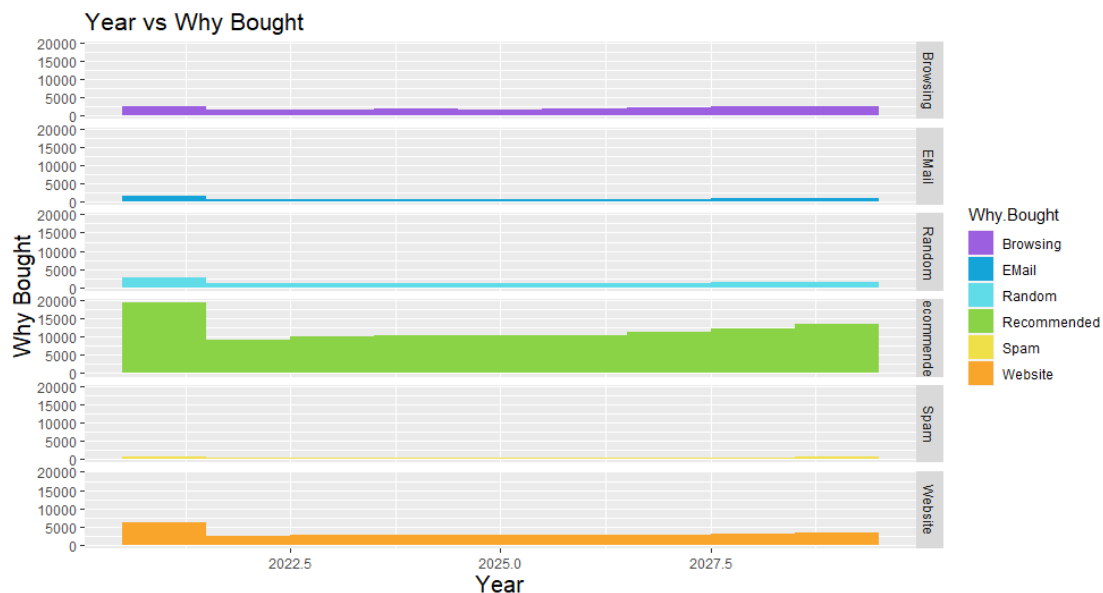


Figure 39: Year vs Why Bought

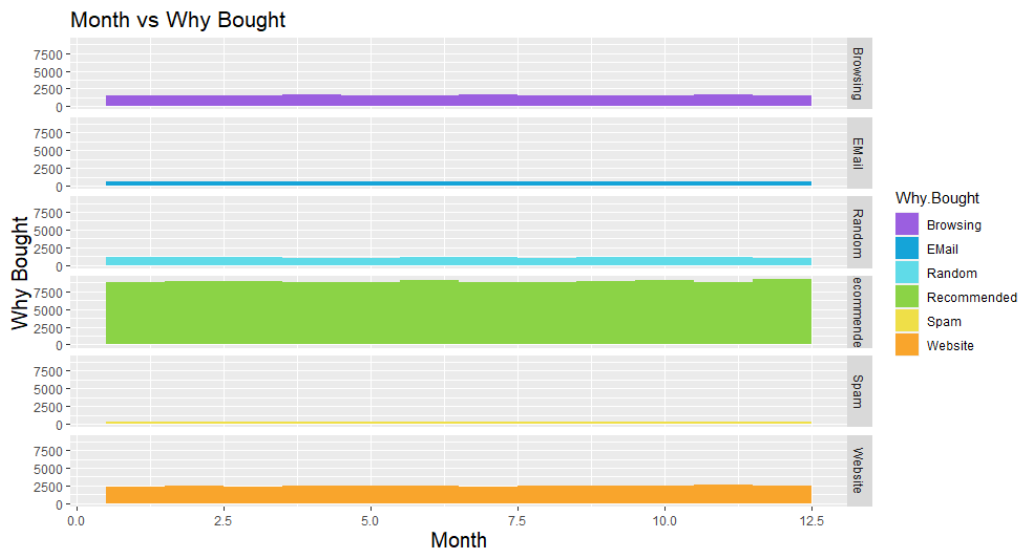


Figure 40: Month vs Why bought

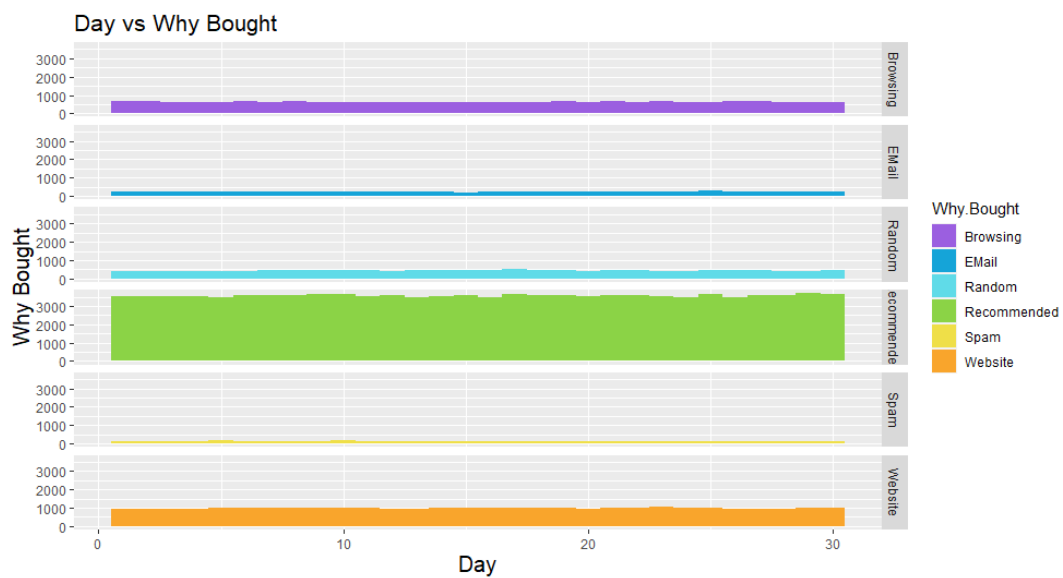


Figure 41: Day vs Why bought

6.2.4 Conclusion

As seen on the table above, the p-values for day and month feature are higher than the threshold p-value of 0.05. thus, the reason for purchases will not differ depending on the day or month of the purchases.

The p-value for the Year feature, however, is $22.e-16$, which is smaller than the threshold p-value of 0.05. thus, the null hypothesis for year is rejected. The reason for purchase will be affected by the year in which the purchases are made.

The rejection and acceptance of hypothesis based on p-values is supported by the visualisations.

7 Reliability of the service and products

7.1 Taguchi loss

Calculating the Taguchi loss function for a refrigerator part with thickness $0.06 \pm 0.04\text{cm}$.

The concept of Taguchi loss is that the cost of scrapping a product increases in a non-linear manner with increasing deviation from the goal value (0.06 in this instance). When a business must produce more products than needed (due to obsolesces) to satisfy a set of requirements, the cost will increase and there will be more waste if the quality is poor. Products that are unreliable and have features that differ from the requirements will reduce service effectiveness at the expense of the company.

All charts in this chapter are a visualisation of this concept and used to support the calculations.

Customers will be satisfied if the thickness of the refrigerator part is between the lower and upper limit. (0.02 and 0.10 in these examples)

7.1.1 At scrap cost of \$45

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

$$45 = k(0.06)^2$$

$$\therefore k = 28\,125$$

Using Taguchi loss function of $L(x) = 28\,125(y - 0.06)^2$, the following chart is obtained:

