Project

Mia du Plessis

Quality Assurance 344

[23580054]

October 2020

Table of content

Introduction

This report contains exploratory data (visual and descriptive) for online retailers. Data is a very valuable resource for a business when used correctly. Information can be obtained from data that can help the company potentially increase revenue.

The main parts of the report are divided into six parts, starting with data wrangling, followed by descriptive statistics, then statistical process control, delivery process optimization, MONAVA testing, and finally service and product reliability.

Certain sales list data should be sanitized by removing invalid data that is considered invalid before performing data analysis techniques. Trends and relationships between purchase dates, classes, customers, and product prices are identified. Descriptive statistics are used to get familiar with the dataset. The focus shifts to product delivery and an X&S chart are created for each class. Calculations of optimal delivery times to reduce losses due to delivery delays and increase service and product reliability are made.

The report ends with a summary of the analysis performed, followed by references.

PART 1: Data Wrangling

All records in the database containing NA values are deleted and thus not used in further calculations. These invalid records may be the result of accidents or administrative problems and should be investigated. We have a lot of sales results in 2022 compared to other years, this could be due to more sales, more data collection, or lost records. This should also be investigated.

Valid data:

The invalid instances with negative values in the price column is removed. The total number of negative values is 5 instances.

The valid data (instances that do not contain any missing data) is separated from the invalid data. The total number of observations is 179983 instances out of 180000 instances.

Thus the new primary key should run from 1 to 179 978.

First 19 instances:

	t	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

Figure 1 Valid data set

Index difference:

The old and new index differs from instance 12345, because instance 12345 is the first instance that contains missing data. This is shown in figure 1 below.

Old row:

12344	90260	34	Luxury	42891.66	2025	8	4	4	Recommended
12345	18973	93	Gifts	NA	2026	6	11	15.5	Website
12346	92286	32	Technology	38167.24	2028	7	6	19.5	Website

New row:

12343	27986	37	Clothing	712.19	2021	10	10	9	Recommended
12344	90260	34	Luxury	42891.66	2025	8	4	4	Recommended
Removed									
12345	92286	32	Technology	38167.24	2028	7	6	19.5	Website
12346	89263	44	Clothing	891.71	2021	7	2	8.5	Recommended

Instance 12 345 took the place of the old instance 12 346.

Invalid data:

The invalid data (instances that contain any NA values) is separated from the valid data. The total number of observations is 17 instances out of 180000 instances.

Invalid Data table:

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

Figure 2 Invalid instances

PART 2: Descriptive Statistics

Data quality report: Continuous features of Valid data:

The date can be used as continuous and categorical and will be used in both the continuous and categorical data quality reports.

```
       ID              AGE              Price                 Year          Month               Day           Delivery.time
 Min.   :11126   Min.   : 18.00   Min.   :     35.65   Min.   :2021   Min.   : 1.000    Min.   : 1.00   Min.   : 0.5
 1st Qu.:32700   1st Qu.: 38.00   1st Qu.:    482.31   1st Qu.:2022   1st Qu.: 4.000    1st Qu.: 8.00   1st Qu.: 3.0
 Median :55081   Median : 53.00   Median :   2259.63   Median :2025   Median : 7.000    Median :16.00   Median :10.0
 Mean   :55235   Mean   : 54.57   Mean   : 12294.10   Mean   :2025   Mean   : 6.521    Mean   :15.54   Mean   :14.5
 3rd Qu.:77637   3rd Qu.: 70.00   3rd Qu.: 15270.97   3rd Qu.:2027   3rd Qu.:10.000    3rd Qu.:23.00   3rd Qu.:18.5
 Max.   :99992   Max.   :108.00   Max.   :116618.97   Max.   :2029   Max.   :12.000    Max.   :30.00   Max.   :75.0
```

Figure 3 Continuous features descriptive statistics

The primary key feature has a cardinality equal to the number of instances. Thus, the primary key is an irrelevant feature. The youngest age is 18 years old and the oldest age is 108 years. The most popular age is 55 years old. The maximum for age could be interpreted as an error or a mistake, because 108 for age is a lot higher than the average living years. The average price is R12294.10. The fastest delivery time is 1 day and the longest delivery time is 30 days. The average time to deliver to customers is 16 days. Price also indicates a maximum of R116618.97, this shows that there are no Prices above R1000000.

