James Goldschmidt

Quality Assurance 344

Project Report

Final

21 October 2022

Table of Contents

Introduction 7

1 8

1.1 Invalid data 8

1.2 Valid Data 8

1.3 New Invalid data frame 9

1.4 New valid data frame 9

2 10

3 26

3.1 26

3.1 X Chart 26

3.2 S Chart 26

Graphs for each class for the oldest 30 samples 26

Technology 26

Clothing 27

Household 27

Luxury 28

Food 28

Gifts 29

Sweets 29

3.2 30

Technology 30

Clothing 30

Household 31

Luxury 31

Food 32

Gifts 32

Sweets 33

4 33

4.1.A 33

Household 34

Luxury 34

Gifts 34

4.1.B 36

Technology 36

Clothing 37

Household 37

Luxury 38

Food 38
Gifts 38
Sweets 39
4.2 39
4.3 40
4.4 41
5 41
5.1 41
5.2 42
6 44
6.1 44
6.2 46
6.3 46
Conclusion 49
References 50

Table of Figures

Figure 1,Invalid data 8
Figure 2, Valid data 8
Figure 3, Assigned invalid data 9
Figure 4, Assigned valid data 9
Figure 5, Total sales for each year 10
Figure 6, Sales per month 11
Figure 7, Sales per day 12
Figure 8, Delivery time distribution for sales 13
Figure 9, Age of customers for sales 14
Figure 10, Reason for purchase 15
Figure 11, Class of product sold 16
Figure 12, Age vs why bought 17
Figure 13, Age of customer vs class of product 18
Figure 14, Month of sale vs class of product 19
Figure 15, Year vs class of product 20
Figure 16, Price vs class of product 21
Figure 17, Delivery time vs class of product 22
Figure 18, Price vs why bought 23
Figure 19, Delivery time vs why bought 24
Figure 20, Delivery time for technology 25
Figure 21, technology mean & standard deviation graph 30 samples 26
Figure 22, Clothing mean & standard deviation graph 30 samples 27
Figure 23, Household mean & standard deviation graph 30 samples 27
Figure 24, Luxury mean & standard deviation graph 30 samples 28
Figure 25, Food mean & standard deviation graph 30 samples 28
Figure 26, Gifts mean & standard deviation graph 30 samples 29
Figure 27, Sweets mean & standard deviation graph 30 samples 29
Figure 28, Technology mean & standard deviation graph all samples 30
Figure 29, Clothing mean & standard deviation graph all samples 30
Figure 30, Household mean & standard deviation graph all samples 31
Figure 31, Luxury mean & standard deviation graph all samples 31
Figure 32, Food mean & standard deviation graph all samples 32
Figure 33, Gifts mean & standard deviation graph all samples 32
Figure 34, Sweets mean & standard deviation graph all samples 33
Figure 35, Household first 3 & last 3 outliers 34
Figure 36, Luxury first 3 & last 3 outliers 34
Figure 37, Gifts first 3 & last 3 outliers 34
Figure 38, Technology consecutive sample points 36
Figure 39, Clothing consecutive sample points 37
Figure 40, Household consecutive sample points 37
Figure 41, Luxury consecutive sample points 38
Figure 42, Food consecutive sample points 38
Figure 43, Gifts consecutive sample points 39
Figure 44, Sweets consecutive sample points 39
Figure 45, Optimal delivery hours 40
Figure 46, Type II error 41
Figure 47, MANOVA 1 supporting graphs 42
Figure 48, MANOVA 2 supporting graphs 43

Figure 49, Taguchi loss function 1 44
Figure 50, Taguchi loss function 2 45
Figure 51, Machine layout 46
Figure 52, Binomial formula 47

List of Tables

Table 1, Delivery process time x-chart 26
Table 2, Delivery process time s-chart 26
Table 3, Consecutive points and index 36
Table 4, Machine data 46

Introduction

The report covers the use of statistical analysis of visuals and data from an online retailer's sales. Six parts make up the report, starting off with data manipulation and wrangling, secondly descriptive statistics, thirdly statistical process control (SPC), fourthly sample processing of data and optimization of the delivery process, fifthly using MANOVA testing to find connections and lastly the reliability of deliveries of the product and service.

Cleaning of the data results in valid data to be analysed and used throughout the report. Purchase dates, classes, clients' information and product prices are used to find trends and relationships through analysis. Delivery time is the focus for the x- and s-charts for each class to perform descriptive statistics. Calculations are done to find the optimised delivery time to result in the lowest cost paid by the company to delivery their products.

The conclusion finishes the report with a reflection on the six parts of the report.

1

1.1 Invalid data

	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
12345	12345	18973	93	Gifts	NA	2026	6	11	15.5	Website
16321	16321	81959	43	Technology	NA	2029	9	6	22.0	Recommended
19541	19541	71169	42	Technology	NA	2025	1	19	20.5	Recommended
19999	19999	67228	89	Gifts	NA	2026	2	4	15.0	Recommended
23456	23456	88622	71	Food	NA	2027	4	18	2.5	Random
34567	34567	18748	48	Clothing	NA	2021	4	9	8.0	Recommended
45678	45678	89095	65	Sweets	NA	2029	11	6	2.0	Recommended
54321	54321	62209	34	Clothing	NA	2021	3	24	9.5	Recommended
56789	56789	63849	51	Gifts	NA	2024	5	3	10.5	Website
65432	65432	51904	31	Gifts	NA	2027	7	24	14.5	Recommended
76543	76543	79732	71	Food	NA	2028	9	24	2.5	Recommended
87654	87654	40983	33	Food	NA	2024	8	27	2.0	Recommended
98765	98765	64288	25	Clothing	NA	2021	1	24	8.5	Browsing
144444	144444	70761	70	Food	NA	2027	9	28	2.5	Recommended
155555	155555	33583	56	Gifts	NA	2022	12	9	10.0	Recommended
166666	166666	60188	37	Technology	NA	2024	10	9	21.5	Website
177777	177777	68698	30	Food	NA	2023	8	14	2.5	Recommended

Figure 1,Invalid data

All the entries in the dataset that has NA in the entries indicating that the entries are invalid. These need to be removed to have valid entries left behind.

1.2 Valid Data

	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	Technology	41843.21	2029	10	22	27.0	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	108953.53	2029	11	13	4.0	Recommended
8	8	32624	58	Sweets	389.62	2025	7	2	2.0	Recommended
9	9	51401	82	Gifts	3312.11	2025	12	18	12.0	Recommended
10	10	96430	24	Sweets	176.52	2027	11	4	3.0	Recommended
11	11	87530	33	Technology	8515.63	2026	7	15	21.0	Browsing
12	12	14607	64	Gifts	3538.66	2026	5	13	13.5	Recommended
13	13	24299	52	Technology	27641.97	2024	5	29	17.0	Browsing
14	14	77795	92	Food	556.83	2025	6	3	3.0	Random
15	15	62567	73	Clothing	347.99	2024	3	29	8.5	Website
16	16	14839	47	Technology	54650.41	2027	12	30	18.5	Recommended
17	17	96208	44	Technology	14739.09	2028	3	17	13.0	Recommended
18	18	39674	69	Technology	22315.17	2026	8	20	20.5	Recommended
19	19	98694	74	Sweets	546.48	2025	5	9	2.0	Recommended
20	20	99187	54	Luxury	81620.21	2027	9	14	3.0	Recommended
21	21	59365	72	Gifts	3314.76	2028	4	30	13.0	Recommended
22	22	37221	24	Sweets	220.91	2021	3	8	3.0	Recommended
23	23	78120	23	Gifts	2378.31	2023	3	10	12.0	Recommended

Figure 2, Valid data

All the remainder entries without any NA entries indicating that all entries in this table are valid.

1.3 New Invalid data frame

	c(1:length_invalid_data)	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
12345	1	12345	18973	93	Gifts	NA	2026	6	11	15.5	Website
16321	2	16321	81959	43	Technology	NA	2029	9	6	22.0	Recommended
19541	3	19541	71169	42	Technology	NA	2025	1	19	20.5	Recommended
19999	4	19999	67228	89	Gifts	NA	2026	2	4	15.0	Recommended
23456	5	23456	88622	71	Food	NA	2027	4	18	2.5	Random
34567	6	34567	18748	48	Clothing	NA	2021	4	9	8.0	Recommended
45678	7	45678	89095	65	Sweets	NA	2029	11	6	2.0	Recommended
54321	8	54321	62209	34	Clothing	NA	2021	3	24	9.5	Recommended
56789	9	56789	63849	51	Gifts	NA	2024	5	3	10.5	Website
65432	10	65432	51904	31	Gifts	NA	2027	7	24	14.5	Recommended
76543	11	76543	79732	71	Food	NA	2028	9	24	2.5	Recommended
87654	12	87654	40983	33	Food	NA	2024	8	27	2.0	Recommended
98765	13	98765	64288	25	Clothing	NA	2021	1	24	8.5	Browsing
144444	14	144444	70761	70	Food	NA	2027	9	28	2.5	Recommended
155555	15	155555	33583	56	Gifts	NA	2022	12	9	10.0	Recommended
166666	16	166666	60188	37	Technology	NA	2024	10	9	21.5	Website
177777	17	177777	68698	30	Food	NA	2023	8	14	2.5	Recommended

Figure 3, Assigned invalid data

The invalid data entries are now combined with a new key/identification that is evenly and consecutively spaced without removing the original identification.

1.4 New valid data frame

	c(1:length_valid_data)	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	1	1	19966	54	Sweets	246.21	2021	7	3	1.5	Recommended
2	2	2	34006	36	Household	1708.21	2026	4	1	58.5	Website
3	3	3	62566	41	Gifts	4050.53	2027	8	10	15.5	Recommended
4	4	4	70731	48	Technology	41843.21	2029	10	22	27.0	Recommended
5	5	5	92178	76	Household	19215.01	2027	11	26	61.5	Recommended
6	6	6	50586	78	Gifts	4929.82	2027	4	24	14.5	Random
7	7	7	73419	35	Luxury	108953.53	2029	11	13	4.0	Recommended
8	8	8	32624	58	Sweets	389.62	2025	7	2	2.0	Recommended
9	9	9	51401	82	Gifts	3312.11	2025	12	18	12.0	Recommended
10	10	10	96430	24	Sweets	176.52	2027	11	4	3.0	Recommended
11	11	11	87530	33	Technology	8515.63	2026	7	15	21.0	Browsing
12	12	12	14607	64	Gifts	3538.66	2026	5	13	13.5	Recommended
13	13	13	24299	52	Technology	27641.97	2024	5	29	17.0	Browsing
14	14	14	77795	92	Food	556.83	2025	6	3	3.0	Random
15	15	15	62567	73	Clothing	347.99	2024	3	29	8.5	Website
16	16	16	14839	47	Technology	54650.41	2027	12	30	18.5	Recommended
17	17	17	96208	44	Technology	14739.09	2028	3	17	13.0	Recommended
18	18	18	39674	69	Technology	22315.17	2026	8	20	20.5	Recommended
19	19	19	98694	74	Sweets	546.48	2025	5	9	2.0	Recommended
20	20	20	99187	54	Luxury	81620.21	2027	9	14	3.0	Recommended
21	21	21	59365	72	Gifts	3314.76	2028	4	30	13.0	Recommended
22	22	22	37221	24	Sweets	220.91	2021	3	8	3.0	Recommended

Figure 4, Assigned valid data

The valid data entries are now combined with a new key/identification that is evenly and consecutively spaced without removing the original identification.

2

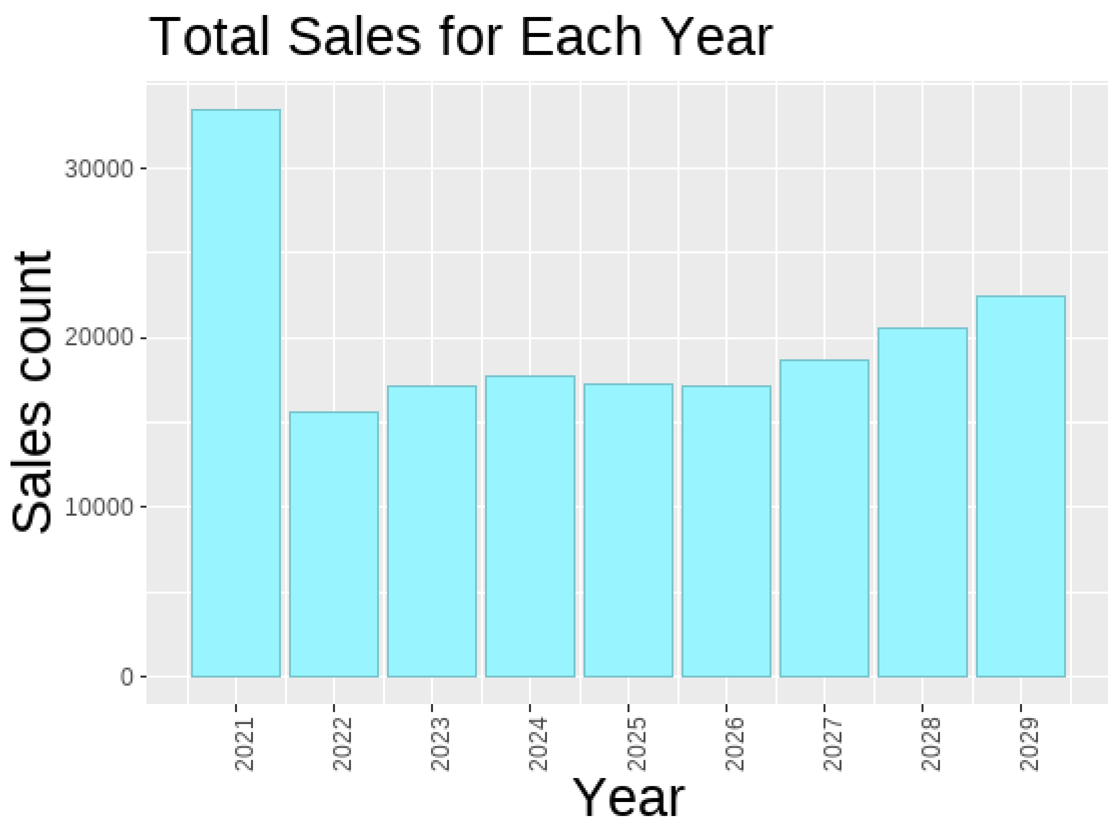


Figure 5, Total sales for each year

Figure 5 indicates that year 2021 was the year with the highest amount of sales. Between 2021 and 2022 there is a large drop of sales where 2022 has the lowest amount of sales compared to all the years. From 2022 to 2029, there is an upward trend of the amount of sales. An investigation should be down to identify why 2021 was such a good year for sales and what caused the huge drop so that history is less likely to repeat itself.

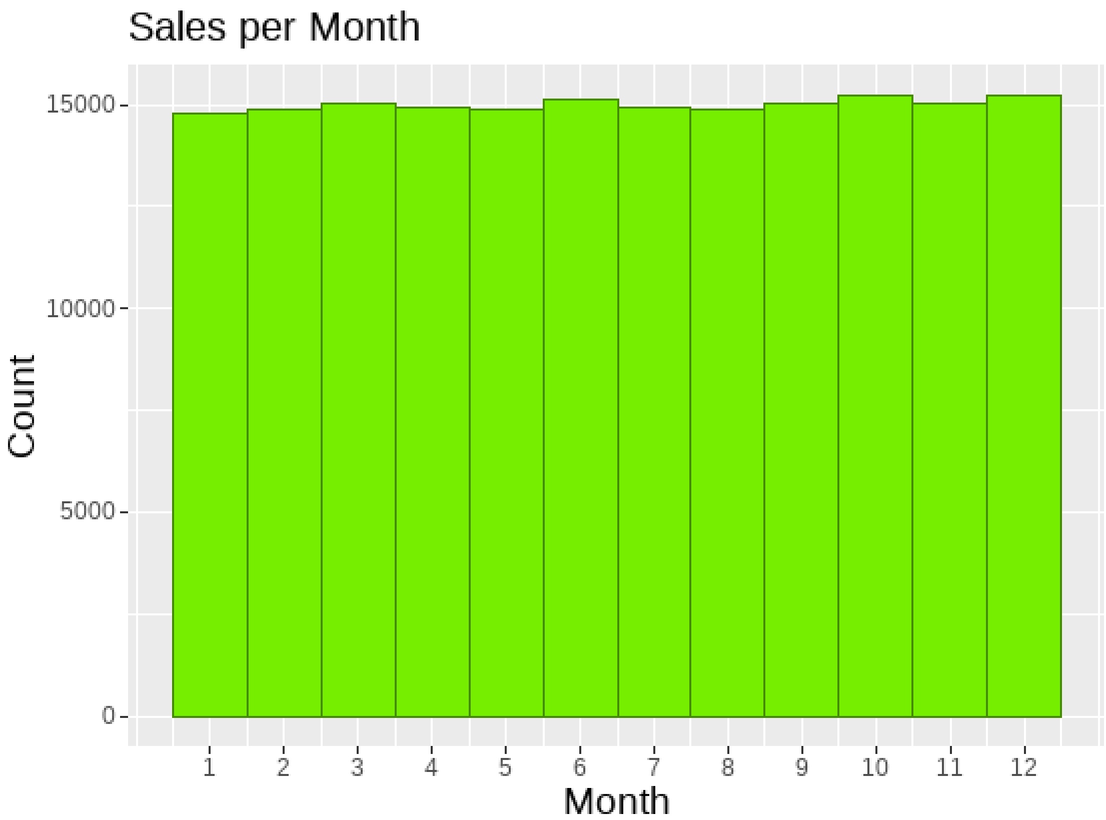


Figure 6, Sales per month

Figure 2 indicates that over the years that the sales are even throughout the 12 months. This puts less strain on the sales system as there seems like there are not large and small demand periods resulting in forecasting being easier and more predictable. Further investigation into the sales per month for each year to confirm this observation.