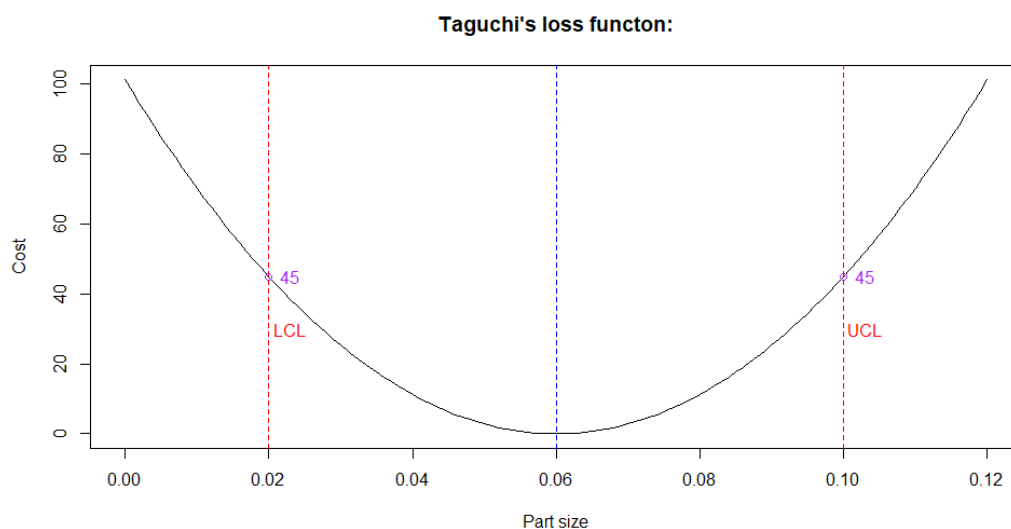


Figure 42: Taguchi loss function at \$45 scrap cost

7.1.2 At scrap cost of \$35

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

$$35 = k(0.06)^2$$

$$\therefore k = 21\,875$$

Using Taguchi loss function of $L(x) = 21\,875(y - 0.06)^2$, the following chart is obtained:

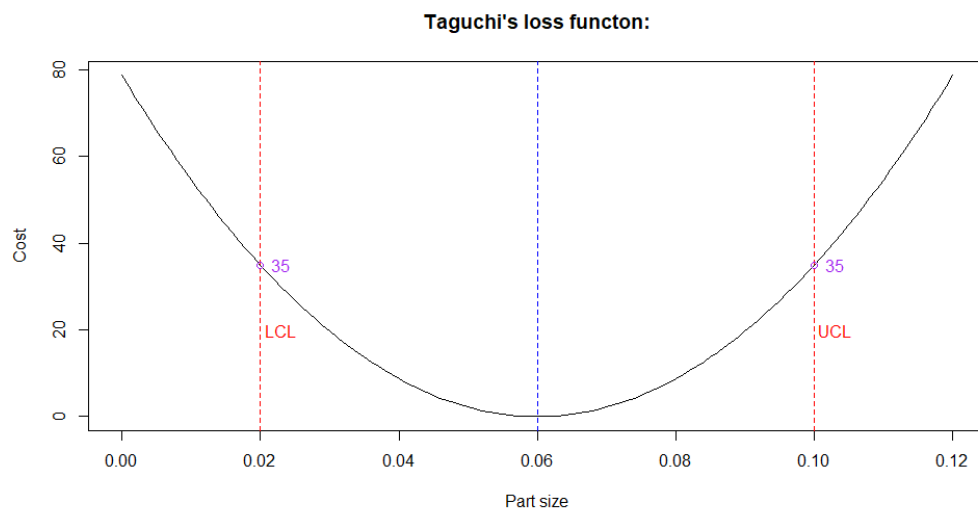


Figure 43: Taguchi loss function at \$35 scrap cost

7.1.3 Process deviation from target is reduced to 0.027

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

$$L(0.027) = \$15.9468 \approx \$15.95$$

On graph:

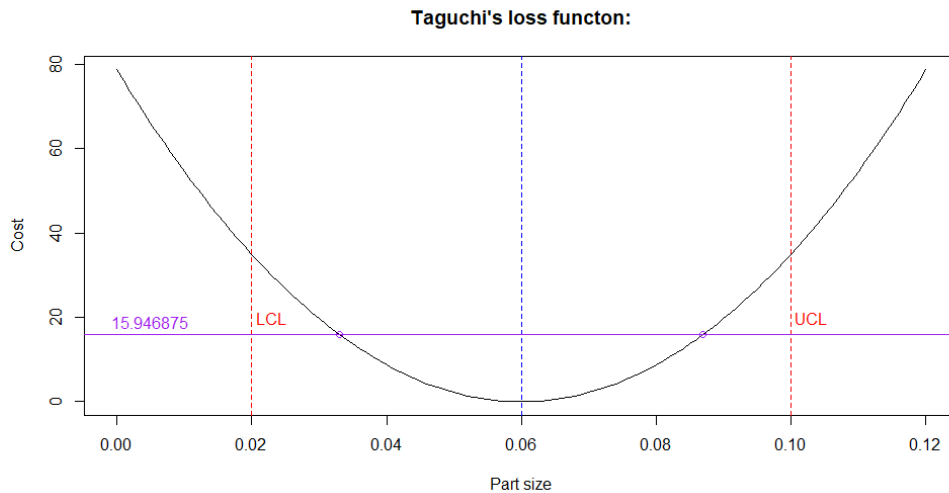


Figure 44: Taguchi loss at process deviation of 0.027 cm

Thus, with a process deviation of 0.027cm the company makes a loss of \$15.95 per item.