Data quality report for Categorical features of the Valid dataset:

Class name	Count	Miss_val	Card	Modes	Mode_freq	Mode_perc
ID	179978	0	15000	41842	27	0.0150018
Class	179978	0	7	Gifts	39149	21.752103
Year	179978	0	9	2021	33443	18.581716
Month	179978	0	12	12	15225	8.4593673
Day	179978	0	30	17	6126	3.4037493
WhyBought	179978	0	6	Recommended	106985	59.443376

. ID, Year, Month and Day are categorical features. The month feature's variables are used to indicate the months from January (month 1) to December (month 12).

ID's cardinality is not the same as the number of instances. This could be a mistake because if every ID differed, the cardinality would've been equal to the number of instances. This implies that almost 30000 instances have the same ID.

Gifts is the class with the most popular sale. The date on which there was the most frequent purchases was in the year 2021, 17th of December. A reason for this could be that it is close to Christmas, indicating that people are buying gifts that time of the year as Christmas gifts.

The reason for most purchases on the online store was because of a recommendation by someone. This information is valuable for the company as this is an indication that good quality service and customer satisfaction will lead to more sales as people tend to recommend products if they are satisfied with the quality of the products.

Graphs

Delivery time

Figure 4 Delivery times for sales

Delivery Time is a bit normally distributed between 35 days and 70 days, with a mean of 47 days. The most frequent delivery times are around 2 to 3 days. Frequent deliveries at 8 days can also be observed. There is a negative trend in the count for delivery days between 15 and 30 days.

Class

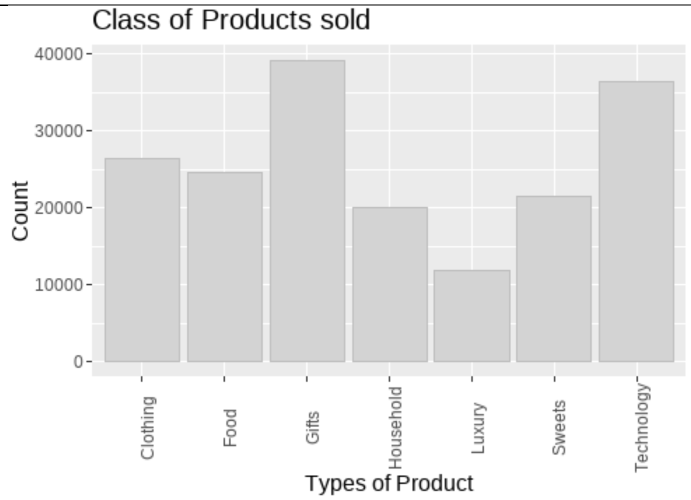


Figure 5 count of classes of products sold

The most frequent product bought is gifts and the second most is technology. The least frequent product bought is luxury products. Gifts and technology must be ensured to be of great quality to customers since these products are the most frequently bought products.

Reasons for bought

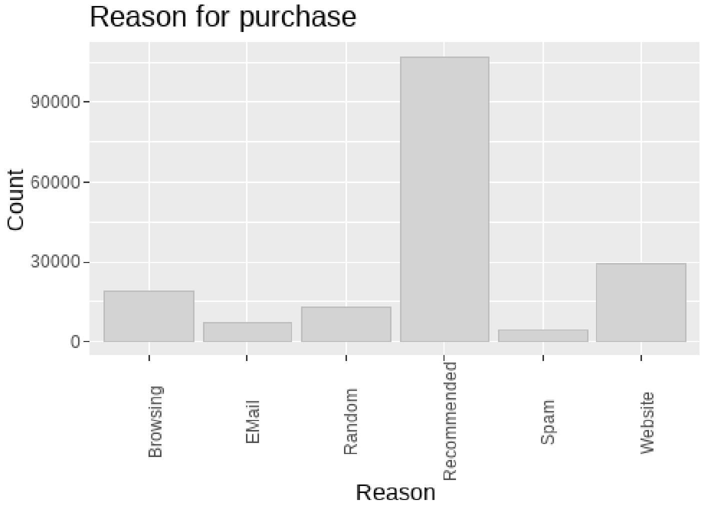


Figure 6 Count of reasons for purchase

The main reason for purchase is because of recommendation. The company can thus find this useful to rather focus on customer satisfaction rather than focusing on spam and emails send to customers to obtain more sales as the least frequent reasons for purchase are because of spam and emails. Service must be excellent to customers to improve the dissemination of information to the public as product advertising.