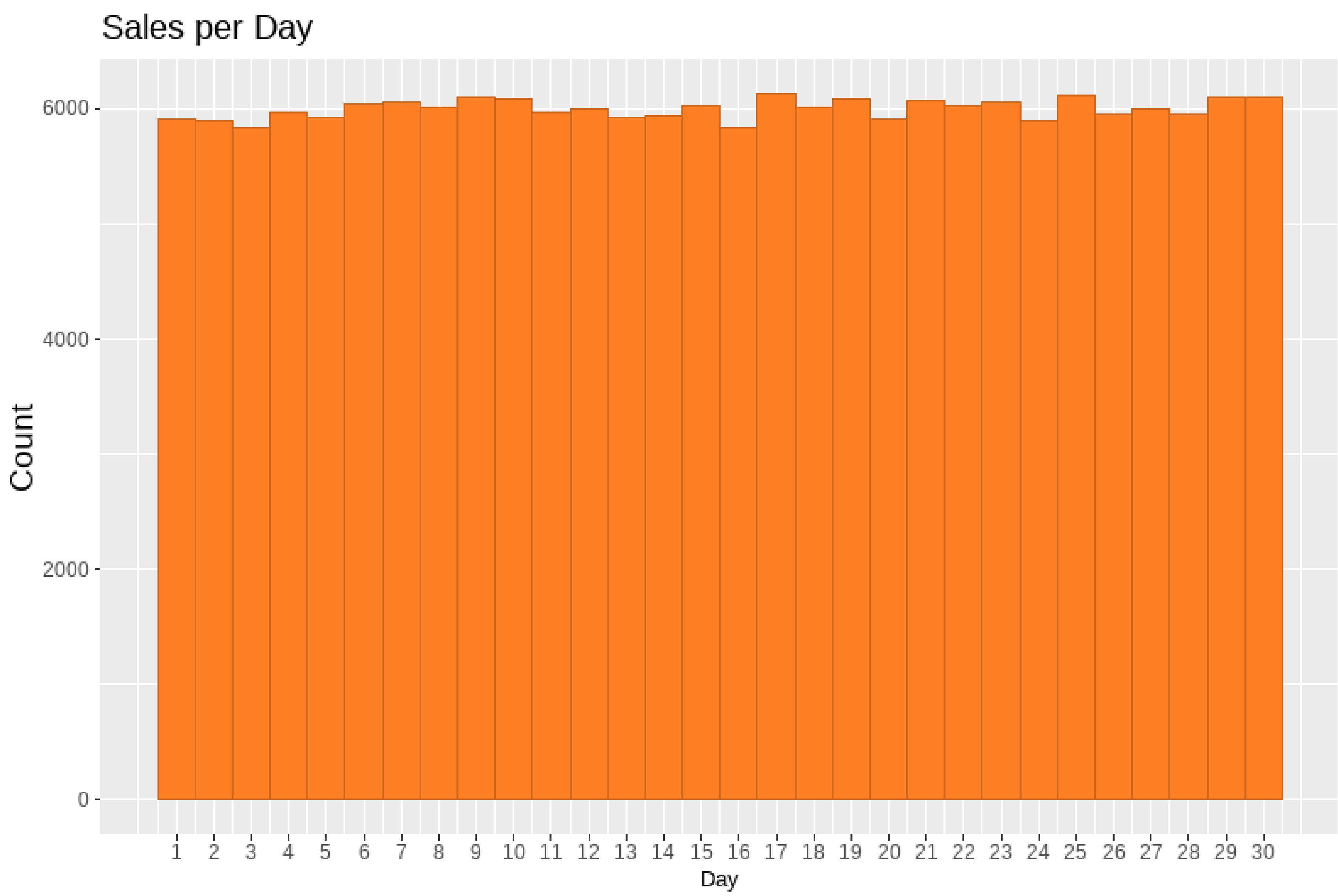


Figure 7, Sales per day

Figure 2 indicates that over the years that the sales are even throughout the 12 months. This puts less strain on the sales system as there seems like there are not large and small demand periods resulting in forecasting being easier and more predictable. Further investigation into the sales per month for each year to confirm this observation.

Figure 8, Delivery time distribution for sales

Figure 2.3 indicates that there are roughly two normal distribution graphs for the delivery time for sales. Further investigation needs to be taken into identifying what measures can be done to lower the delivery time for the domain between 35 and 65 days to bring them closer to the first normal distribution between 1 and 30 days. The largest amount of delivery times occure between 2 and 5 days and 7 and 11 days.

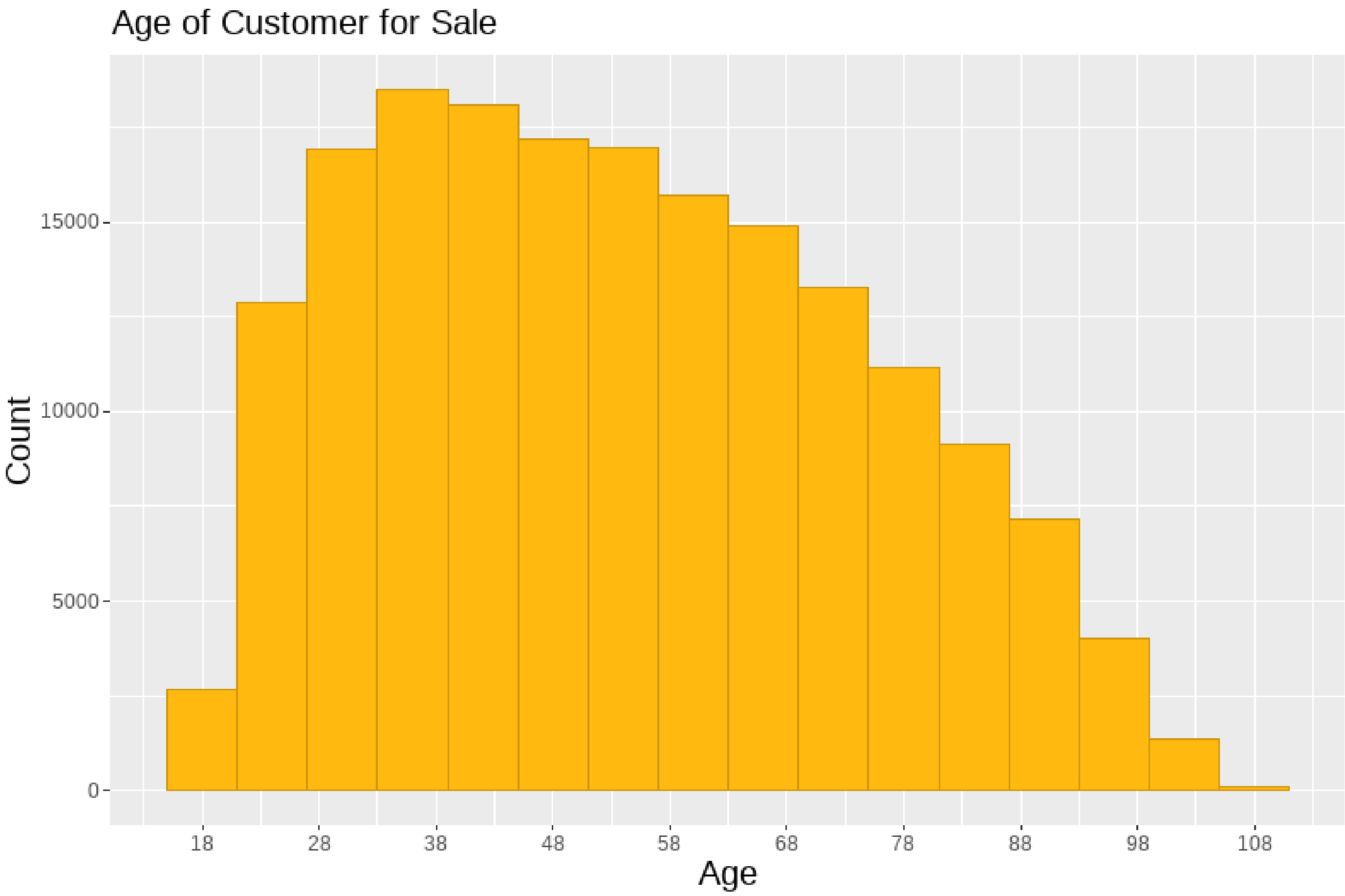


Figure 9, Age of customers for sales

Figure 2. has a unimodal (right-tailed) distribution of sales for the different ages of customers. This is expected with most sales coming from the ages between 28 and 57 years of age where most people have children and spend more before they leave the house and start spending their own money. Work should be down to make it more appealing to the ages between 20 and 28 as these values are lower than the peak and it represents a group of people that are generally not thinking of savings/retirement and do not have children and therefore can spend more of their monthly income.

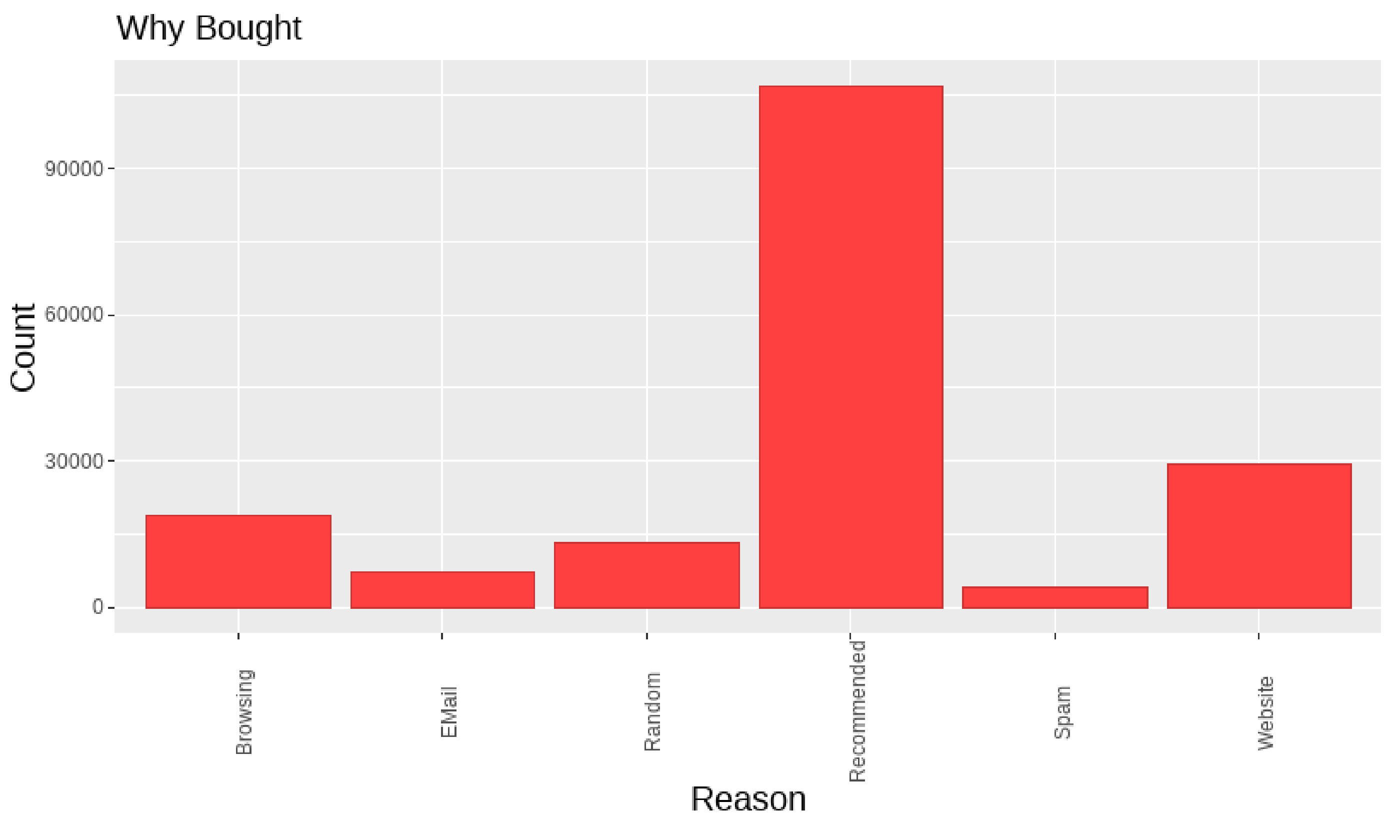


Figure 10, Reason for purchase

Figure 2. allows the business to compare itself to that of other businesses in the same sector and regarding its customers reason for purchases the different class of products from them. From looking at the figure the business can identify the recommended category as its most successful area in promoting its products followed by websites. Money can therefore be focused and more efficiently used to get the most returning and new customers in these areas as well as the areas that other companies are experiencing large amounts from.

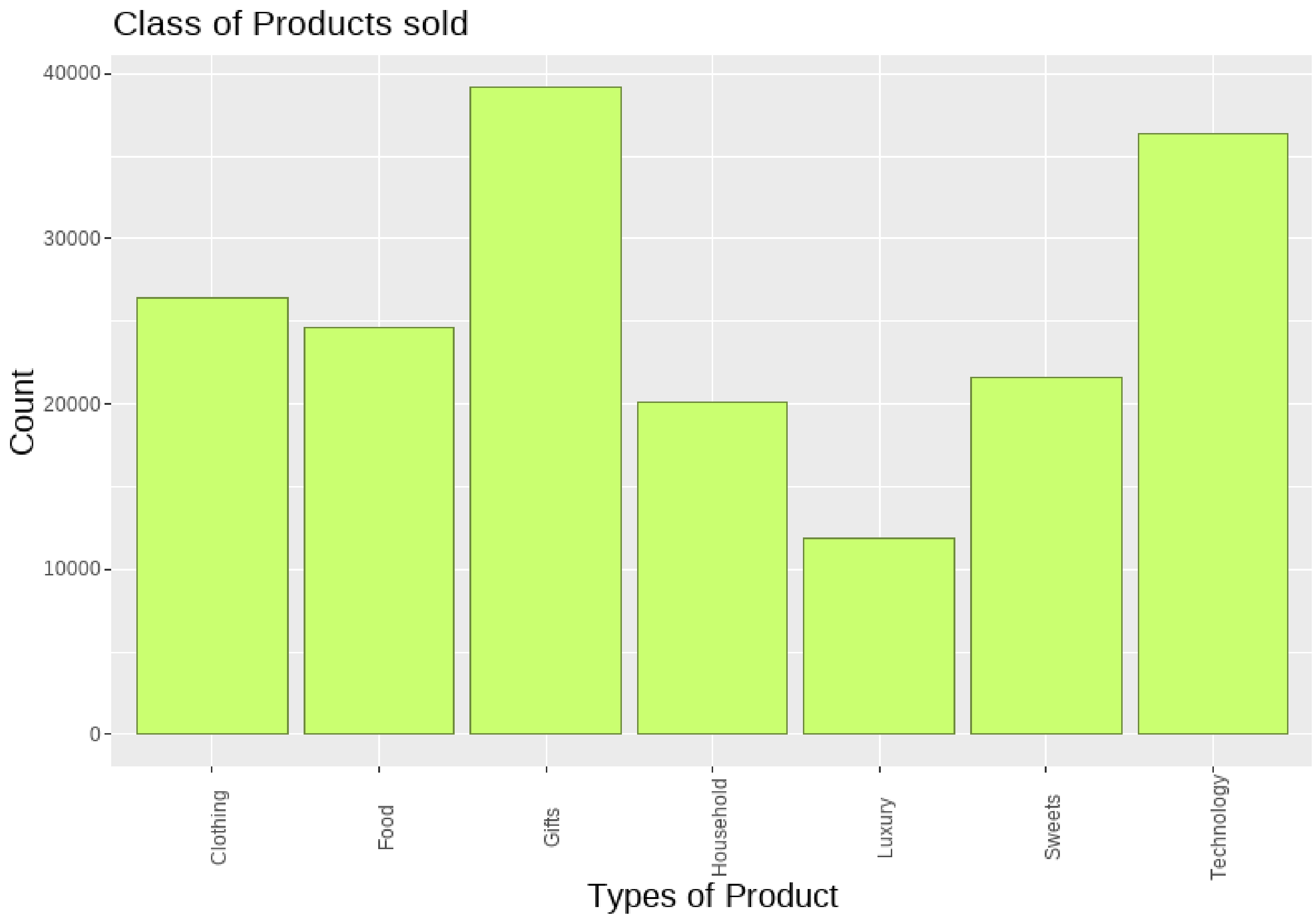


Figure 11, Class of product sold

Figure 2. indicates the classes that the business makes the most sales and what its current customer base targets. This therefore results in better advertisements and forecasting for products. These classes are of gifts and technology.

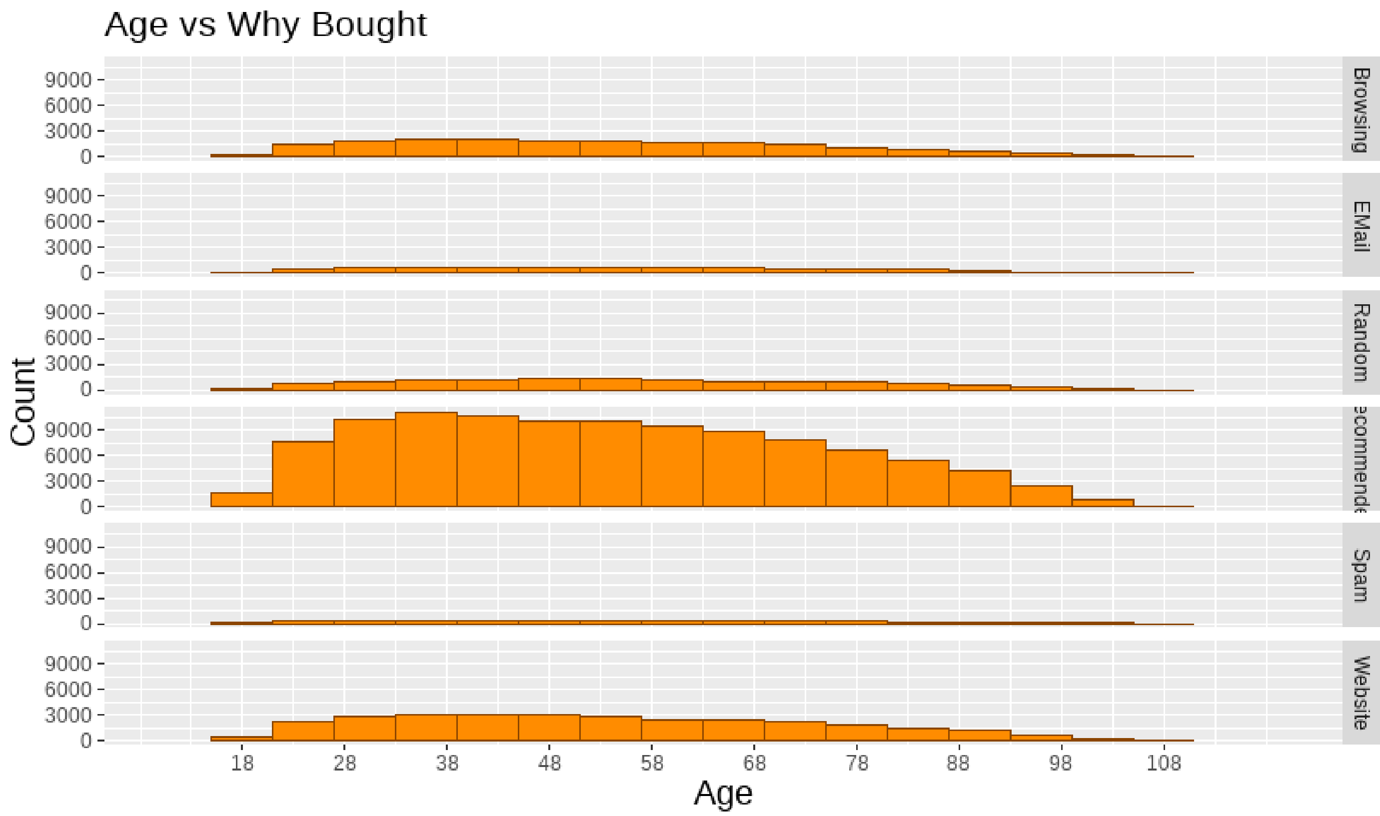


Figure 12, Age vs why bought

Figure 2. confirms the same observation made in the age that customers are in as all classes have a unimodal distribution (skewed right) just varying degree. This figure also confirms that identified in the reason for purchase figure as the categories recommended and website have the largest volume.