7.2 System reliability

Production systems need to be reliable to ensure that service levels to customers are upheld. When machines break down it can lead to stockouts and frustrated customers. The effect of having 2 machines per stage of the production system is evaluated here.

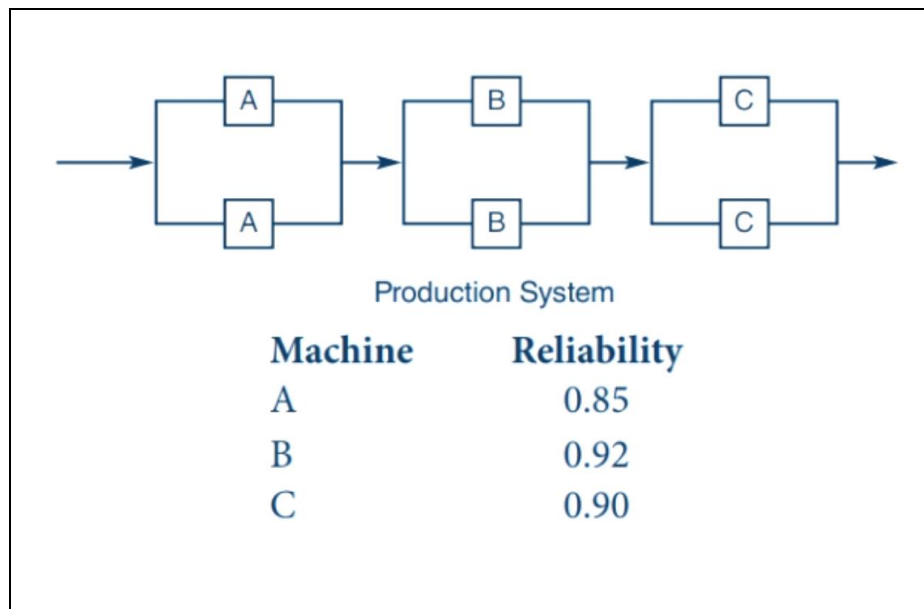


Figure 45: Reliability of machines

7.2.1 Reliability if only one machine at A, B and C is used

$$\text{Reliability} = R_a * R_b * R_c$$

$$\text{Reliability} = 0.85 \times 0.92 \times 0.90 = 0.7038$$

If only one machine per production category is used the reliability of the total system is 0.7038. This means the production line will run 70.38% of the time.

7.2.2 Reliability if two machines at A, B and C is used

When two machines of each compartment is working it will result in the process reliability being higher.

When components are connected in parallel, the combined reliability is $1 - P(\text{both fail})$, using this concept the total reliability of the system can be calculated as follows:

$$R_{AA} = 1 - (1 - 0.85)^2 = 0.9775$$

$$R_{BB} = 1 - (1 - 0.92)^2 = 0.9936$$

$$R_{CC} = 1 - (1 - 0.90)^2 = 0.99$$

Then the total reliability can be calculated by multiplying the new serie equivalent reliabilities together.

$$R_{AA} \times R_{BB} \times R_{CC} = P(\text{System})$$

$$0.9775 \times 0.9936 \times 0.99 = P(\text{System})$$

$$0.9615316 = P(\text{System})$$

Using 2 machines in parallel leads to a 26% improvement in production system reliability. This is due to the production system still being able to run if one of the machines break, as the other one will still be able to produce – allowing the rest of the production line to run.

7.3 Binomial distribution

7.3.1 Case 1: 20 vehicles available

$$P(\text{Vehicles}) = \text{Reliability of Vehicles} = 0.990$$

$$P(\text{Drivers}) = \text{Reliability of Drivers} = 0.998$$

$$\therefore \text{Total Reliability} = P(\text{Vehicles}) \times P(\text{Drivers}) \times 365 = 0.98834$$

$$\text{Total Reliability} = 360.7449$$

7.3.2 Case 2: 21 vehicles available

$$P(\text{Vehicles}) = \text{Reliability of Vehicles} = 0.999$$

$$P(\text{Drivers}) = \text{Reliability of Drivers} = 0.998$$

$$\therefore \text{Total Reliability} = P(\text{Vehicles}) \times P(\text{Drivers}) \times 365 = 0.99788$$

$$\text{Total Reliability} = 364.229$$

7.3.3 Conclusion:

The availability of an extra vehicle will result in 3.48 additional days available for delivery. This will improve the service level of the business, help control the Delivery Times and lead to higher levels of customer satisfaction. The price of an extra vehicle needs to be considered, however, as the difference is marginal. Especially as most classes have expected delivery times that has a larger scope between the limits than 3.48.