Age

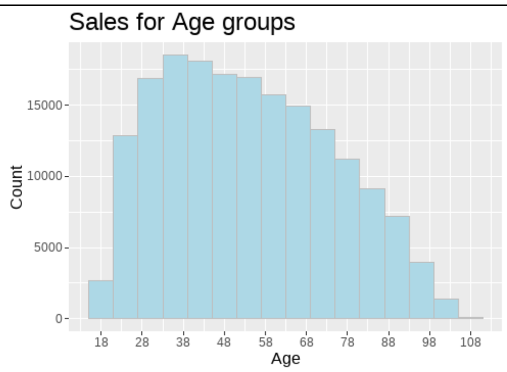


Figure 7 Sales per age group

A unimodal, right-tailed distribution is observed. There is a tendency toward very low values. The age group that has the highest purchasing count is the group of ages between 33 and 38. This could be because people in this age group often have families and a stable income which leads to a higher count in purchasing, thus high sale counts for the company. The skew tendency might be because as people get older, they buy less as they still use items that they purchased at a young age. The sales management team must advertise among younger age groups through online material as these age groups tend to use online platforms more than older age groups.

Year

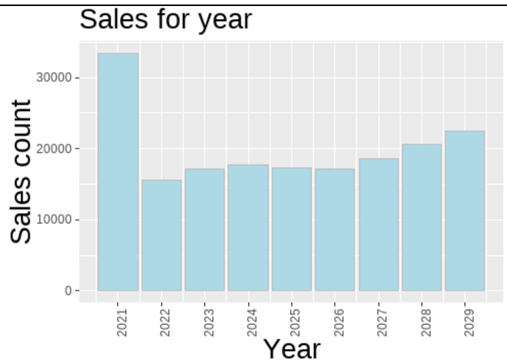


Figure 8 Sales per year

The figure above indicates that year 2021 had the most sales. By excluding the year 2021, a positive trend can be identified from year 2022 to year 2029.

Month

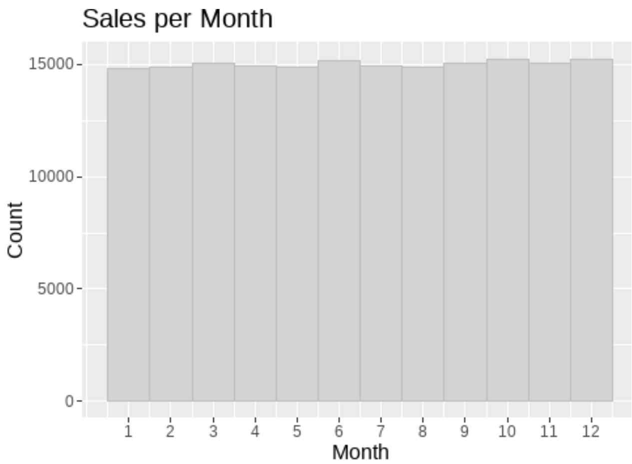


Figure 9 Sales per month

The graph of sales per month shows a uniform distribution, thus no trends can be identified. This could be useful in forecasting sales for the coming months.

Day

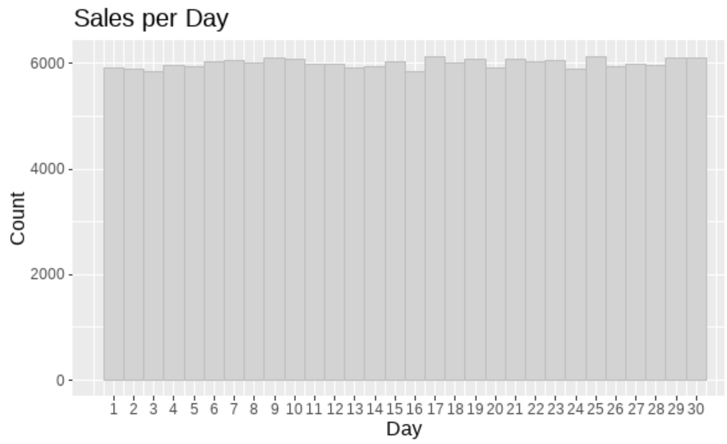


Figure 10 Sales per day

A uniform distribution can be seen, thus the sale count for each day does not seem to follow a trend. Could be useful in forecasting sales for the following days. The least sales seem to be at the beginning of each month.