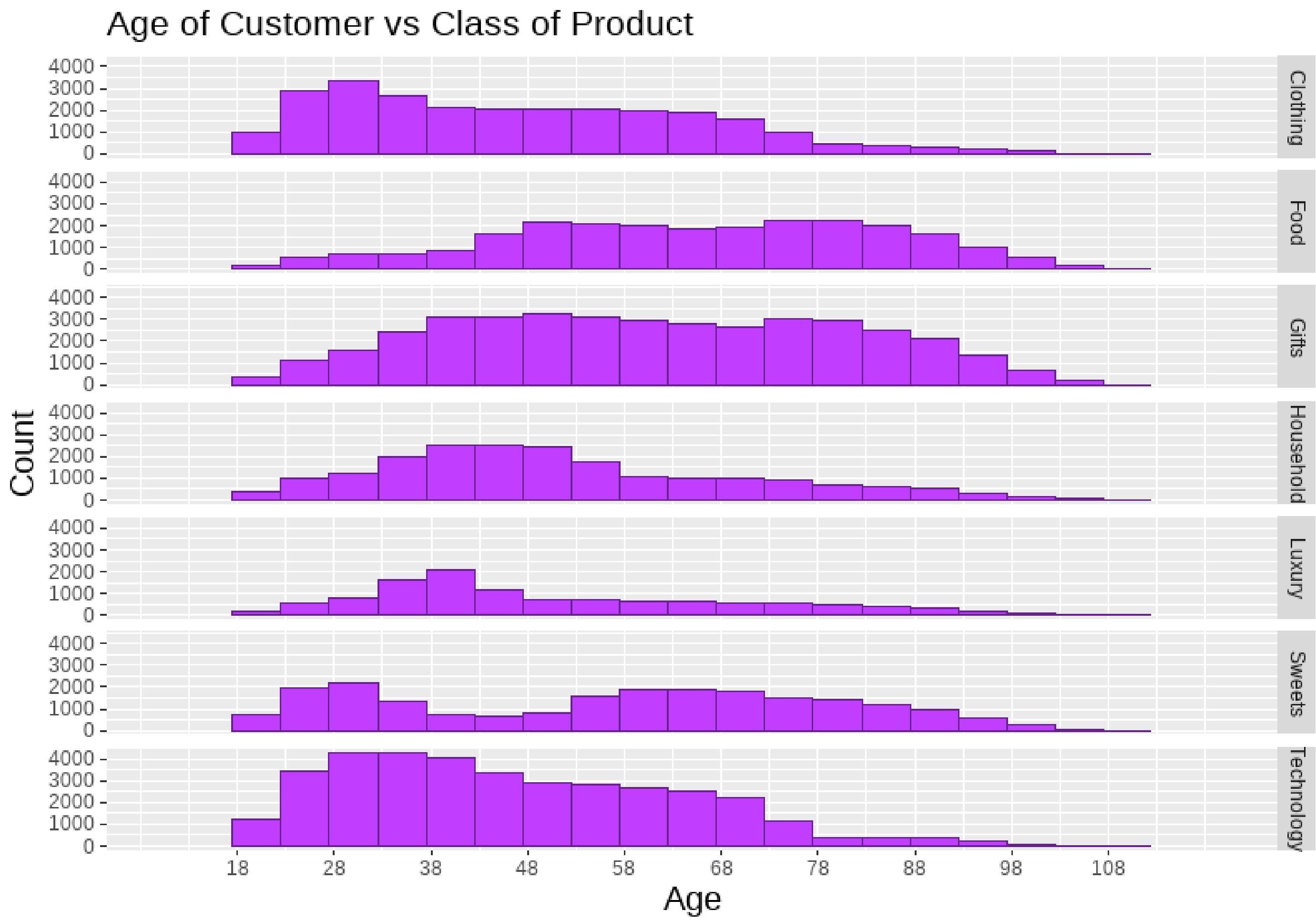


Figure 13, Age of customer vs class of product

Figure 2. allows the business to make age-appropriate advertisements and direct them to these ages as there are clear areas where different ages buy different classes of products. Gifts has the largest volume again and is generally universal in its distribution as most people of all ages buy gifts for one another. The same can be said about the food class as all ned food to survive. As age increases the classes of clothing, household, luxury and technology expectantly decrease as this category of people generally start spending less in these classes as they get old and are more often bought by people with youth and money readily available.

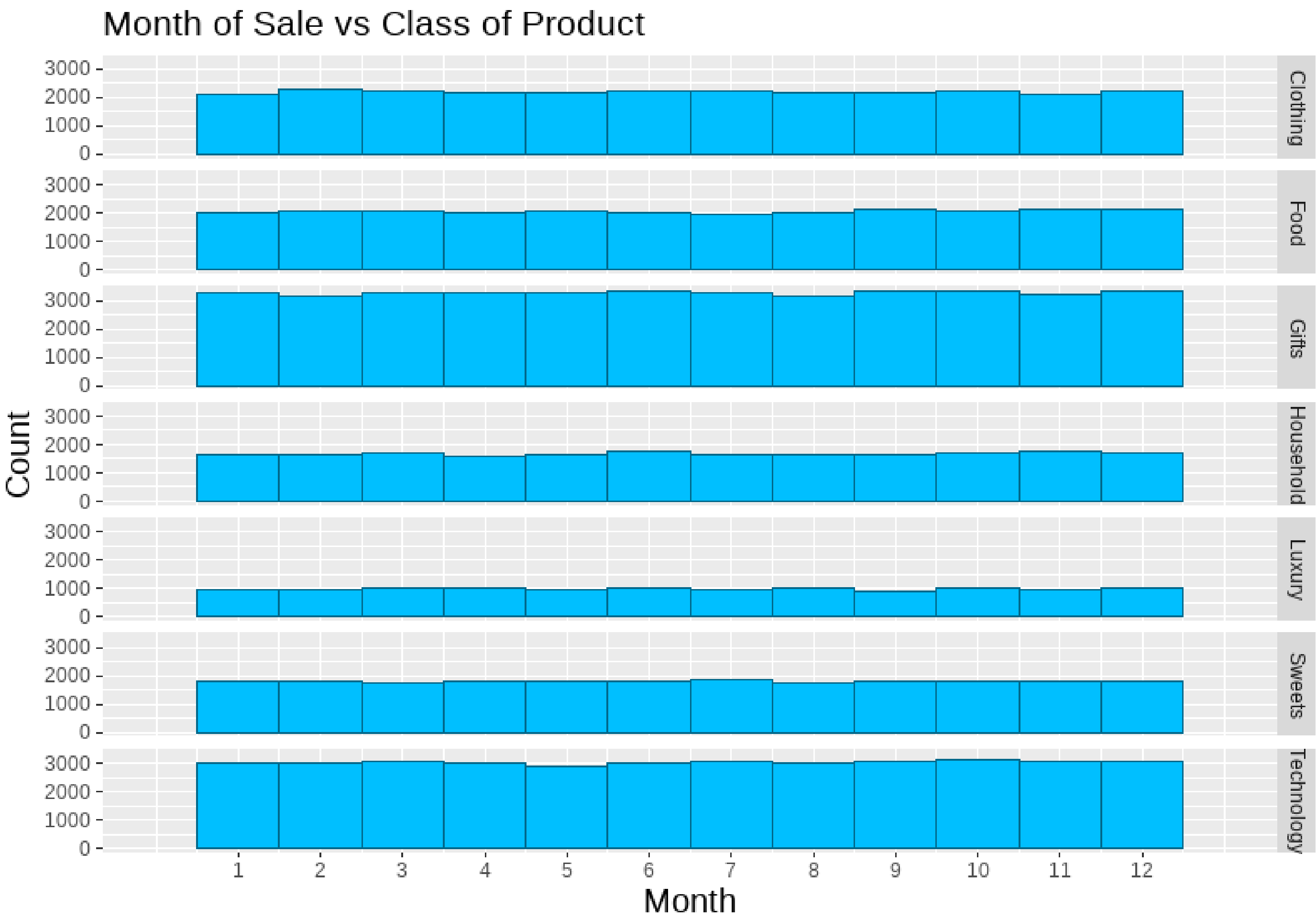


Figure 14, Month of sale vs class of product

Figure 2. further confirms what was identified in sales per month as all classes experience fairly the same volumes of sales through the year. The volumes of sales for the different classes are also confirmed above.

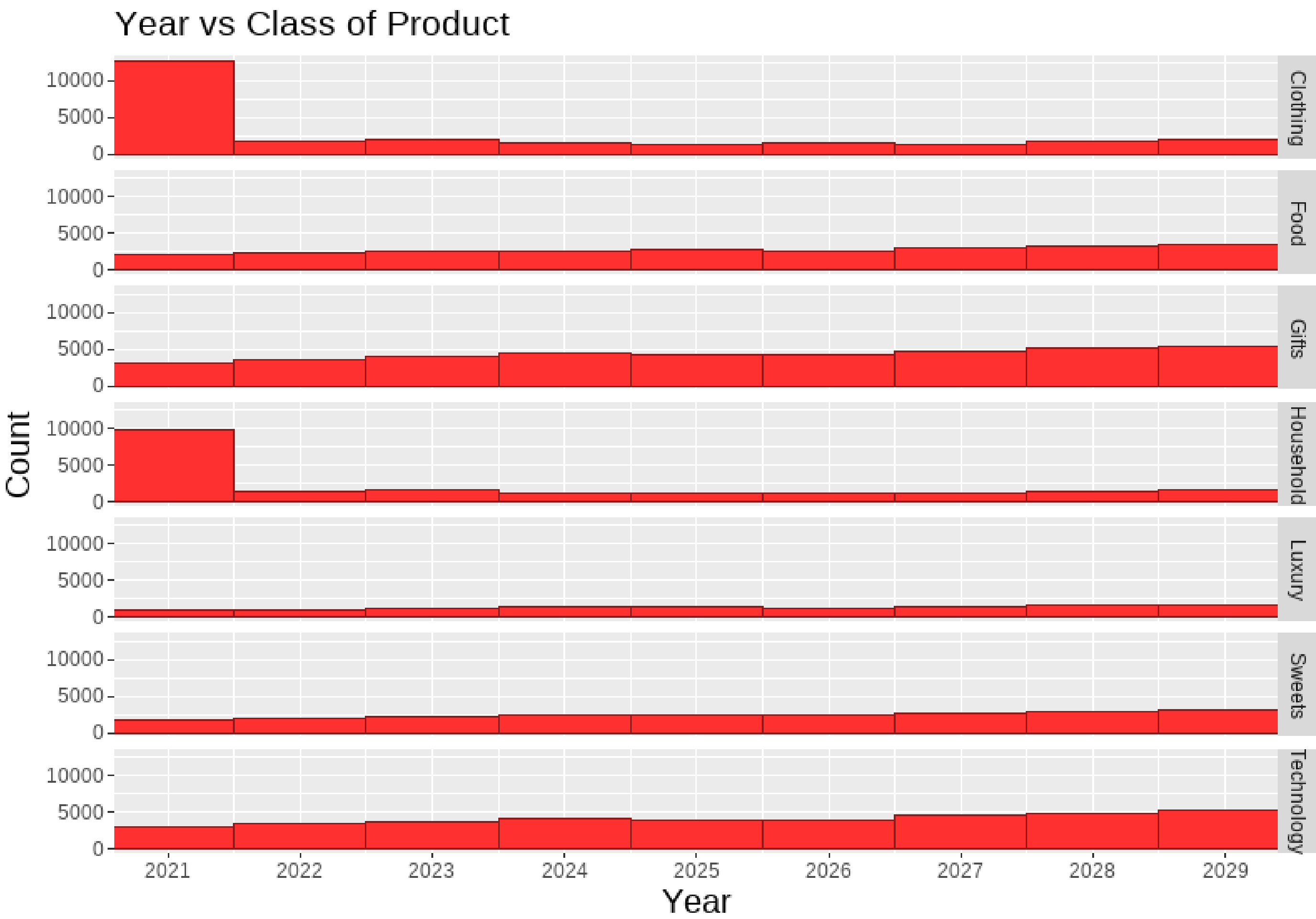


Figure 15, Year vs class of product

Figure 2. further confirms the demand over the years and has gone further in that now one can see that the volumes of products through the years have also been included. The large spike in sales in 2021 can be attributed to that of classes clothing and households as all the other classes experience a general upward trend of sales throughout the years from 2021 to 2029.

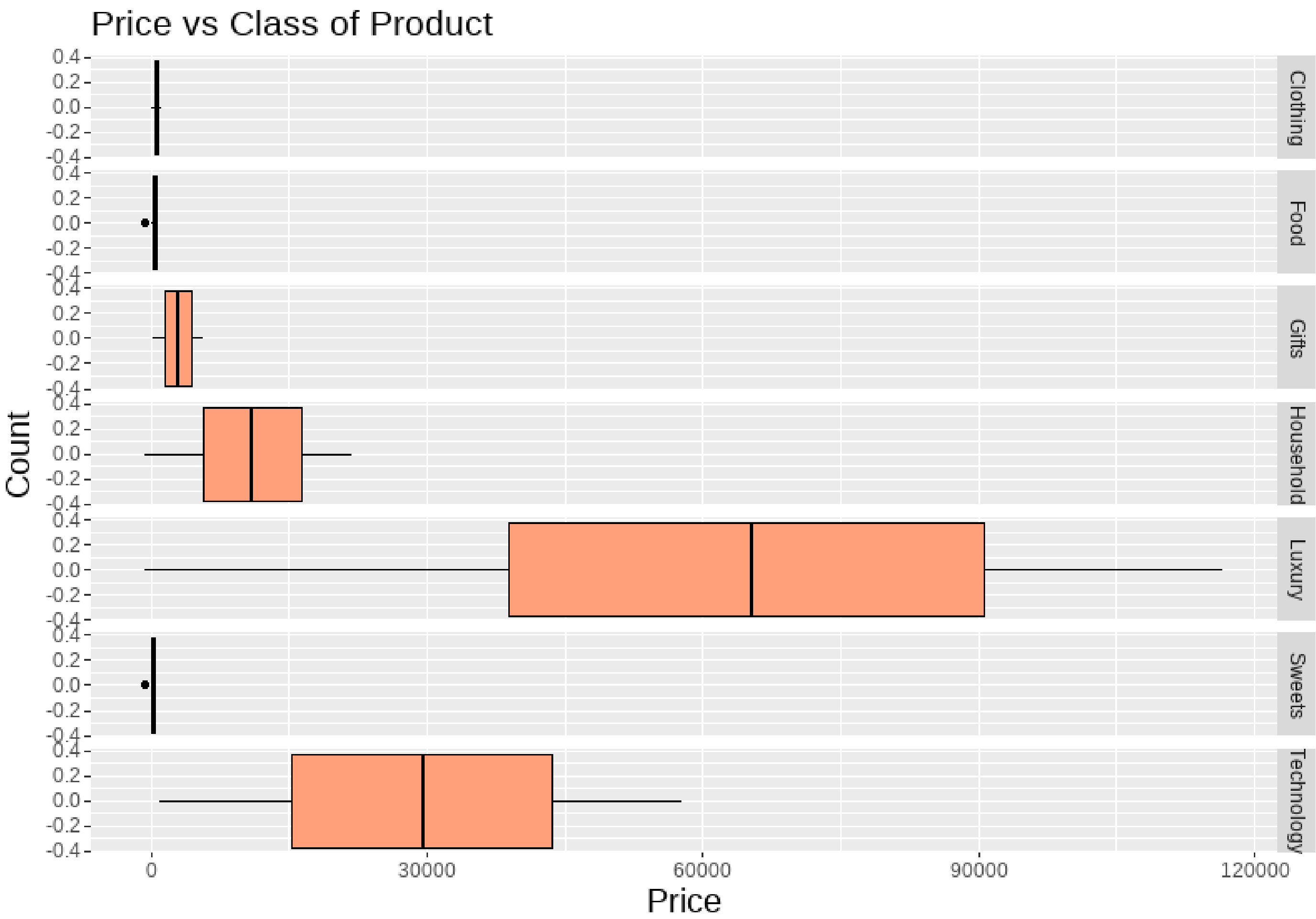


Figure 16, Price vs class of product

Figure 2. indicates that luxury has the highest mean prices amongst classes as well as the largest distribution of prices as it extends from 0 all the way to just over $110 000. This is followed by the technology class with a mean price close to $30 000. The third highest mean is that of the household class products, the highest price is close to $25 000 with a mean of closer to $10 000. This indicates that the products are expensive for most individuals but not for the ultra wealthy.

Figure 17, Delivery time vs class of product

Figure 2. indicates that household products take the longest delivery time and this can be possible from the fact that it is not a necessity such as food. This is followed by technology and gifts. The other classes of products have low delivery times as they are generally kept in stock and are small enough to keep a good level of stock for the demand that they experience.

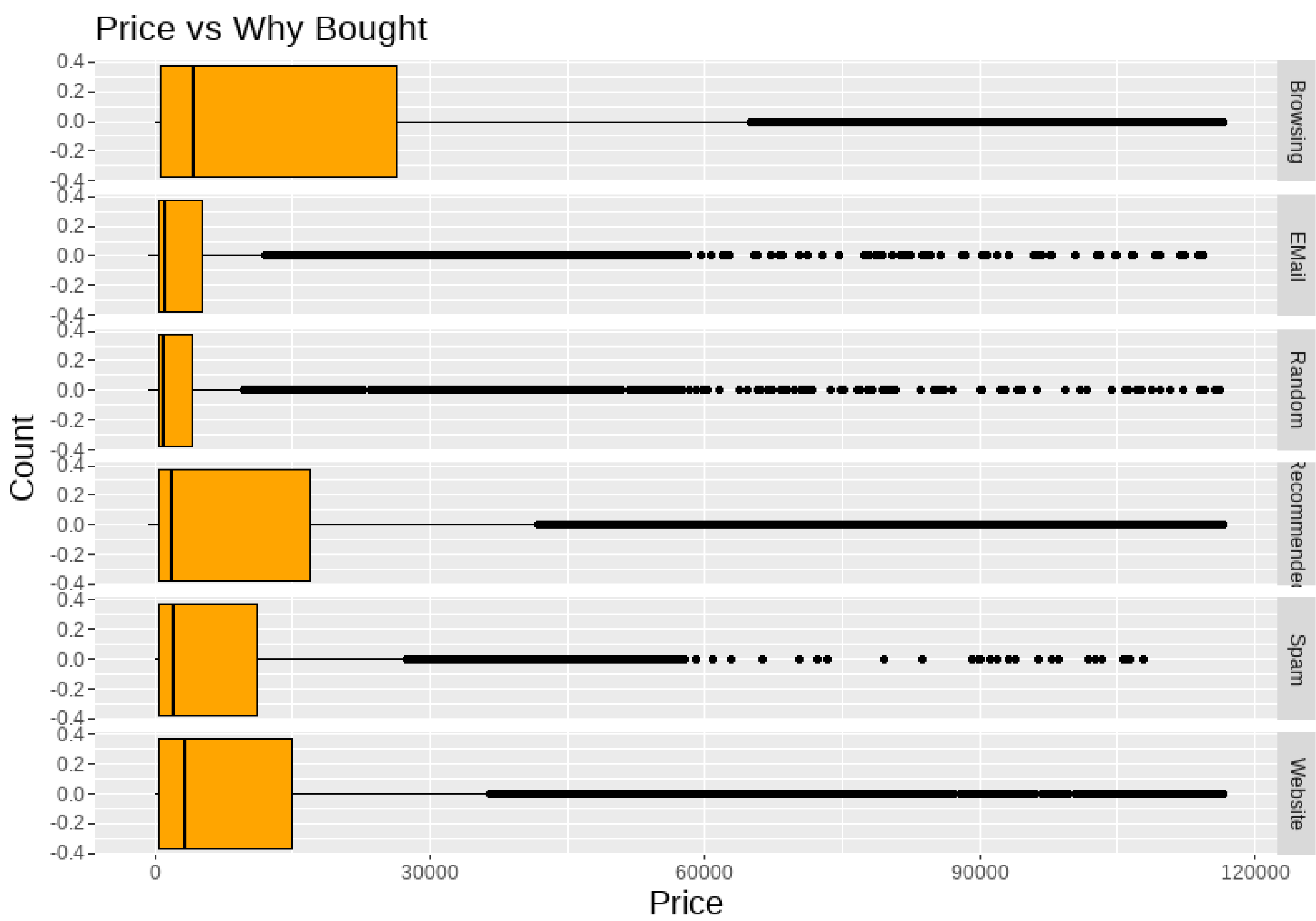


Figure 18, Price vs why bought

The distribution of entries for the price compared to the reason why bought. Most reasons showing that they have a large amount of outliers in the top end of their range