8 Conclusion

The sales value data was cleaned of missing and negative values to ensure all graphs and calculations are correct and reliable. The data was split into 2 separate sets and saved as Excel files for future use. The valid data set was used to create visual representations that aided in the review process and helped to identify trends between features. These identified trends were used to advise the business on future processes, such as: how to optimize future delivery times and which clients to target. X- and S charts were created for statistical process control and to identify which classes are out of control. It was evident that that gifts, luxury items and household products are not controlled. Recommendations were made for these out-of-control features to improve the control level. A new logistics partner was recommended to solve some of these control issues. The MANOVA test supported this. A type I error had a far lower likelihood of occurring than a type II error. Therefore, the business should emphasize ensuring that things are delivered on time rather than presuming that they are. Samples outside the control limits were isolated and analysed on their own to provide more insight into potential business problems. Finally, calculations were done to show the effect of the Taguchi Loss function on the total production cost, that parallel machines in the production line will be beneficial to the business, and how various amounts of delivery drivers influence the business.

9 References

Comtois, D., 2022. *Cran R*. [Online]

Available at: [https://cran.r-](https://cran.r-project.org/web/packages/summarytools/vignettes/introduction.html)

[project.org/web/packages/summarytools/vignettes/introduction.html](https://cran.r-project.org/web/packages/summarytools/vignettes/introduction.html)

Data Carpentry, 2020. *Data Carpentry*. [Online]

Available at: <https://datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2>

[Accessed 24 10 2022].

M, B., 2020. *geeksforgeeks*. [Online]

Available at: [https://www.geeksforgeeks.org/binomial-distribution-in-r-](https://www.geeksforgeeks.org/binomial-distribution-in-r-programming/#:~:text=Binomial%20distribution%20in%20R%20is,not%20affect%20the%20next%20outcome.)

[programming/#:~:text=Binomial%20distribution%20in%20R%20is,not%20affect%20the%20next%20outcome.](https://www.geeksforgeeks.org/binomial-distribution-in-r-programming/#:~:text=Binomial%20distribution%20in%20R%20is,not%20affect%20the%20next%20outcome.)

[Accessed 22 10 2022].

Philips, B., 2013. *BMJ Blogs*. [Online]

Available at: [https://blogs.bmj.com/adc/2013/05/13/statsminiblog-continuous-vs-](https://blogs.bmj.com/adc/2013/05/13/statsminiblog-continuous-vs-categorical/#:~:text=Categorical%20variables%2C%20aka%20discrete%20variables,BMI%2C%20temperature%2C%20neutrophil%20count.)

[categorical/#:~:text=Categorical%20variables%2C%20aka%20discrete%20variables,BMI%2C%20temperature%2C%20neutrophil%20count.](https://blogs.bmj.com/adc/2013/05/13/statsminiblog-continuous-vs-categorical/#:~:text=Categorical%20variables%2C%20aka%20discrete%20variables,BMI%2C%20temperature%2C%20neutrophil%20count.)

STHDA, 2020. *Statistical tools for high-throughput data analysis*. [Online]

Available at: <http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance>

[Accessed 23 10 2022].

Taylor, C., 2017. *ThoughtCo*. [Online]

Available at: [https://www.thoughtco.com/what-is-a-p-value-](https://www.thoughtco.com/what-is-a-p-value-3126392#:~:text=The%20answer%20to%20this%20is,we%20choose%20a%20threshold%20value.)

[3126392#:~:text=The%20answer%20to%20this%20is,we%20choose%20a%20threshold%20value.](https://www.thoughtco.com/what-is-a-p-value-3126392#:~:text=The%20answer%20to%20this%20is,we%20choose%20a%20threshold%20value.)

[Accessed 18 10 2022].

Yi, M., 2021. *Chartio*. [Online]

Available at: <https://chartio.com/learn/charts/histogram-complete-guide/>

[Accessed 23 October 2022].