SPLOM for Age, Price and Delivery time

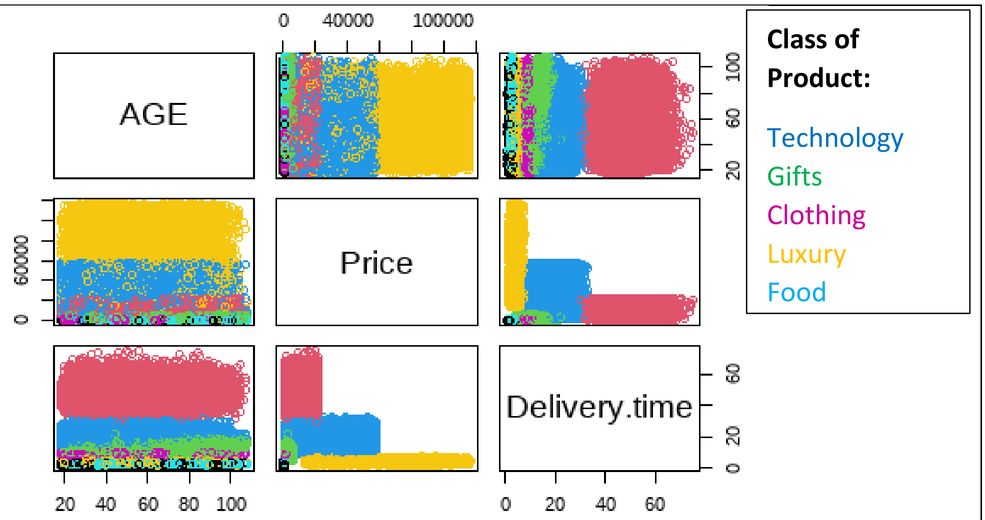


Figure 11 SPLOM

According to the SPLOM, there is a good separation between Price and delivery. The separation shows that luxury items (which are the most expensive) have the shortest delivery times. Household items, averagely priced, take the longest to deliver. A trend can be identified, the more expensive the item, the faster the delivery time. This can be seen when comparing technology, luxury items and gifts.

A relationship between Age vs Price and Age vs Delivery time cannot be identified when looking at the SPLOM. This SPLOM makes it easy to simultaneously compare features with each other at once. Making it easy when deciding into which comparisons a deeper look should be taken.