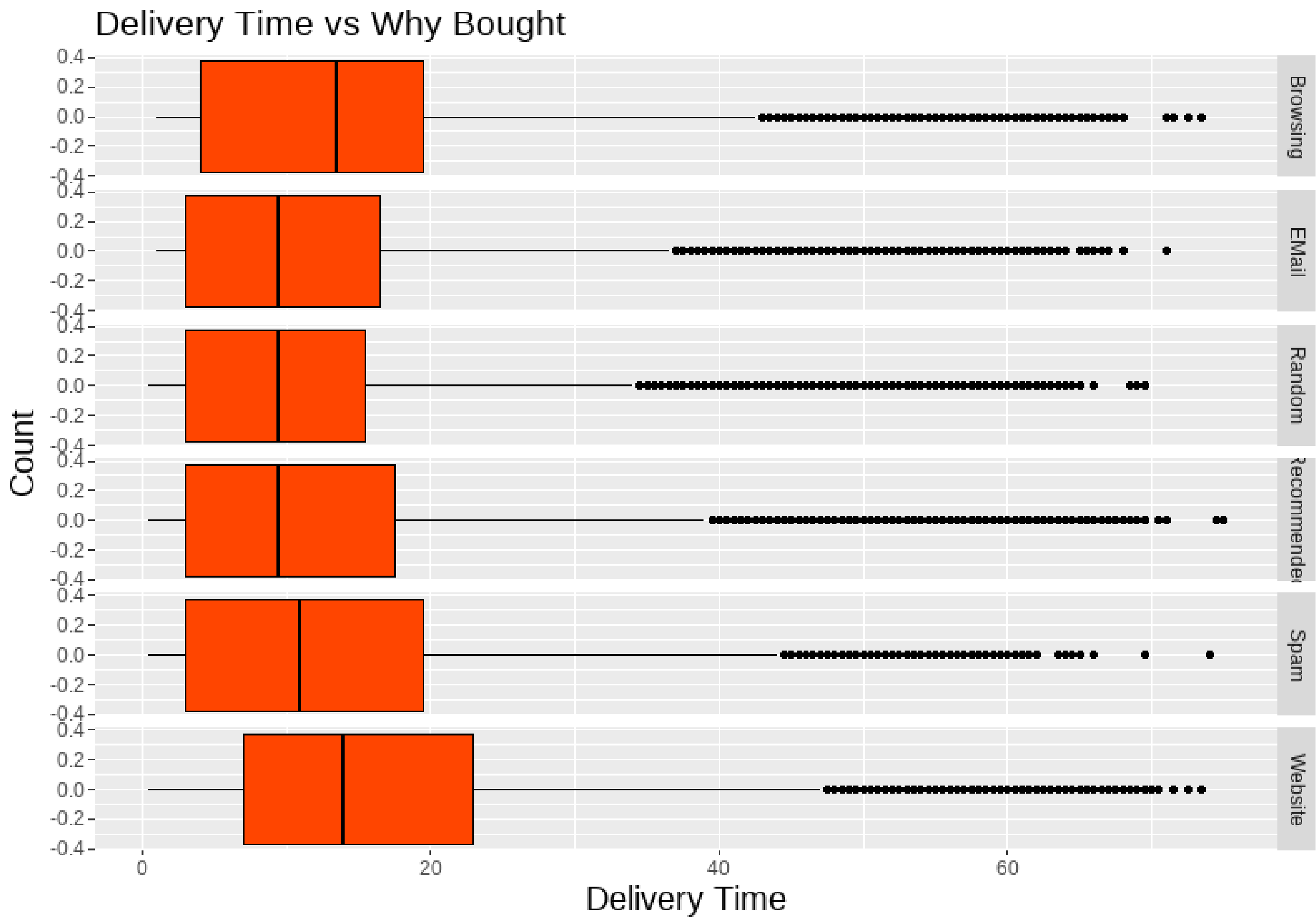


Figure 19, Delivery time vs why bought

The distribution of entries for the delivery time compared to the reason why bought. Most reasons showing that they have a large amount of outliers in the top end of their range

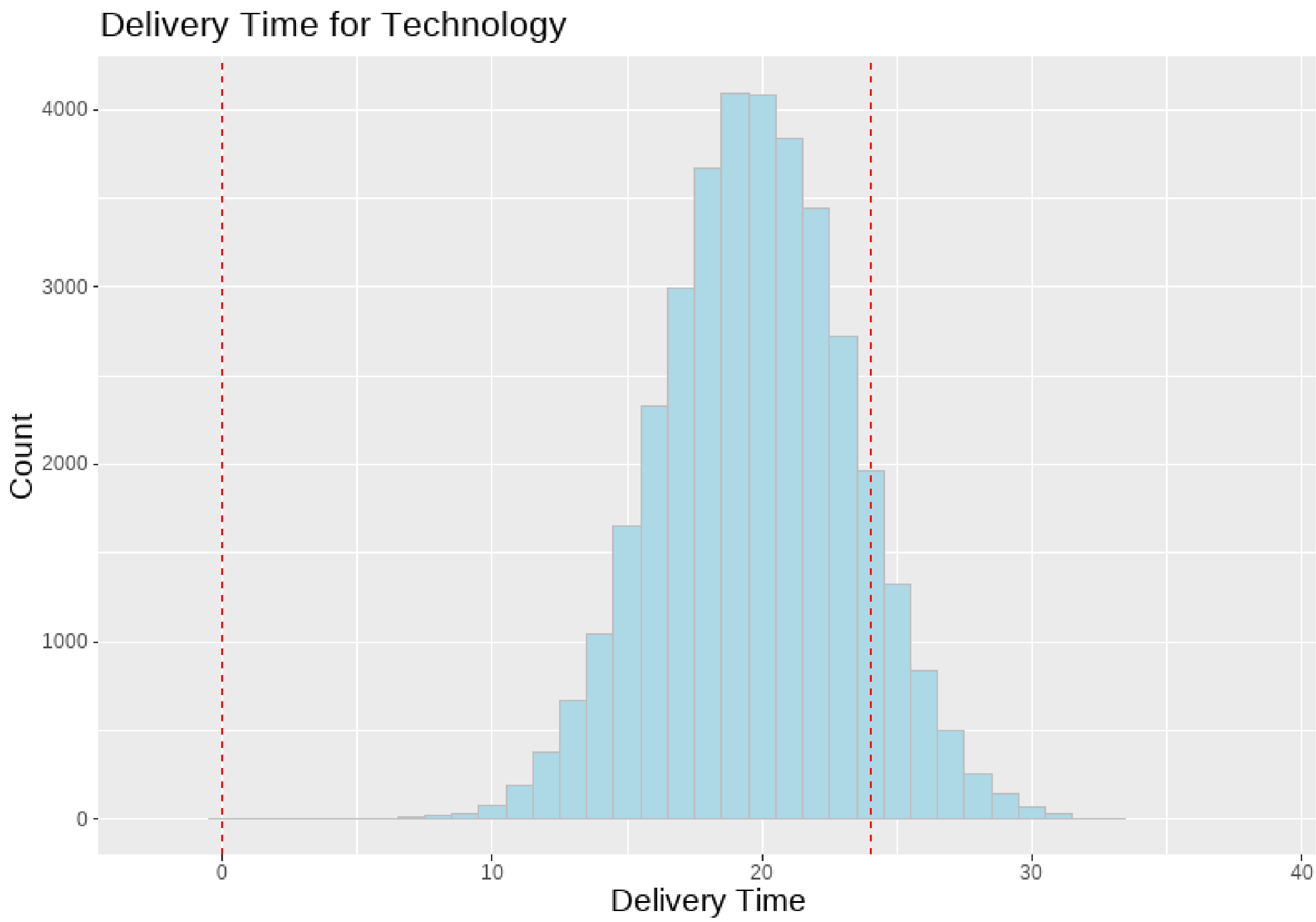


Figure 20, Delivery time for technology

Figure 2. is made using a USL and LSL value of 24 and 0 respectively. A reason for the fact that the LSL is not below 0 is that time can never be below 0 (negative).

USL <- 24

LSL <- 0

Cp <- 1.142207

Cpk <- 0.3796933

Cpu <- 0.3796933

Cpl <- 1.90472

With a Cp value greater than one, it results in the process in being capable. As can be seen in figure 2. that the graph is not centered between the LSL and USL which is a result from the fact that the Cpk is less than the Cp. To improve the process will result in moving the graph so that it becomes centered between the USL and LSL.

3

3.1

3.1 X Chart

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22.9731	22.10688	21.24066	20.37444	19.50822	18.64201	17.77579
Clothing	9.404681	9.259787	9.114894	8.97	8.25106	8.680213	8.535319
Household	50.24618	49.0182	47.79021	46.56222	45.33423	44.10625	42.87826
Luxury	5.493524	5.240868	4.988212	4.735556	4.482899	4.230243	3.977587
Food	2.70933	2.63622	2.56311	2.49	2.41689	2.34378	2.27067
Gifts	9.487909	9.11231	8.73671	8.361111	7.985512	7.609913	7.234313
Sweets	2.896798	2.757124	2.617451	2.477778	2.338104	2.198431	2.058758

Table 1, Delivery process time x-chart

3.2 S Chart

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.179912	4.551784	3.923656	3.295528	2.6674	2.039272	1.411143
Clothing	0.8664496	0.7613819	0.6563142	0.5512465	0.4461789	0.3411112	0.2360435
Household	7.343248	6.452789	5.562329	4.67187	3.781411	2.890952	2.000493
Luxury	1.51086	1.32765	1.144439	0.9612289	0.7780185	0.5948081	0.4115978
Food	0.4371911	0.3841763	0.3311615	0.2781467	0.2251319	0.1721171	0.1191023
Gifts	2.246048	1.973687	1.701326	1.428965	1.156604	0.8842432	0.6118823
Sweets	0.8352331	0.7339508	0.6326685	0.5313862	0.4301039	0.3288216	0.2275393

Table 2, Delivery process time s-chart

Graphs for each class for the oldest 30 samples

Technology

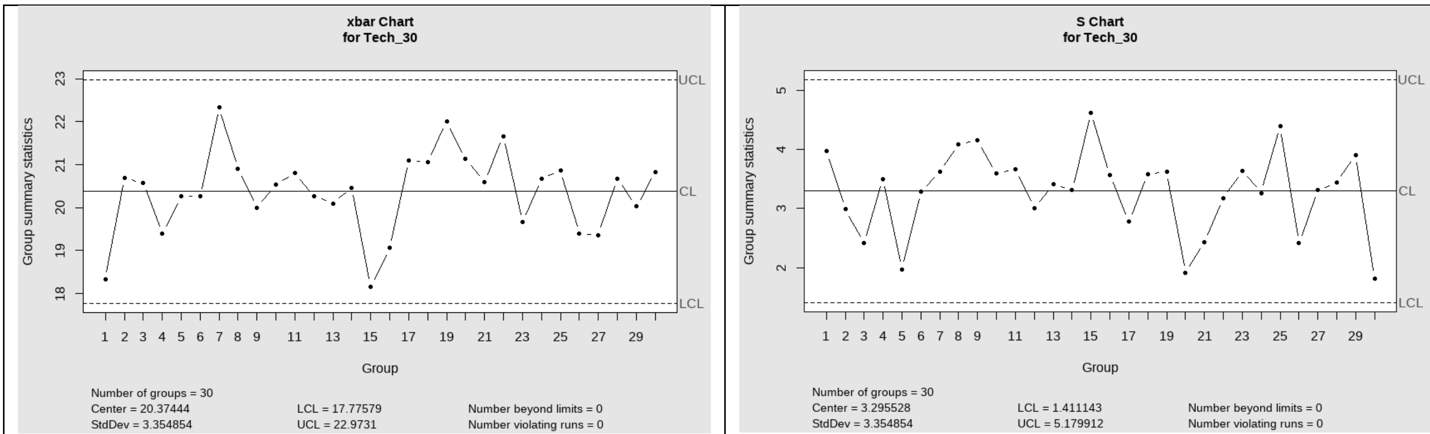
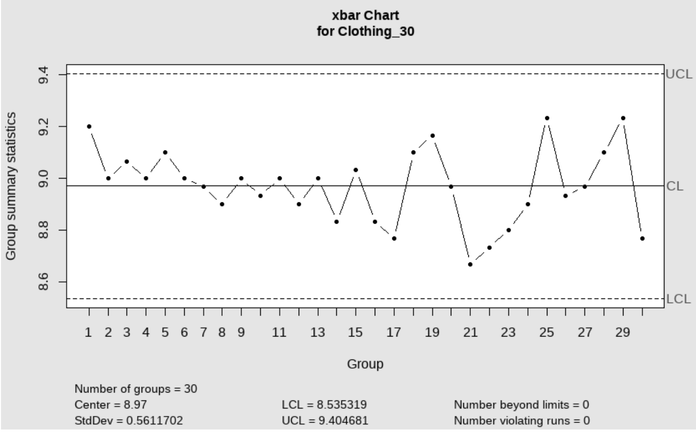


Figure 21, technology mean & standard deviation graph 30 samples

For the first 30 mean samples in the class of technology shows that there are no points outside the control limits for the mean of the samples with most points near the centre line (20.37) with a few random points closer to the control limits (17.78 and 22.97). The standard deviation graph, most points are near the centre with no outliers as expected. All points are random and are in control in the first 30 samples of technology. The standard deviation chart is stable and in control and allows for the mean chart to be evaluated.

Clothing



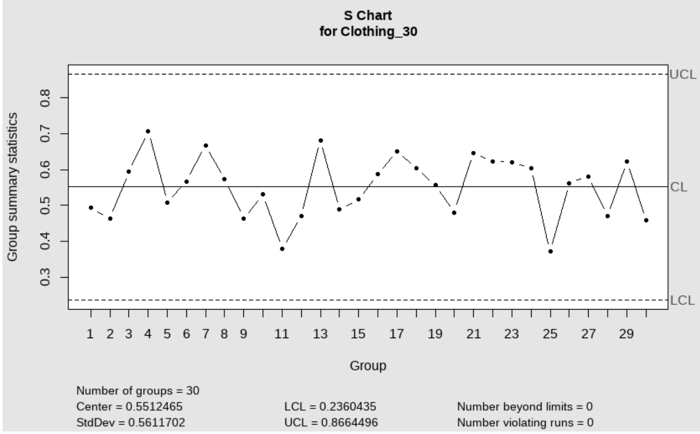
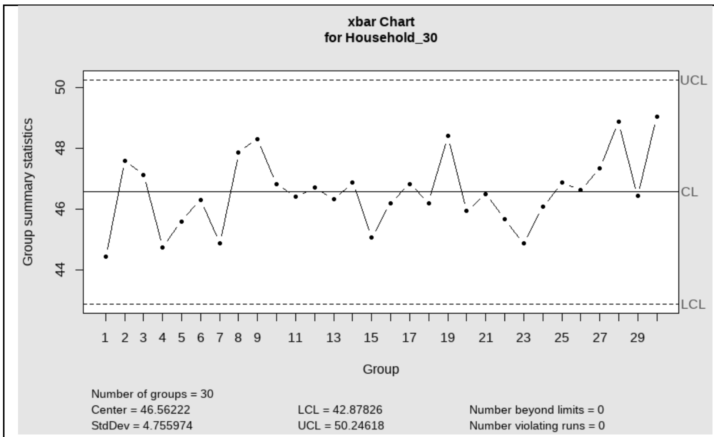


Figure 22, Clothing mean & standard deviation graph 30 samples

For the first 30 mean samples in the class of clothing, many are located near the centre line with a few points closer to the control limits. This is further proven in the standard deviation of the corresponding sample points as the centre is a small value (0.55) compared to that of the other categories. The standard deviation gives the impression that the data is not out of control and does not have much variation resulting in the mean graph being evaluated.

Household



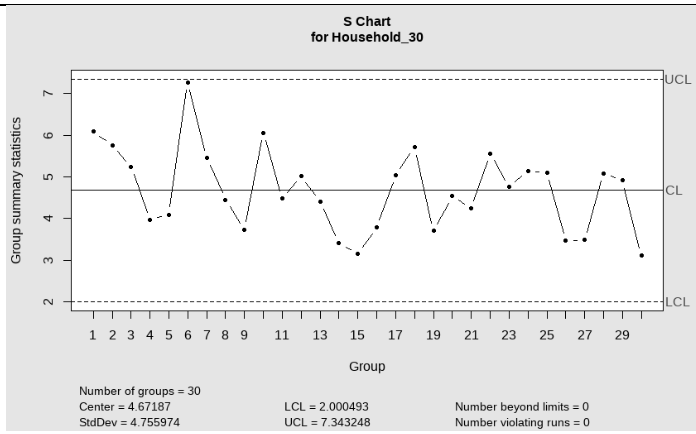


Figure 23, Household mean & standard deviation graph 30 samples

For the first 30 mean samples in the class of household it gives the impression that it is in control as most points are close to the centre line with a few between it and the control limits. There are no points beyond the control limits in either of the two graphs and therefore all is in control.

Luxury

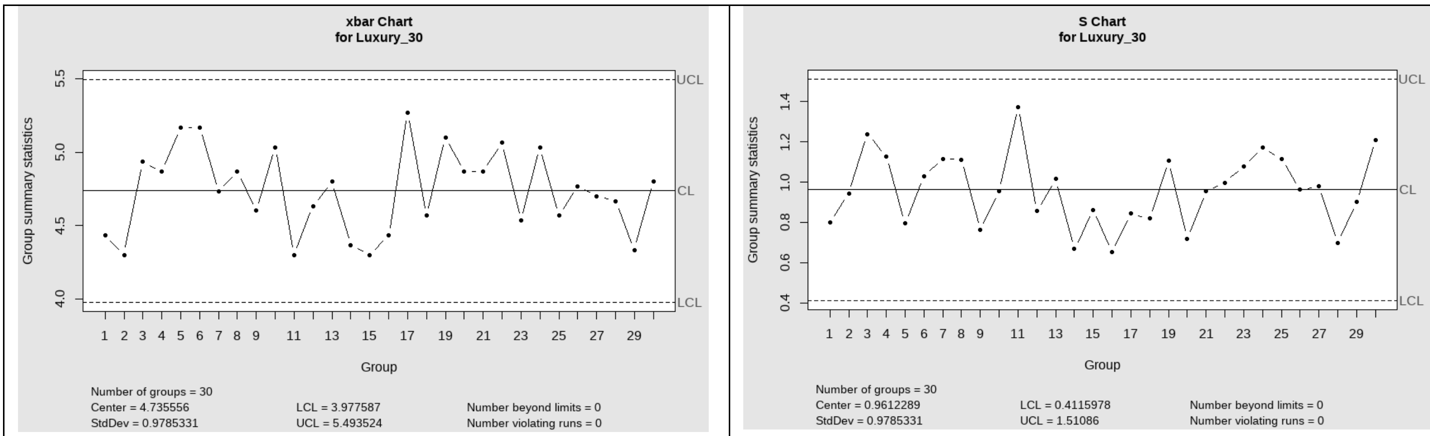


Figure 24, Luxury mean & standard deviation graph 30 samples

For the first 30 mean samples in the class of luxury the sample points are spread out more than the previous graphs but there are no concerns as all points in both graphs are within the control limits. The mean graph has roughly equal distribution of points above and below its centre line (4.74) with most points being within 1.5 sigma from the centre line therefore showing a good distribution of data in the luxury class. The luxury class can be considered in control.

Food

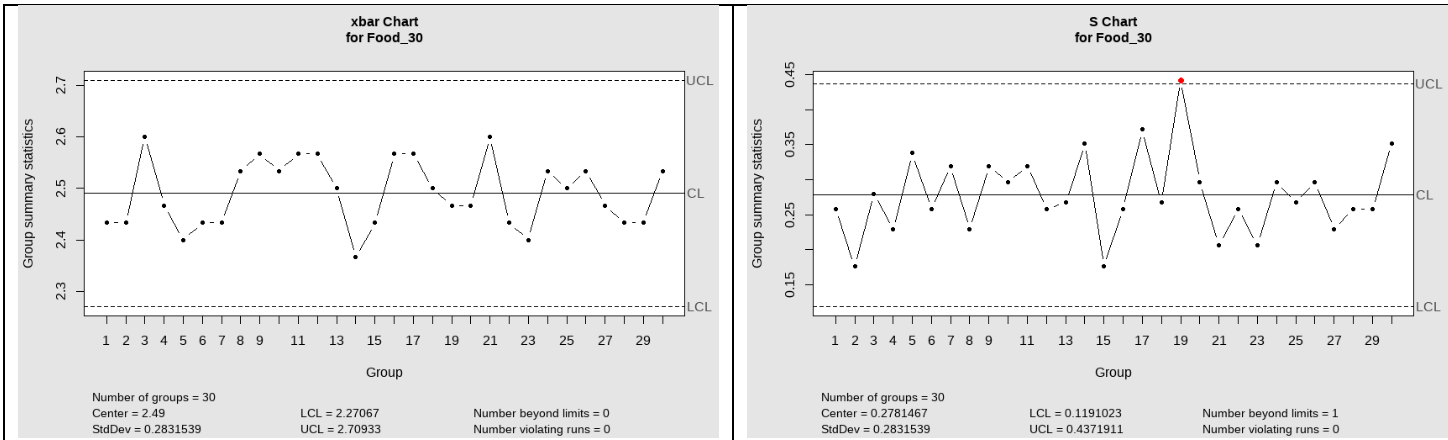


Figure 25, Food mean & standard deviation graph 30 samples

For the first 30 mean samples in the class of food, all seems in control with even amounts of points above and below the centre showing a good distribution of data in the food class. When looking at the standard deviation graph one notices that sample point 19 is beyond the limits and therefore needs to be removed.

Gifts

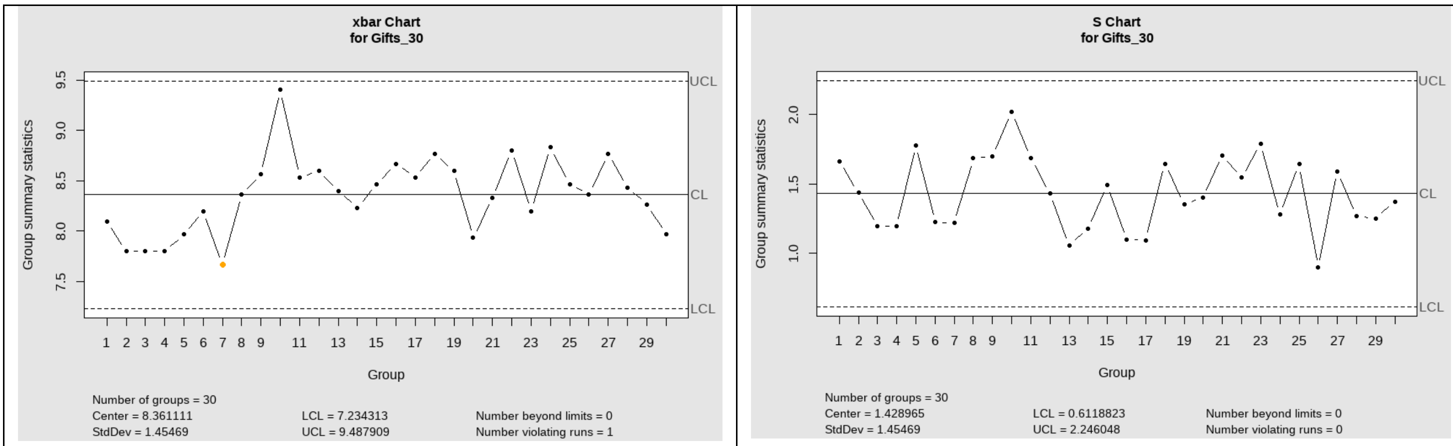


Figure 26, Gifts mean & standard deviation graph 30 samples

For the first 30 mean samples in the class of gifts, the mean graph has no sample points outside the control limits and most points are close to the centre line with a few closer to the control limits. There is also a similar distribution of points above and below the centre line. The standard deviation graph shows that the sample pints are all within the control limits and have a good distribution of their points resulting in the mean graph being used for evaluation purposes.

Sweets

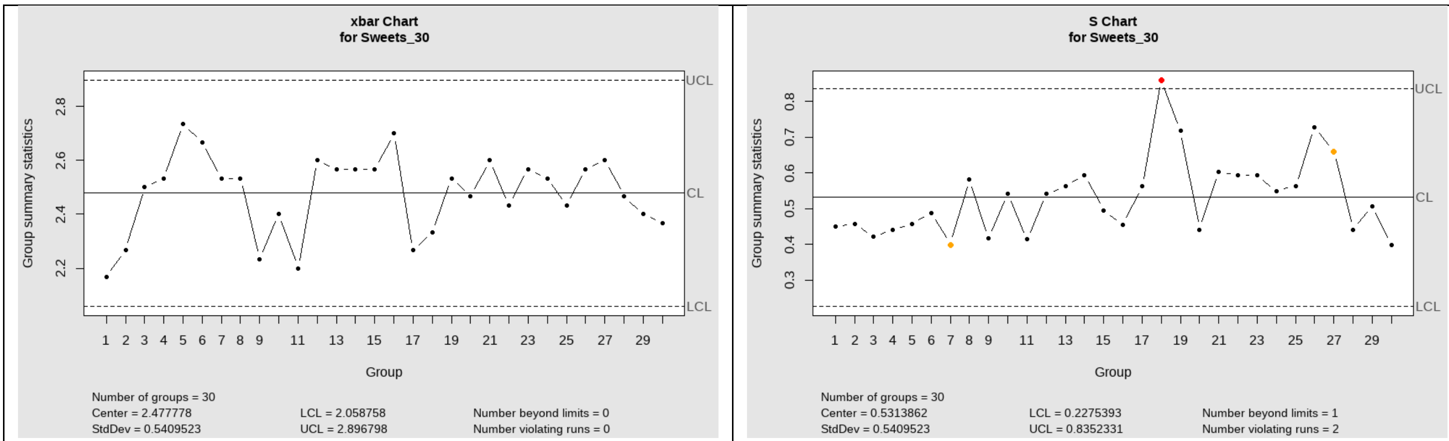


Figure 27, Sweets mean & standard deviation graph 30 samples

For the first 30 mean samples in the class of sweets, the mean graph has no sample points outside the control limits but in the standard deviation graph, sample number 18 is outside the control limits and needs to be removed. Besides this point the rest of the points look fairly well distributed.

3.2

Technology

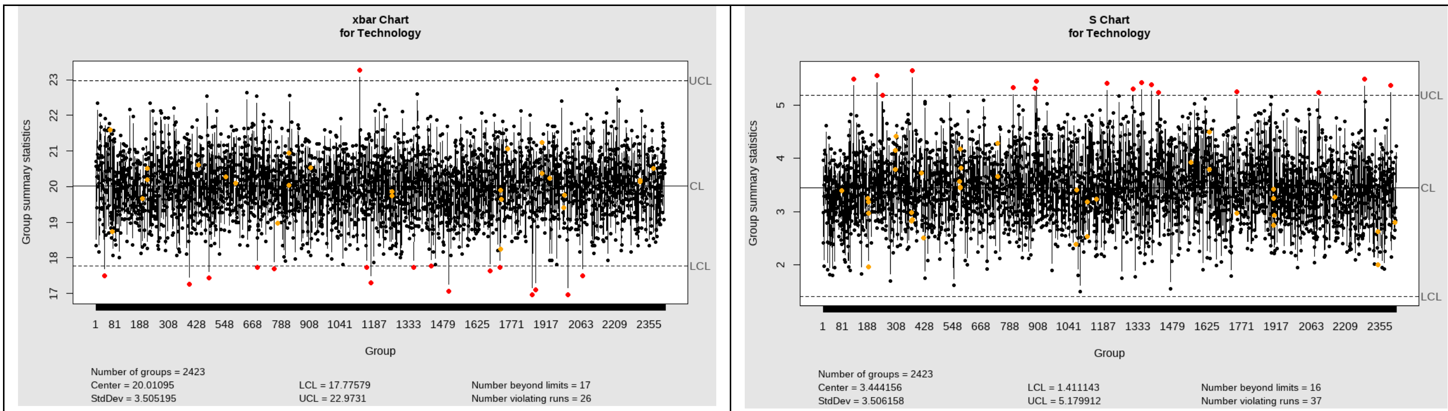


Figure 28, Technology mean & standard deviation graph all samples

A good distribution of data that is in the samples as many are close to the centre line with a healthy amount of points closer to the control limits. There are only 17 points; which represents 0.7% of the total samples; that are beyond control limits from the 30 samples and gives the impression that the data is in control. The standard deviation graph gives the same impression.

Clothing

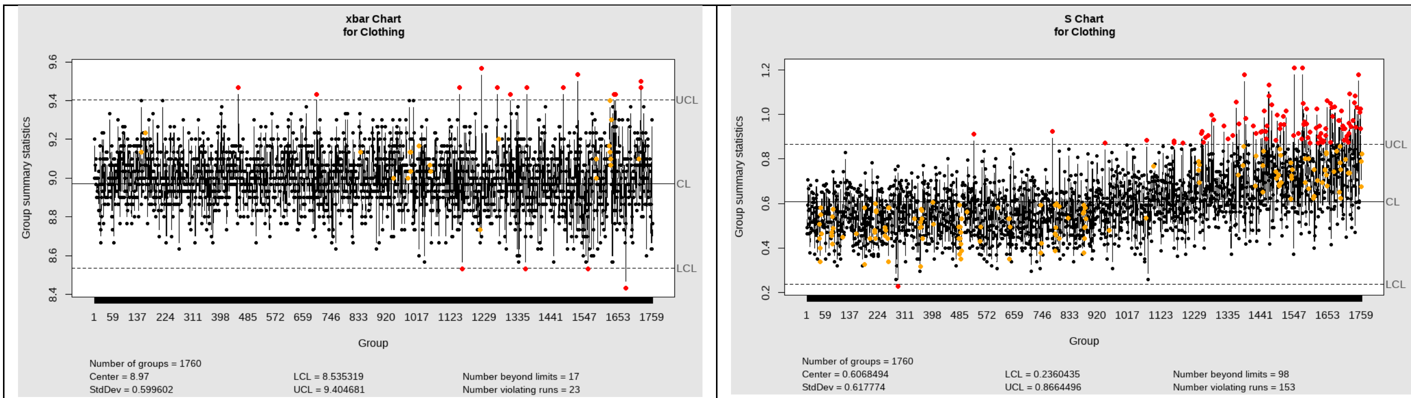


Figure 29, Clothing mean & standard deviation graph all samples

The clothing mean sample points resemble that of technology but closer look at the standard deviation graph indicates that the more resent clothing purchases show a large consistent trend of exceeding the standard deviation control limits set by its 30 sample data.

Household

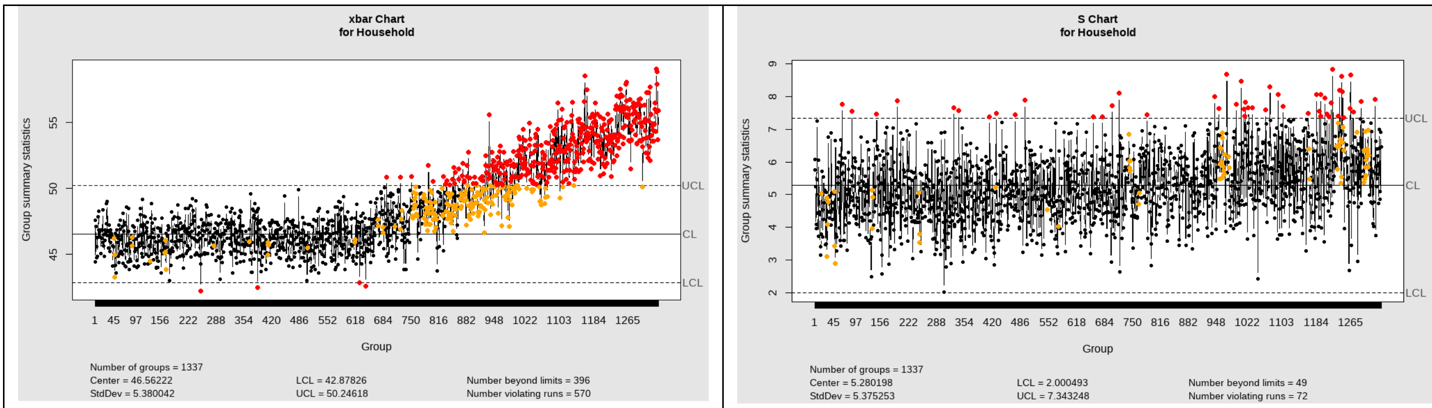


Figure 30, Household mean & standard deviation graph all samples

The household mean points give a good representation of random points with few exceeding the control limits until point number 760. From here onwards there is a continuous upward trend of the data that as the samples become newer/younger (samples ordered oldest to newest based on year, month, day, x) the delivery time increases. This can be seen in the mean figure above where all the red points represent samples that exceed the limits. This needs to be investigated as it most likely results in transportation costs increasing from sample number 760 and therefore decreasing profit. As a result of a large number of mean sample points above its UCL it results in the standard deviation shifting its centre value up. The delivery time for household products is not stable and not in control from the information provided in the two graphs above.

Luxury

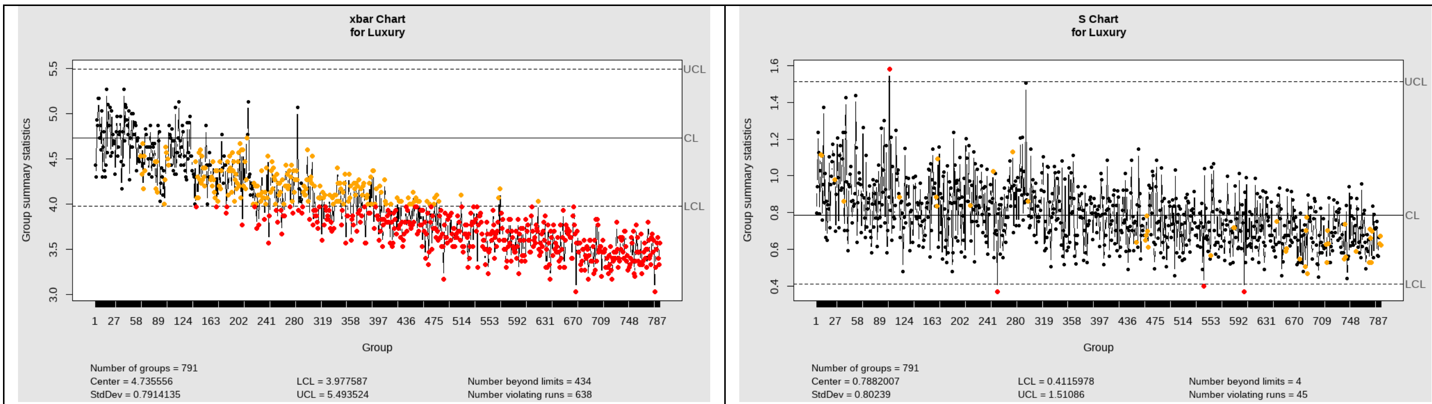


Figure 31, Luxury mean & standard deviation graph all samples

Luxury mean samples' graph indicates a continuous downward trend of the samples' delivery costs. As can be seen that over half the samples (434) extend below the LCL set by its 30-sample experiment. This is a positive trend as the delivery cost continuously decreases and will likely result in an increase in profit. The experiment is clearly out of control with the amount of sample points exceeding the limits. The standard deviation graph represents a fairly good graph seeing that only 4 points exceed the control limits but there is a large difference in UCL and LCL values as the UCL is over 50 percent larger than LCL with respect to the new centre line. These points also exhibit a downward trend of point values as they are mostly between the centre line and the LCL near the end samples. Both may be attributed to the fact that the mean samples tend to be closer spaced as they near the end of the samples.

Food

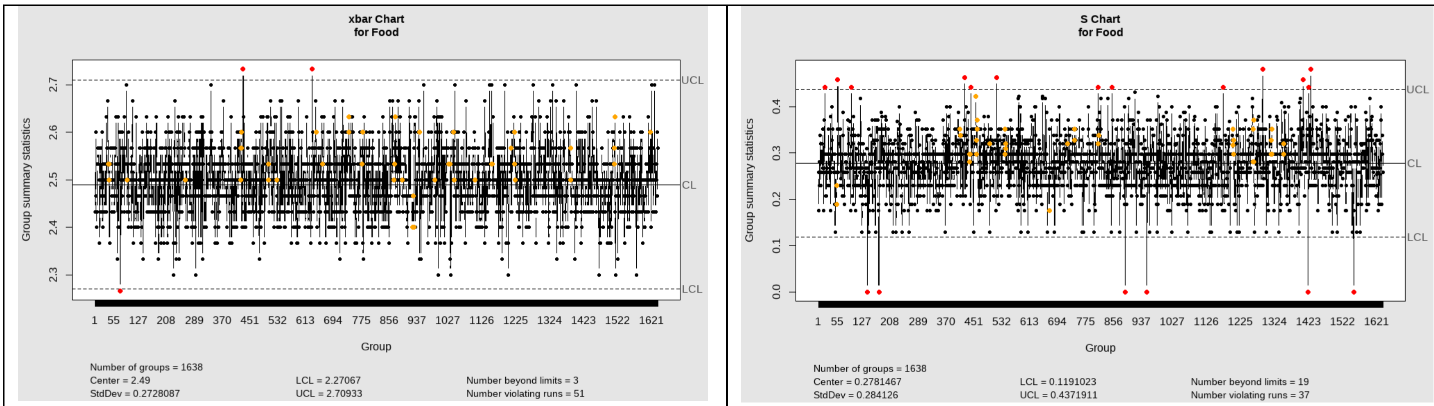


Figure 32, Food mean & standard deviation graph all samples

Food samples represents a continuous trend of having mean sample points scattered between the centre and the control limits with points re-occurring with the same values and can be seen forming lines as time goes on. The delivery costs have been kept fairly similar as the years have gone by and the small amount of sample points outside the control limits (3) indicates the food delivery costs are in control. The standard deviate graph has a very similar layout to that of its mean graph with a slight increase of sample points that exceed the control limits (17) and more points being closer to the centre lines.