Delivery time vs class of product

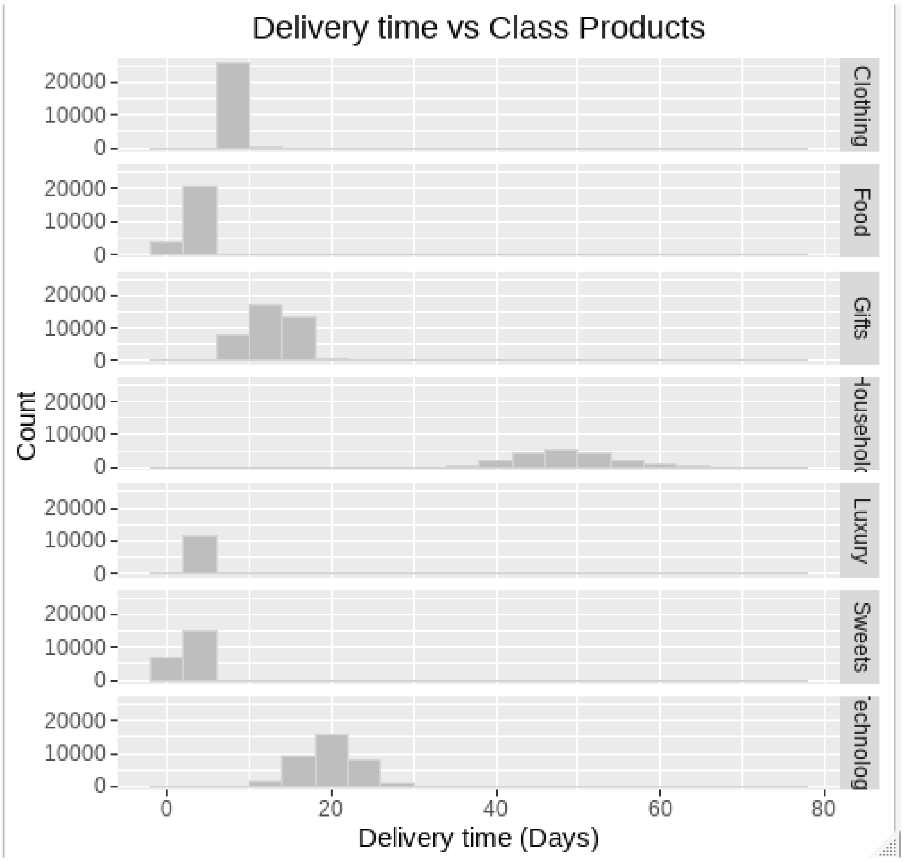


Figure 12 Plot of Delivery time vs Class of products

According to the graphs the items food, luxury, and sweets have the shortest delivery time. The short delivery period of food and sweets could be due to small product size, smaller batches and shorter lifetimes which make it easy and fast to travel. Household items are the items which have the longest delivery times. It could be beneficial for the company if the reasons for long delivery for Household items take longer to try and shorten these times. This could be due to the big sizes of household products which makes transport more complicated and expensive. The distribution of household products seems to follow a normal distribution, this could be due to size and distance variation.

Age vs Reason for sale

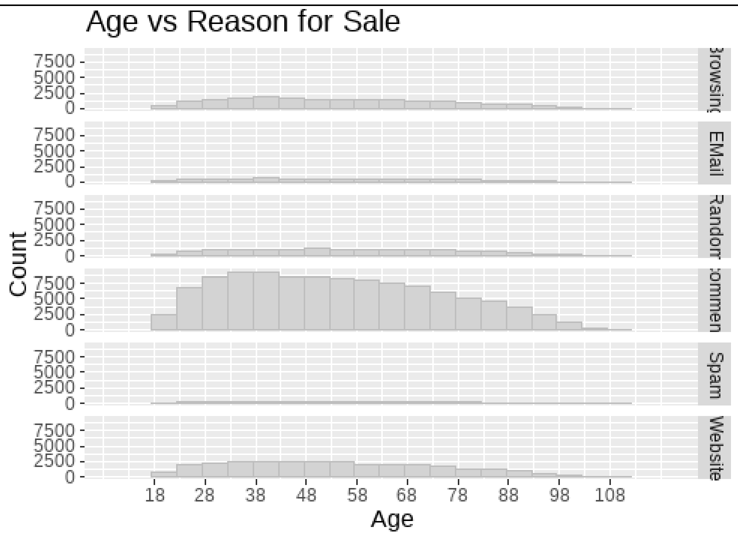


Figure 13 Age vs Reason for Sale

There seem to be a skewed to the right unimodal distribution for the reason of purchase, especially when looking at reasons “recommend”, “browsing” and “website”. The reason for buying seems to be distributed across all ages, tending to lower counts of sales as age increases. Either the company should focus on keeping customers in the age groups between 28 and 58 happy or/and find ways to increase sales for older age groups.