Gifts

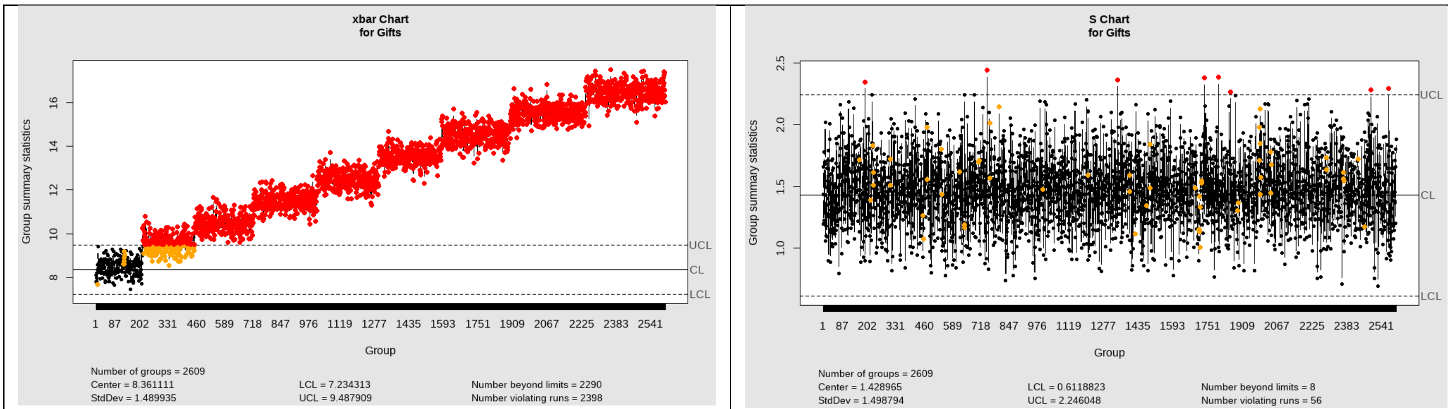


Figure 33, Gifts mean & standard deviation graph all samples

The gifts’ mean delivery time shows a continuous upward trend of re-occuring trends that represent phases as time goes on in the delivery cost. As the sales data is from 2021 to 2029 and there are nine ‘blocks’ of repeating form, each ‘block’ of the mean sample points represents one year and can be seen that each following year has a consistent increase in its mean cost. This continuous upward trend leads to sample points being outside the first 30 sample points derived UCL very quickly as there are 2290 out of 2609 samples that are beyond limits. When compared to the 30 sample point limits the data is out of control by the amount of sample points beyond control limits. A potential solution is to compare trends from years separately and not all data together. The standard variation of all points shows points randomly placed around the centre line with very few outside the control limits (8) and decreasing amounts of sample points as the control limits are approached which is good.

Sweets

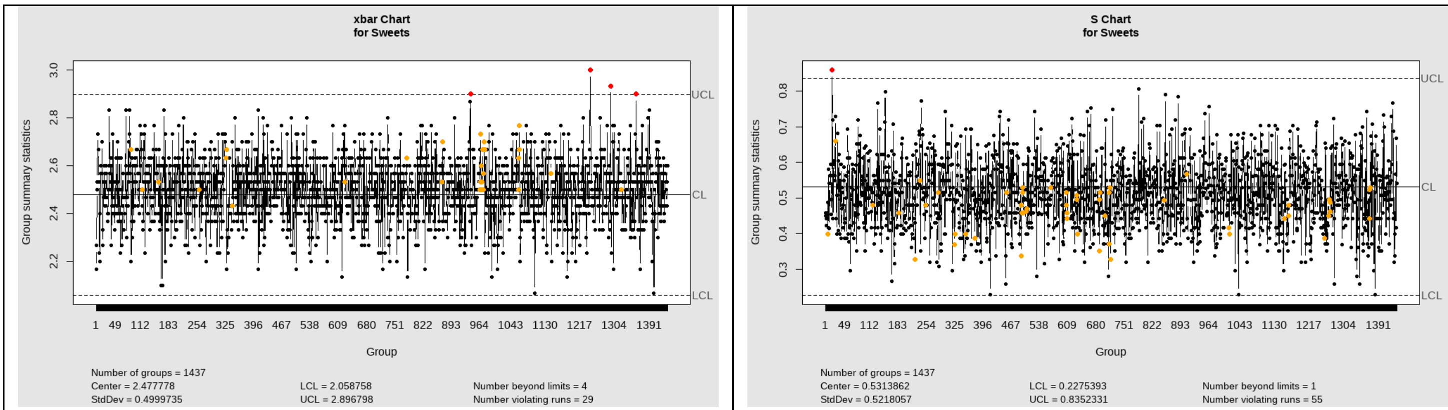


Figure 34, Sweets mean & standard deviation graph all samples

The sweets mean sample points resemble the same type of graph from the food class in that there are points throughout the sample that have the same value and start to form lines the more there are of that value. There are less random values as they are repeated but there is a sense that there is a small variation in their spacing which could suggest that there are products (that fall under sweets) that cost the same to deliver that are ordered fairly consistently with a few random orders. By the low number of outliers (4) in the mean sample points it indicates that the class of sweets can be considered in control. The standard deviation of sample points further confirms that the class of sweets is in control as there is only one outlier with most points located close to the centre line and decreasing as they get closer to the control limits. There is a very faint period trend in the standard deviation graph that can be caused varying product delivery cost due to different products being ordered at different times of the year.

4

Class	Total	1st Outlier	2nd Outlier	3rd Outlier	3rd last Outlier	2nd last Outlier	Last Outlier
Technology	17	37	398	483	1872	2009	2071
Clothing	17	455	702	1152	1677	1723	1724
Household	396	252	387	629	1335	1336	1337
Luxury	434	142	171	184	789	790	791
Food	3	75	NA	NA	NA	432	633
Gifts	2290	213	216	218	2607	2608	2609
Sweets	4	942	NA	NA	1243	1294	1358

As was identified in part 3, categories technology, clothing, food and sweets are in control as they all have low numbers of mean sample points outside the control limits (outliers) whereas household, luxury and gifts are clearly out of control with their high number of outliers. These can represent problems within the business and need to be investigated whether they are justifiable or not.

4.1.A

Only the outlier graphs from categories that have been identified of being out of control are plotted below. The deliveries for technology, clothing, food and sweets are expectable minimal deliveries that will be outside the control limits due to rare conditions that can be examined after the out-of-control categories are thoroughly investigated as they have large amounts of deliveries that are outside the control limits and therefore have delivery times that are not within the specified range between the control limits.

Household

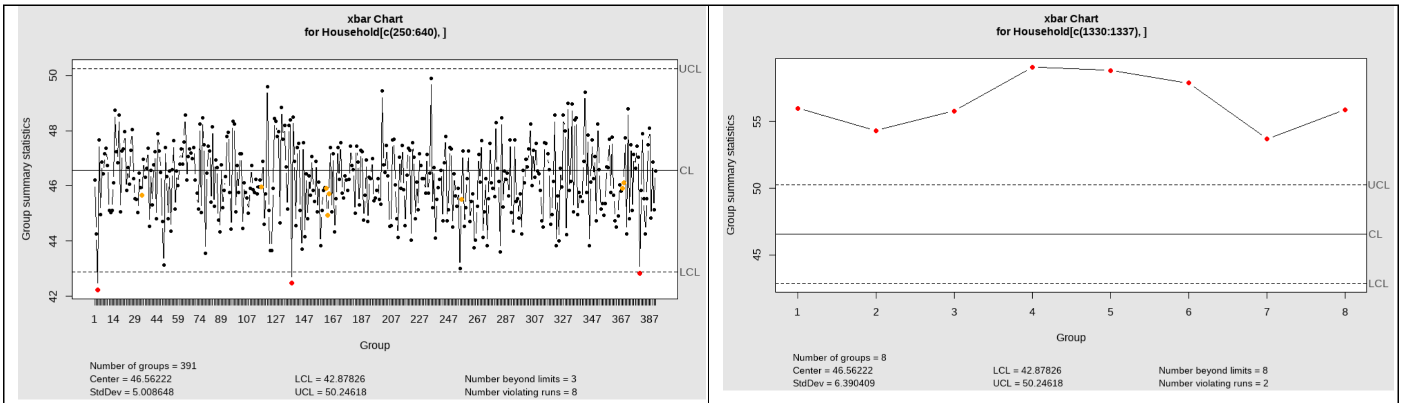


Figure 35, Household first 3 & last 3 outliers

The first three outliers seem like all are random whereas the last three outliers are all consecutive, above the UCL as well as the range of sample points that precede them.

Luxury

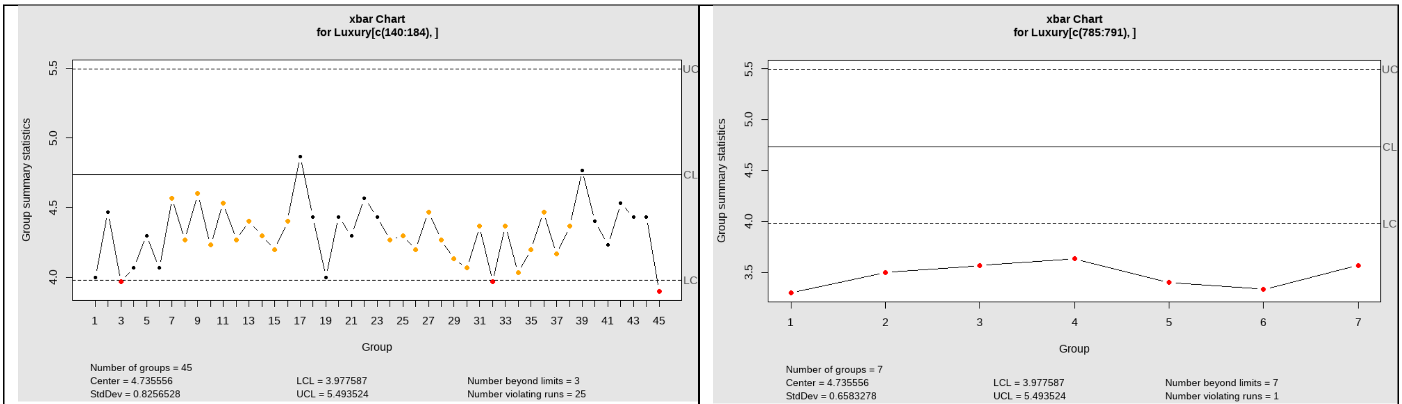


Figure 36, Luxury first 3 & last 3 outliers

The first three outliers seem like all are random whereas the last three outliers are all consecutive, below the LCL as well as the range of sample points that precede them.

Gifts

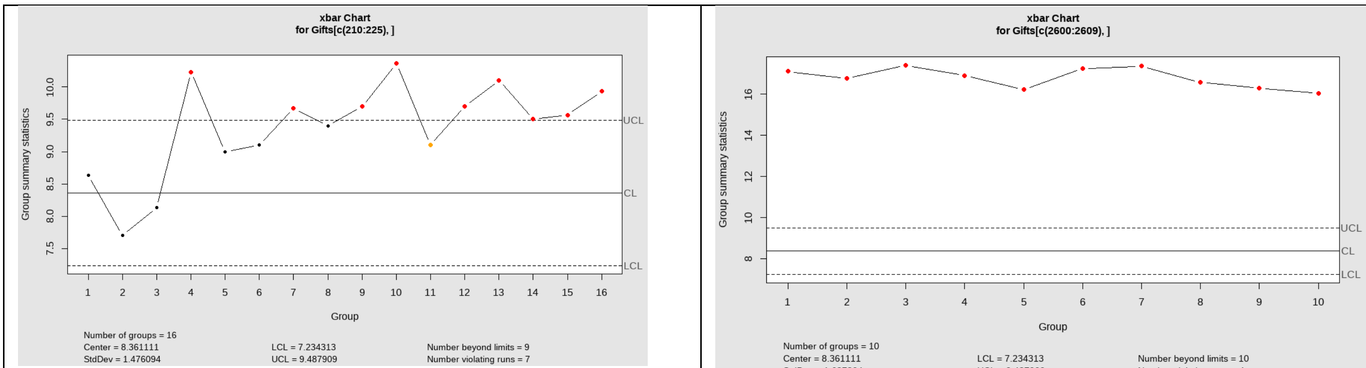


Figure 37, Gifts first 3 & last 3 outliers

The first three outliers do not seem random as they are spaced very close to one another and the last three outliers are all consecutive, significantly above the UCL as well as the range of sample points that precede them.

4.1.B

Most consecutive samples from the standard deviation sample data between -0.3*sigma and 0.4*sigma control limits

Class	Maximum of Consecutive Sample Points	End of Sample/s Index Numbers
Technology	6	372; 1776; 1747
Clothing	4	1013
Household	3	45; 198; 545; 588; 647; 843
Luxury	4	63
Food	6	441
Gifts	5	254; 307; 603; 1651
Sweets	4	94; 189; 971

Table 3, Consecutive points and index

A reason that can be attributed to the low number of maximum consecutive sample points with the mean data is that of the dependant small range between -0.3*sigma and 0.4*sigma that if increased would result in larger numbers.

Technology

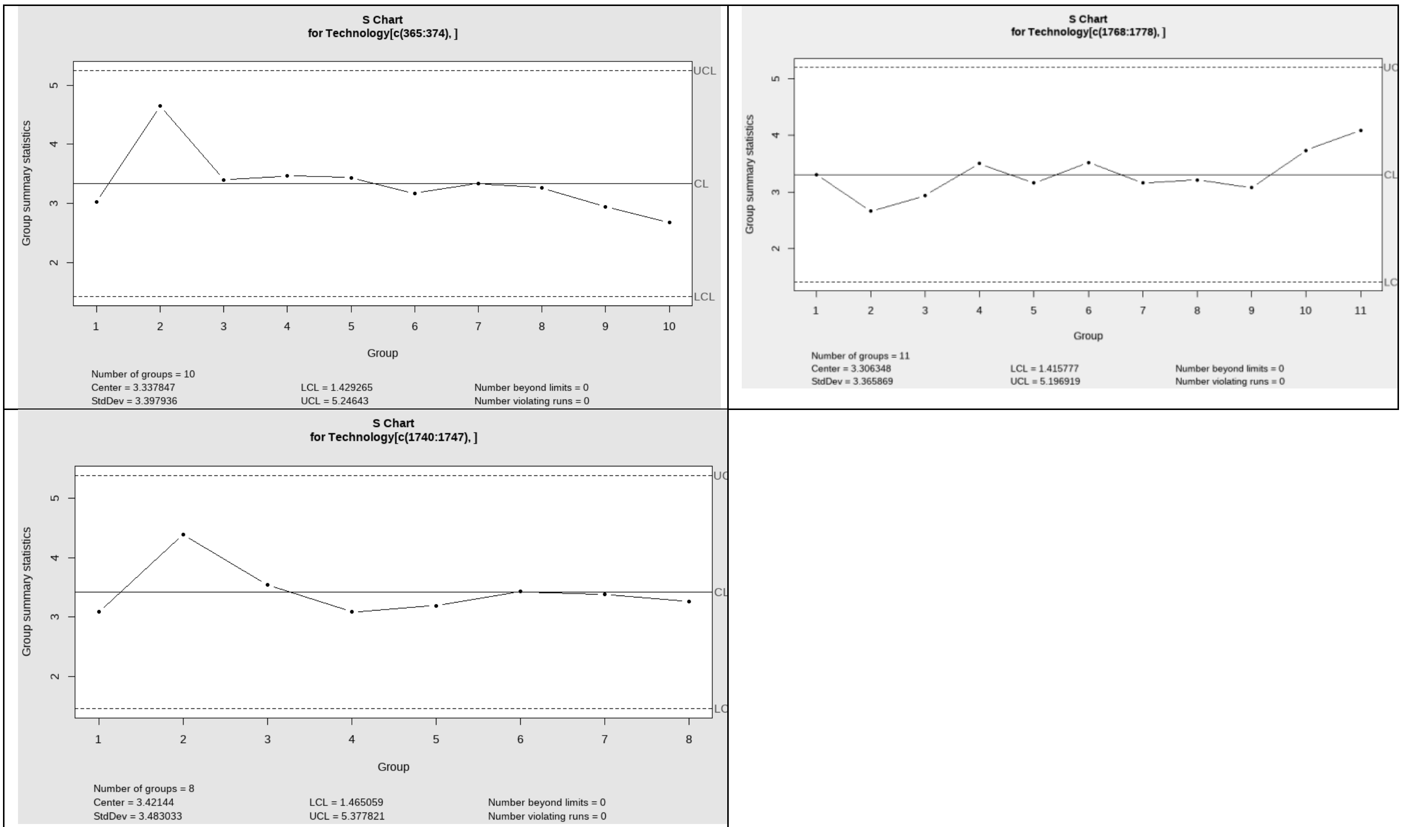


Figure 38, Technology consecutive sample points

Technology with 3 examples of 6 consecutive points between the stated limits.

Clothing

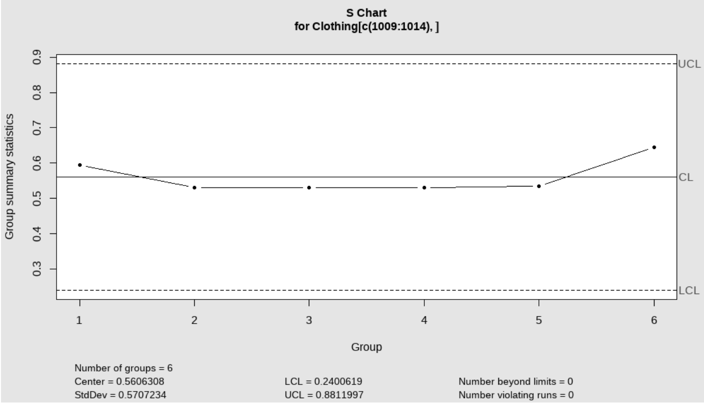


Figure 39, Clothing consecutive sample points

Clothing with 1 example of 4 consecutive points between the stated limits.

Household

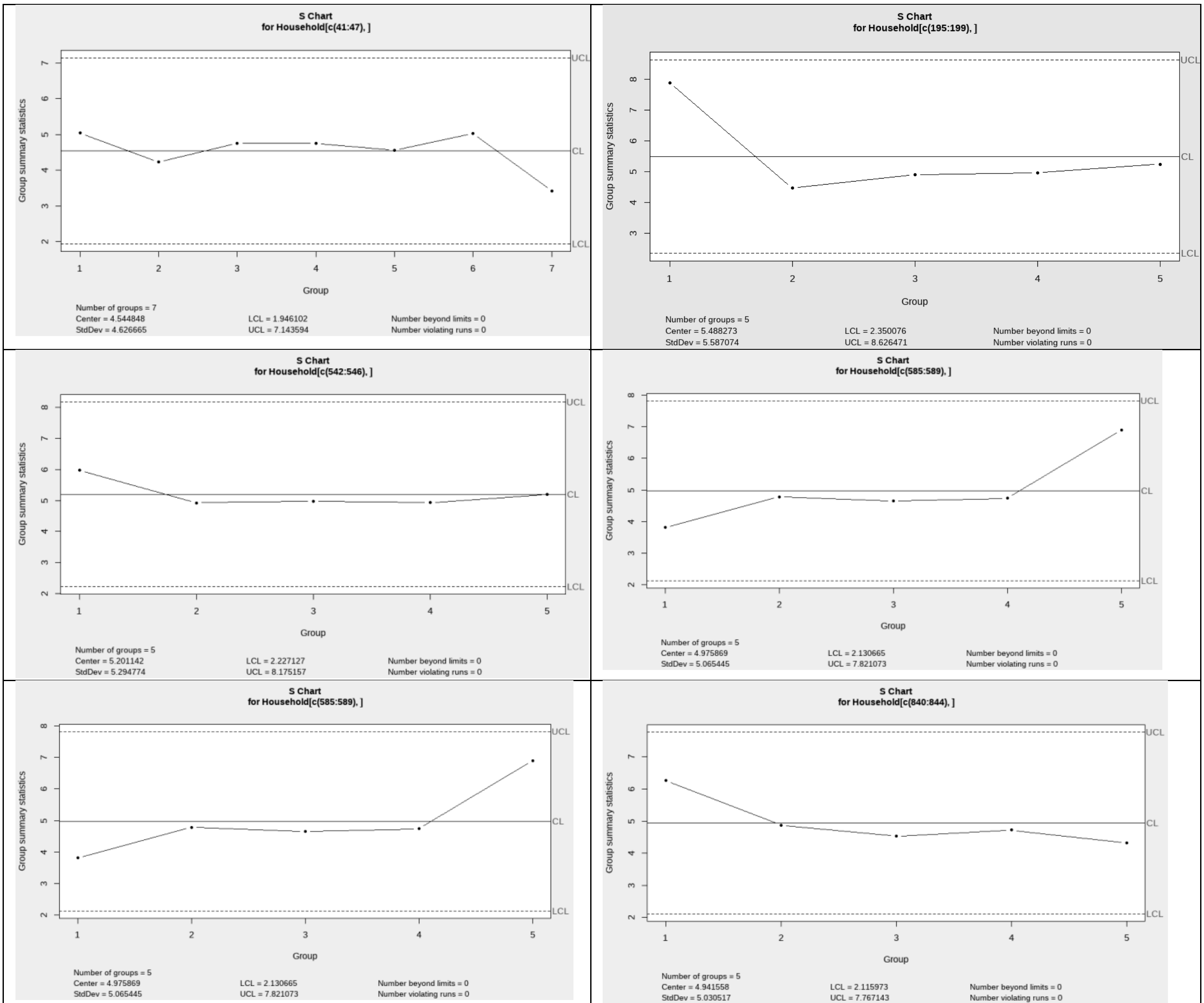


Figure 40, Household consecutive sample points

Household with 6 examples of 3 consecutive points between the stated limits.

Luxury

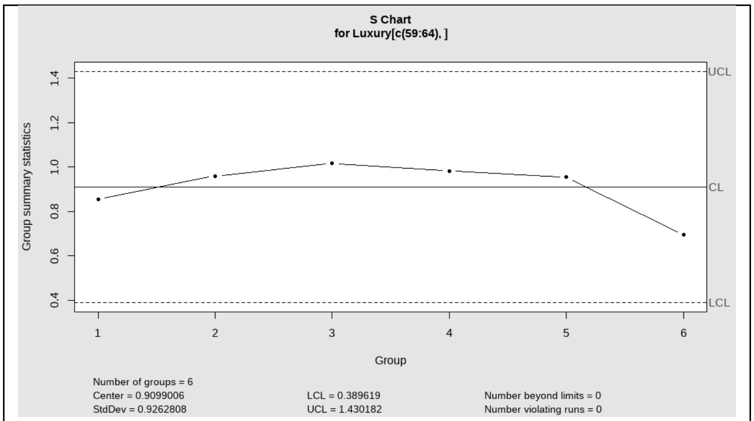


Figure 41, Luxury consecutive sample points

Luxury with 1 example of 4 consecutive points between the stated limits.

Food

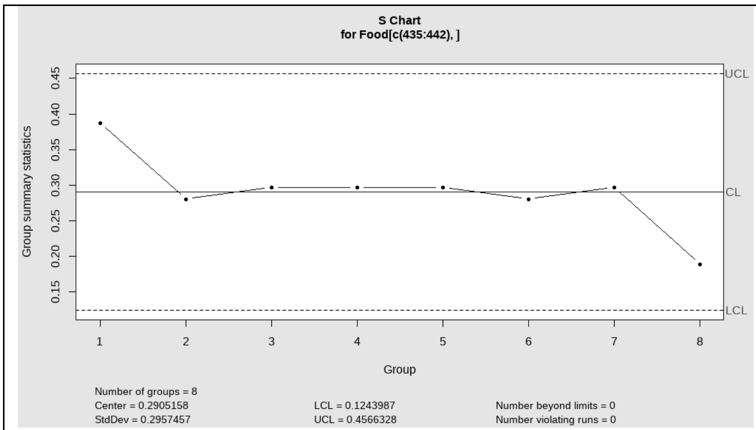
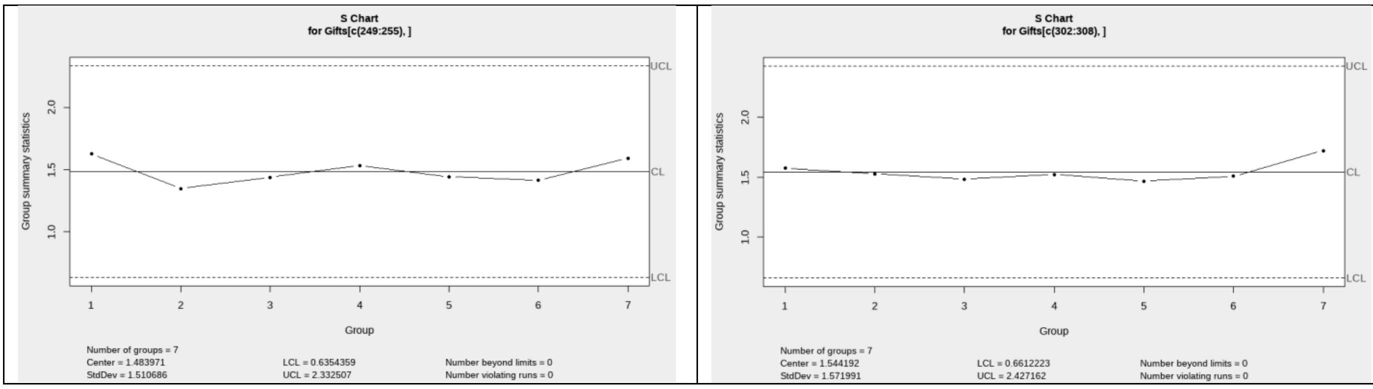


Figure 42, Food consecutive sample points

Food with 1 example of 6 consecutive points between the stated limits

Gifts



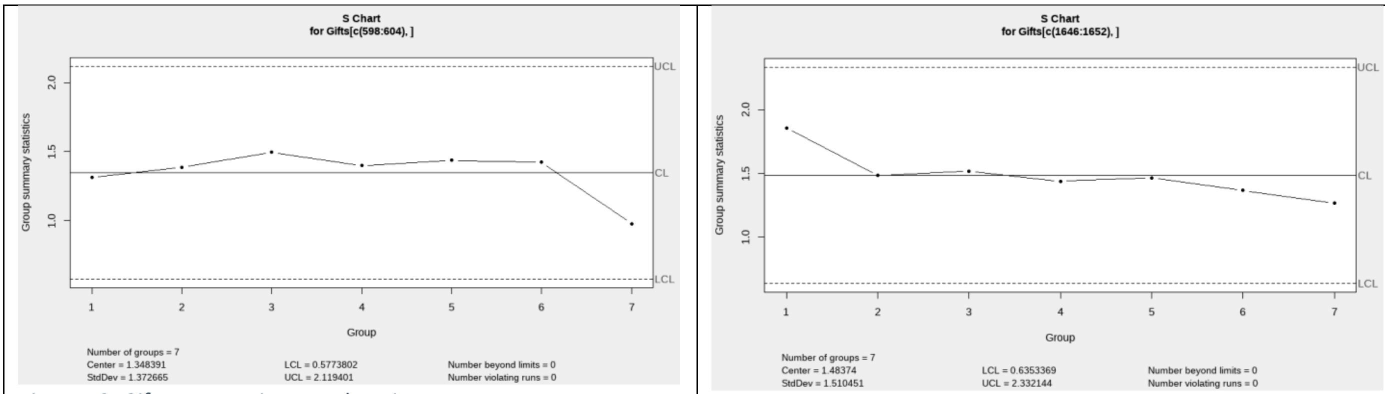


Figure 43, Gifts consecutive sample points

Gifts with 4 examples of 5 consecutive points between the stated limits

Sweets

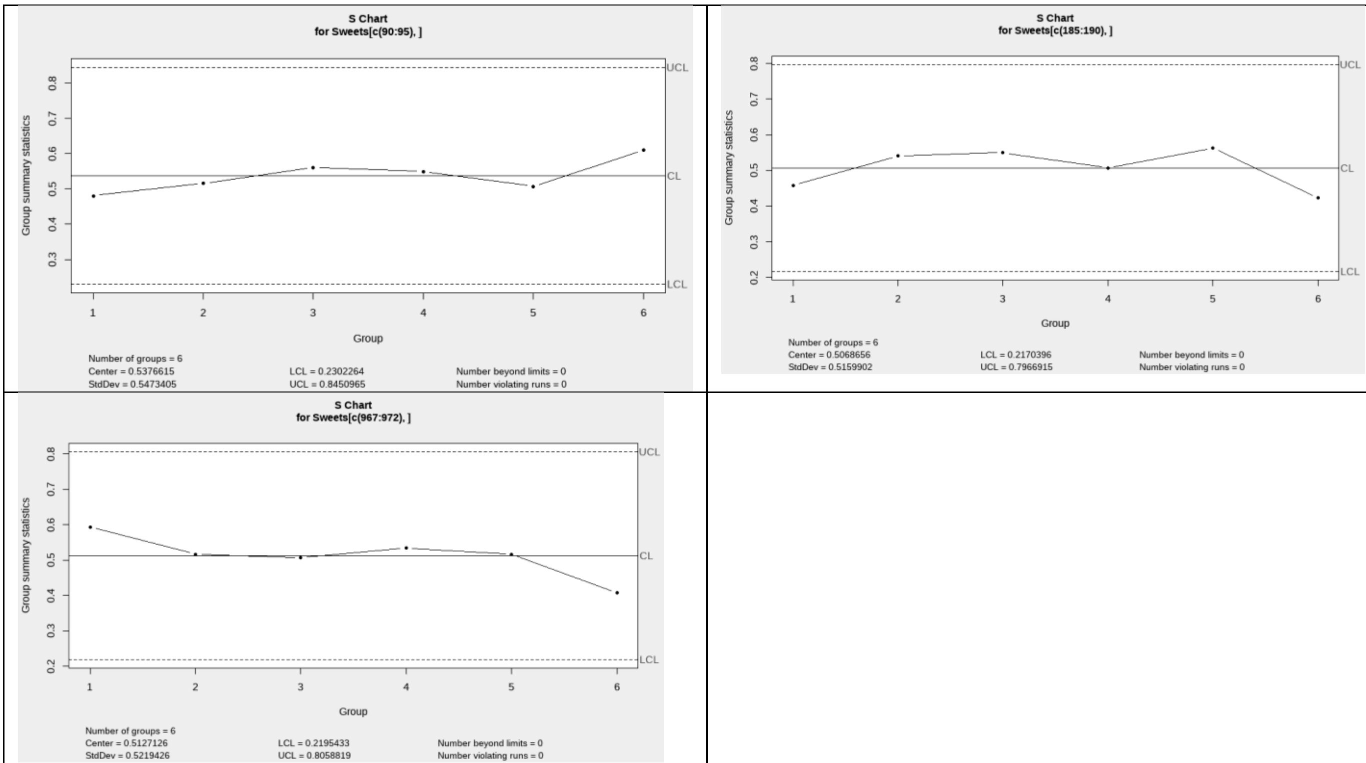


Figure 44, Sweets consecutive sample points

Sweets with 3 examples of 4 consecutive points between the stated limits

4.2

Estimate the likelihood of making a Type I (Manufacturer's) Error for 4.1.A and 4.2.B:

- H_o : the process is in control and centred on the centreline calculated using the first 30 samples
- H_1 : the process is not in control and off the centreline or has increased or decreased in variation

4.1.A

The probability of making a Type I error is 0.27% (0.002699796) and it represents the probability that the data is satisfactory for the control limits but is classified as being beyond the control limits.

4.1.B

The probability of making a Type I error is 73% (0.7266668) and it represents the probability that the standard deviation data is satisfactory for the control limits (0.4*sigma and -0.3*sigma) but is classified as being beyond the control limits.

4.3

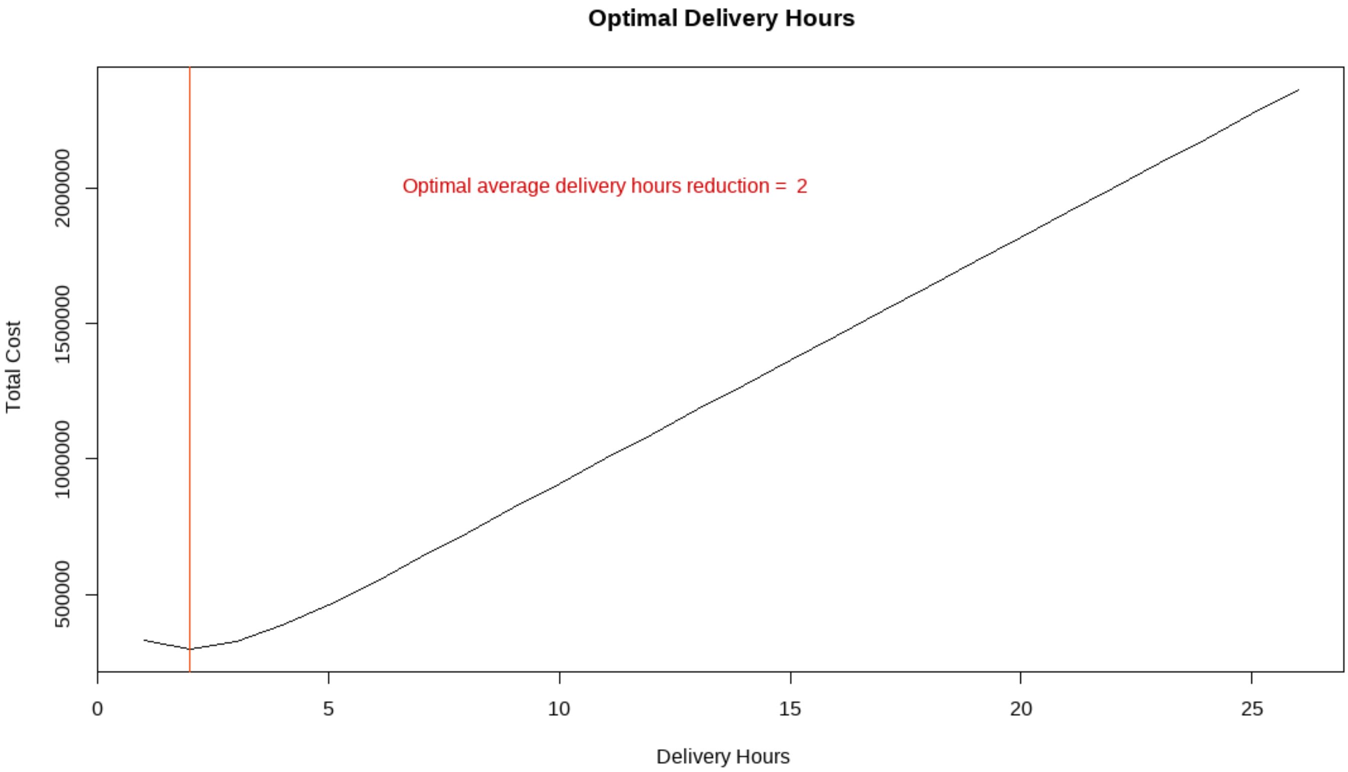


Figure 45, Optimal delivery hours

The delivery hours data has a mean of 20.01 hours with 1356 lost sales due to being over 26 hours. There is a cost of R329/item-late-hour as well as a cost to reduce the delivery time by one hour, which is R2.5/item/hour. Taking these two costs into account the best cost saving solution is found. The calculations returned that the best solution will be to decrease the average delivery time by 2 hours which would result in it being 18 hours. The Taguchi loss function and the loss function are not the same as the graph is not symmetrical.

4.4

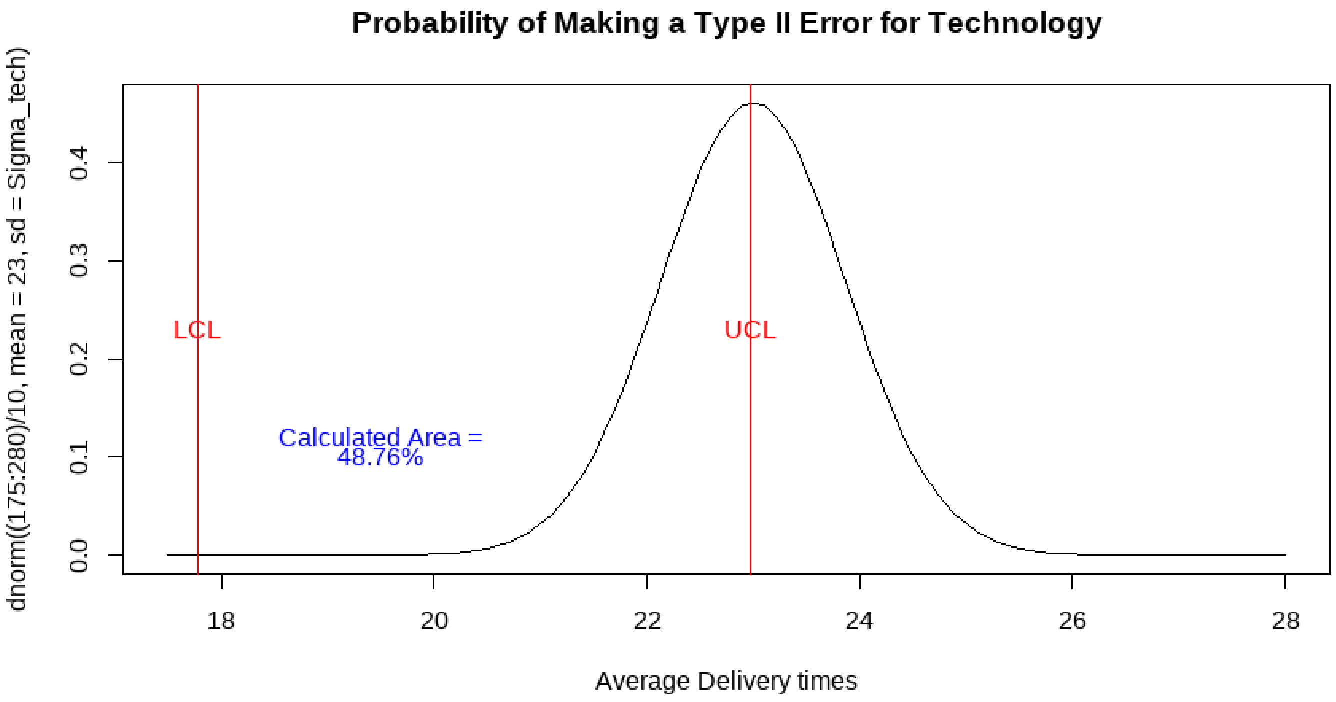


Figure 46, Type II error

Making a type II error is accepting the null hypothesis when the null hypothesis is false. In this case it would be to say that the product is delivered on-time but the product was in actual fact delivered late. As indicated above, the probability of a type II error using a mean of 23 hours is equal to 48.76% (0.4876147). This is a high probability and the business needs to make sure that this is decreased by making sure that deliveries that are late are classified as late and not on-time. A reason for this being close to 50% is because the mean was moved from +-20 to 23 and the 3*sigma of the actual technology data is close to 3*sigma.

5

For the MANOVA test, a p-value of 0.05 is used as this is the most common p-value used through p-value tests.

5.1

H_0 = For the categories price, delivery times and age, they make no significant change to the buying pattern in terms of product class.

H_1 = At least one class has an influence on the buying pattern.

Dependent variables: price, delivery time an age.

Independent variable: class of each product.

- MANOVA test: does a class have any influence
  A p-value of 2.2e-16 is returned which is much smaller than 0.05 and therefore one needs to reject the null hypothesis as at least one class is significantly different because at least one dependent variable differs in its average.

Analysis of each dependent class

- Price: a value of 2.2e-16 is returned as is smaller than 0.05 which results in the conclusion that the price differs depending on the class of the product.

- Delivery time: a value of 2.2e-16 is returned as is smaller than 0.05 which results in the conclusion that the delivery time differs depending on the class of the product.
- Age: a value of 2.2e-16 is returned as is smaller than 0.05 which results in the conclusion that the age of the person buying the product differs depending on the class of the product.

Graphs supporting results:

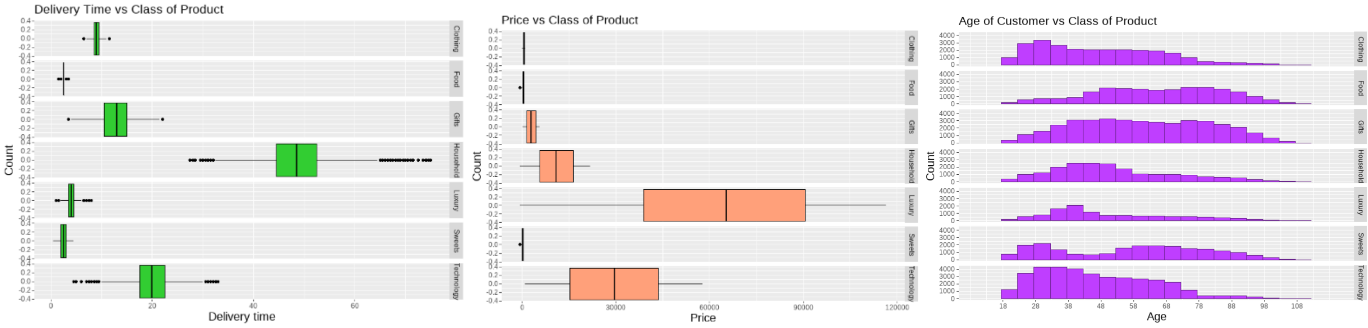


Figure 47, MANOVA 1 supporting graphs

Follow link to supporting part 2 graphs.

Conclusion: The delivery time for the different classes of products vary greatly between them. The price of each class of product varies greatly in areas such as luxury, technology and household. The age of customers who buy the products vary amongst the classes as different classes of products appeal to different age groups.

5.2

H_0 = For the categories price, delivery times and age, they make no significant change to the buying pattern in terms of why bought.

H_1 = At least one class has an influence on the buying pattern in terms of why bought.

Dependent variables: price, delivery time an age.

Independent variable: why bought

- MANOVA test: does a class have any influence
  A p-value of 2.2e-16 is returned which is much smaller than 0.05 and therefore one needs to reject the null hypothesis as at least one class is significantly different because at least one dependent variable differs in its average.

Analysis of each dependent class

- Price: a value of 2.2e-16 is returned as is smaller than 0.05 which results in the conclusion that the price differs depending on the class of the product.
- Delivery time: a value of 2.2e-16 is returned as is smaller than 0.05 which results in the conclusion that the delivery time differs depending on the class of the product.
- Age: a value of 2.2e-16 is returned as is smaller than 0.05 which results in the conclusion that the age of the person buying the product differs depending on the class of the product.

Graphs supporting results:

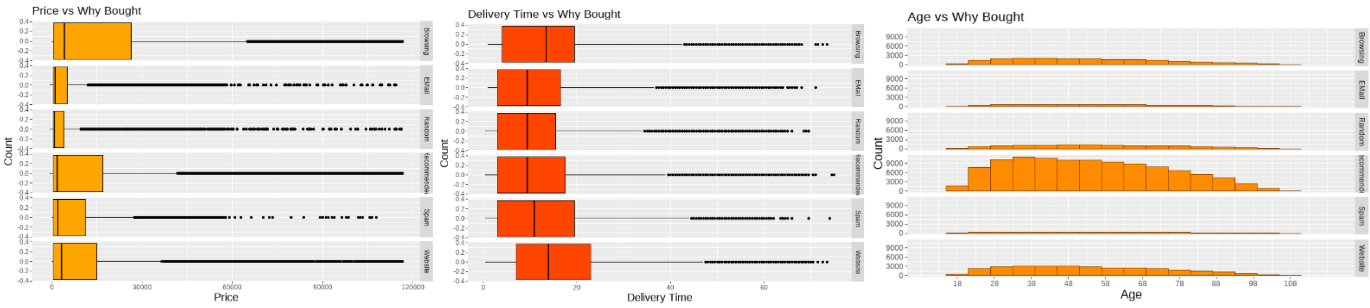


Figure 48, MANOVA 2 supporting graphs

Follow link to supporting part 2 graphs.

Conclusion:

The age has a clear difference in distribution in the different reasons for being bought. The delivery times have varying means and distribution of points for each of the reasons for buying. The summary data of the price of each reason for being bought varies greatly.

6

6.1

Problem 6: A blueprint specification for the thickness of a refrigerator part at Cool Food, Inc. is 0.06 +- 0.04 centimeters (cm). It costs $45 to scrap a part that is outside the specifications. Determine the Taguchi loss function for this situation.

Constant:

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

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

45 = k(0.04)^2

k = 45/(0.04)^2

= 28 125

Loss Function:

$L(x)$ = 28 125(x – 0.06)^2

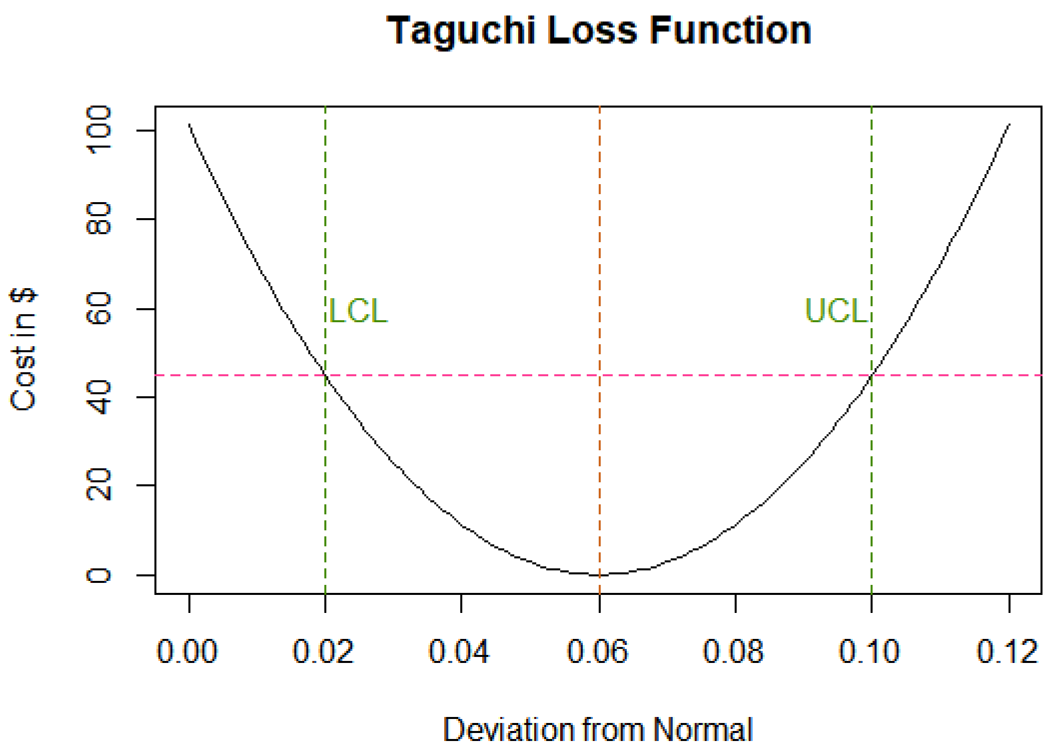


Figure 49, Taguchi loss function 1

For the best result, the business will want to have a deviation of zero from the norm measurement of 0.06 centimetres, as this will have the lowest cost, shown above, as the quality of the product is the best. As the part measurement deviates from the norm the cost increases and results in a smaller profit as the part is also now less desirable. As the question mentions there is a $45 cost to scrap a part and this results in $45 being spent no matter how far out the tolerances the part is. Unfortunately, in reality it is very hard to achieve no tolerances and the smaller the tolerances the more expensive the product will become and will outweigh the cost of the small deviations. This will therefore lead to a test to find what the best tolerances when taking into the cost of higher precision.

Problem 7: A team was formed to study the refrigerator part at Cool Food, Inc. described in Problem 6. While continuing to work to find the root cause of scrap, they found a way to reduce the scrap cost to $35 per part.

A. Determine the Taguchi loss function for this situation.

Constant Calculation:
$L(x) = k(x - T)$^2
35 = k(0.04)^2
k = 35/(0.04)^2
= 21 875
Loss Function Calculation:
$L(x) = k(x - T)$^2
$L(x)$ = 21 875(x – 0.04)^2

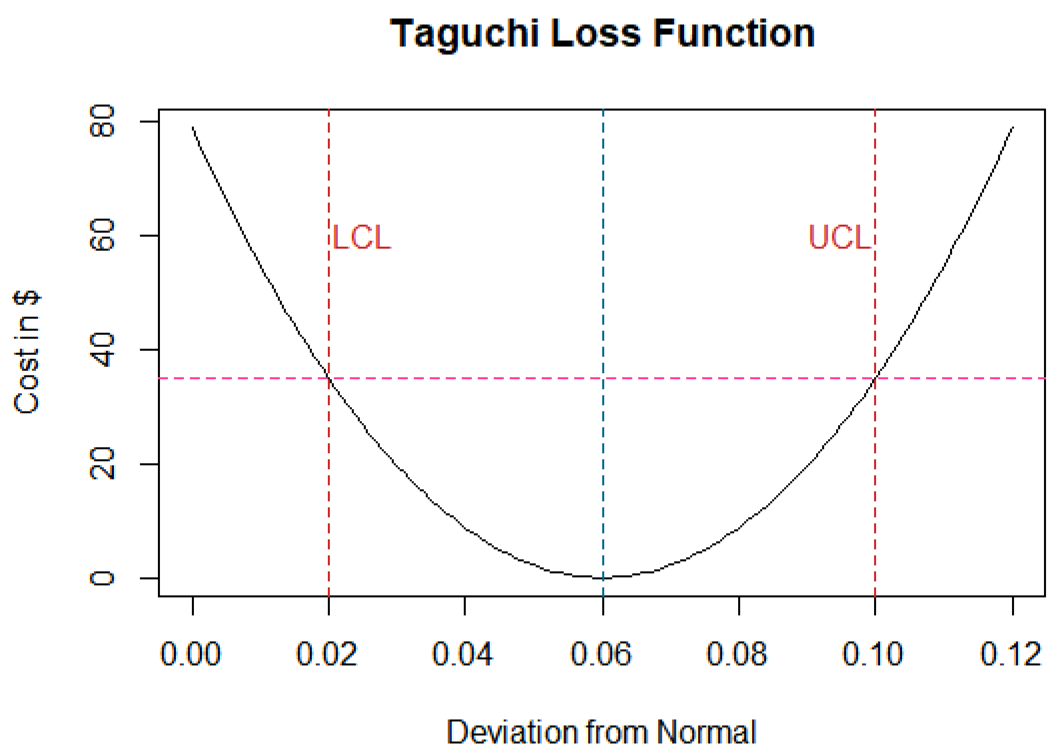


Figure 50, Taguchi loss function 2

The business still has the same tolerance to work in but has a smaller cost to scrap the part and the cost increases more gradually with increasing tolerances from the norm of 0.06 centimetres when compared to the scrap value of $45. This means that they will make more profit from identical parts that are from the two experiments with the scrap value of $35 compared to $45 because of the more gradual cost increase. This still does not move from the fact that lower quality parts will cost the business more as shown above.

B. If the process deviation from target can be reduced to 0.027 cm, what is the Taguchi loss?

$L(0.027)$ = 21 875(0.027)^2

$L(0.027)$ = $15.946875

This reduction in the process deviation from 0.04 to 0.027 centimetres will result in the business making a loss of $15.946875/part for parts that are outside the tolerance of 0.027 centimetres compared to $35/part with a tolerance of 0.04 centimetres. Even though the parts will be more similar to one another because of the smaller tolerances, this may increase obsoletes if the process/machinery is not improved and it may result in decreased levels of service.

6.2

Problem 27: Magnaplex, Inc. has a complex manufacturing process, with three operations that are performed in series. Because of the nature of the process, machines frequently fall out of adjustment and must be repaired. To keep the system going, two identical machines are used at each stage; thus, if one fails, the other can be used while the first is repaired (see accompanying figure).

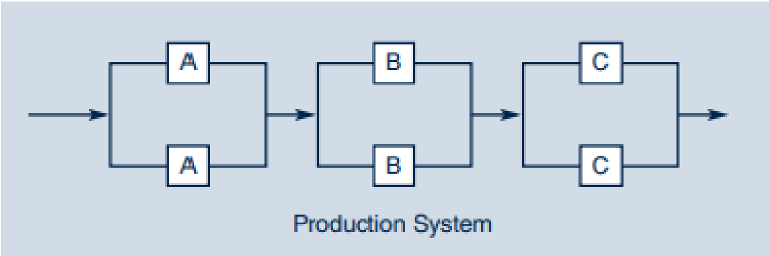


Figure 51, Machine layout

The reliabilities of the machines are as follows:

Machine	Reliability
A	0.85
B	0.92
C	0.90

Table 4, Machine data

A. Analyze the system reliability, assuming only one machine at each stage (all the backup machines are out of operation).
   System Reliability = Reliability(Machine A)* Reliability(Machine B)* Reliability(Machine C)
   System Reliability = 0.85*0.92*0.90
   System Reliability = 0.7038
B. How much is the reliability improved by having two machines at each stage
   System Reliability = Reliability(2xMachine A)* Reliability(2xMachine B)* Reliability(2xMachine C)
   System Reliability = (1-(1-0.85)^2)* (1-(1-0.92)^2)*(1-(1-0.90)^2)
   System Reliability = 0.9615
   By having two identical machines in parallel at each of the three stations respectively, it increases the system's reliability by 25.77% (96.15-70.38). This is the advantage of placing identical machines in parallel because unlike just having machines in series (where like electrical circuits, if one component breaks or slows it directly impacts the machines around it) there is still another path the part can take if one machine breaks as there is an operational machine that it can still go through. This is true if both machines are running else the reliability decreases and is treated like a series machine.

6.3

Reliable vehicles probability:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad x = 0, 1, 2, \ldots, n$$

Figure 52, Binomial formula

1. P(0) = [21C0 * p^0 * (1 − p)^25−0] * 1560 = 1560 − (190 + 22 + 3 + 1) = 1344
   p = 0.007071366
2. P(1) = [21C1 * p^1 * (1 − p)^25−1] * 1560 = 190
   p = 0.006596143
3. P(2) = [21C2 * p^2 * (1 − p)^25−2] * 1560 = 22
   p = 0.008908085
4. P(3) = [21C3 * p^3 * (1 − p)^25−3] * 1560 = 3
   p = 0.01216631
5. P(4) = [21C4 * p^4 * (1 − p)^25−4] * 1560 = 1
   p = 0.01970866

Weighted average = (0.007071366*1344+0.006596143*190+0.008908085*22+0.0267331*3+0.01970866*1)/1560 = 0.007057287

Expected reliable delivery days in a year:

P(x < 1) = [21C1 * 0.007071366 * (1 − 0.007071366) 25−1] = **0.9904327**

P(x < 1) = [21C1 * 0.007071366 * (1 − 0.007071366) 25−1] * 365 = **361.5079 days**

1. P(0) = [21C0 * p^0 * (1 − p)^25-0] * 1560 = 1560 - (95 + 6 + 1) = 1458
   p = 0.003242961
2. P(1) = [21C1 * p^1 * (1 − p)^25-1] * 1560 = 95
   p = 0.003084143
3. P(2) = [21C2 * p^2 * (1 − p)^25-2] * 1560 = 6
   p = 0.004468035
4. P(3) = [21C3 * p^3 * (1 − p)^25-3] * 1560 = 1
   p = 0.008241969

Weighted average = (1458*0.003242961 + 95*0.003084143 + 6*0.004468035 + 1*0.008241969)/1560 = 0.003241206

Expected reliable delivery days in a year:

P(x < 1) = [21C1 * 0.003241206* (1 − 0.007071366) 25−1] = **0.9978825**

P(x < 1) = [21C1 * 0.003241206* (1 − 0.007071366) 25−1] * 365 = **361.5079 days**

Total reliable days (combining reliability of vehicles and drivers):

P(reliable) = P(vehicles) * P(drivers) = 0.9904327*0.9978825 = 0.9883354

Expected reliable days = 0.9883354*365 = 360.7424days

This means out of 365 days of the year, 361 days will have reliable delivery

Now increasing vehicles from 21 to 22 and using the same weighted average:

1. P(22) = [22C0 * p^0 * (1 − p)^25-0] = 0.8557211

2. P(21) = [22C1 * p^1 * (1 − p)^25-1] = 0.1338038
3. P(20) = [22C2 * p^2 * (1 − p)^25-2] = 0.009985537
4. P(19) = [22C3 * p^3 * (1 − p)^25-3] = 0.0004731445
5. P(18) = [22C4 * p^4 * (1 − p)^25-4] = 1.597353e-05

Increased delivery reliability:

Total of all reliability = 0.8557211 + 0.1338038 + 0.009985537 + 0.0004731445 + 1.597353e-05 = **0.9999996**

Increased expected reliable days = 0.9999996*365 = 364.9998 = **365 days**

This increase of delivery vehicles from 21 to 22 has led to an increase in reliable days of 4 to total 365 days which means the company should be able to deliver reliably every day of the year.

Conclusion

Data is obtained from the online store which is then cleaned and sorted into valid and invalid data as the valid data is used throughout the project. Graphs are constructed to gain understanding of the distribution of information that is in the data so that it can be analysed statistically.

Graphs indicating sample points of the mean and standard deviation are used as control charts to find whether a class of products was in control or out of control. From this the three classes of household, luxury and gifts were found to be out of control. Further investigation was done to find the extent of the out of control and this will require further investigation into areas that were identified within these classes.

Type I and II were calculated and type II significantly outweighed type I which will need some investigation into what is causing this besides the fact that the mean was shifted to 23.

Impacts of different factors in the data was explored to find out what areas have a correlation to others. This will lead the business making better decisions as they know which areas are affected by one another.

Further investigation into the delivery time area shows the impact of adding another vehicle into the vehicle pool and the delivery reliability it has.

The report proves the importance of statistical analysis and the insightful advantage it can possibly give a business if done right. Furthering the business and the industries around it.

References

Scrucca, L. (2004) ‘qcc: An {R} package for quality control charting and statistical process control’, *R News*, 4(1).

Smith, K.N., Lamb, K.N. and Henson, R.K. (2020) ‘Making Meaning out of MANOVA: The Need for Multivariate Post Hoc Testing in Gifted Education Research’, *Gifted Child Quarterly*, 64(1). doi:10.1177/0016986219890352.