Age vs Class of product

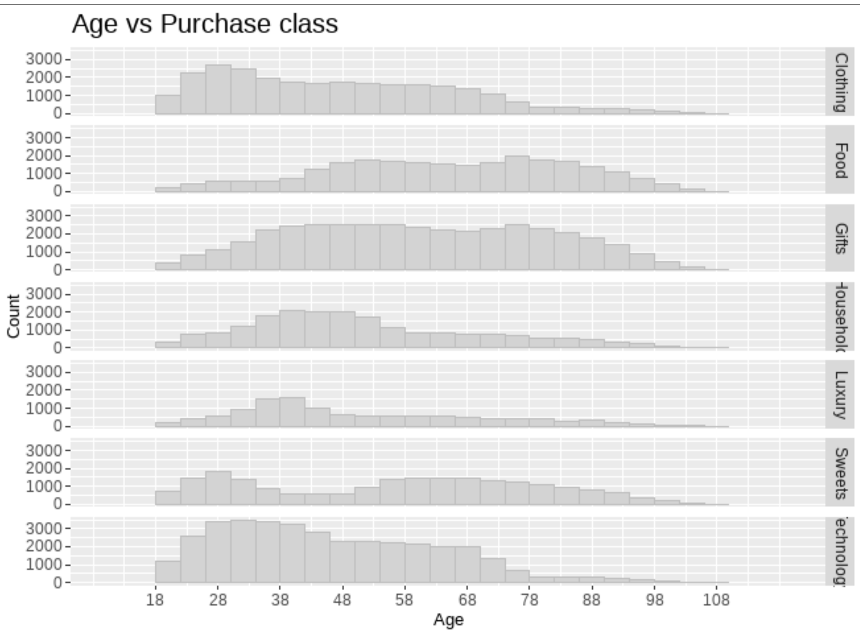


Figure 14 Age vs Purchase class

All the classes go beyond the age of 108 years. Retirement villages might make use of online platforms to purchase products, but a lot less than younger age groups. This might be because younger people are more connected online and have the need to purchase more items. Clothing is distributed exponentially: The age group buying clothing the most frequently is ages 28 to 38. This is due to younger people that are interested in following clothing trends. The Food class has a multimodal distribution: The age group buying food the most frequently is aged 78 to 88. The Gifts class is very distributed across the age demographics; it makes sense since gifts are bought by any age group, it does not increase or decrease depending on your age. Household items are distributed exponentially with a most frequent age range of 38 and 44. It makes sense for a middle-aged person to spend a lot of money on household items as they move to a bigger house as they reach their middle age and as they start to make families. Luxury is also exponentially distributed with a most frequent age range of 38 to 44. This could be due to people in these age groups starting to earn more stable incomes and having saved up for a few years making it possible to spend more money on luxury items.

Technology has a unimodal (right-tailed) distribution: most frequent age between 30 and 34 years old. Technology's mean age group is the youngest among the other classes. It could be beneficial for sales managers to investigate ways to advertise technology in a way that reaches the younger age groups to further increase sales. It can be concluded that younger/middle-aged people tend to buy more clothing, households, luxury and technology conclusion is conducted that the younger age people/middle-aged people tend to buy clothing, households, luxury and technology.

Month vs Class of product

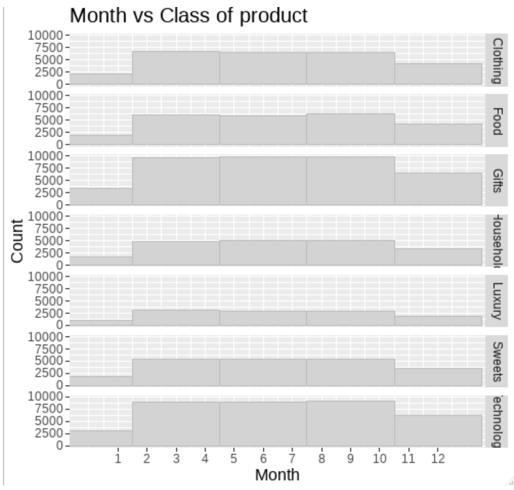


Figure 15 Month vs Class

A clear seasonal trend can be identified between months 2 and 10 for all classes. Thus people tend to spend more on items between February and October, with the classes being irrelevant. It could be beneficial for the company to identify reasons for this seasonal trend to decide when to promote and advertise products to increase revenues.

Price vs Class of Product

Figure 16 Price vs Class

It can easily be assumed, when looking at the boxplots, that the price for luxury items is the highest. Technology is the second most expensive class. These items cost more to manufacture and have a higher value than the other classes such as sweets, food and clothing. It would be more beneficial for the company to focus on selling and advertising more of the luxury and technological items, since they can easily attain more revenue by selling fewer items than selling a lot more food items to get the same revenue as the higher priced classes. Luxury items are distributed among price axis, this is an indication of variability in luxury items' prices.

Process capability indices:

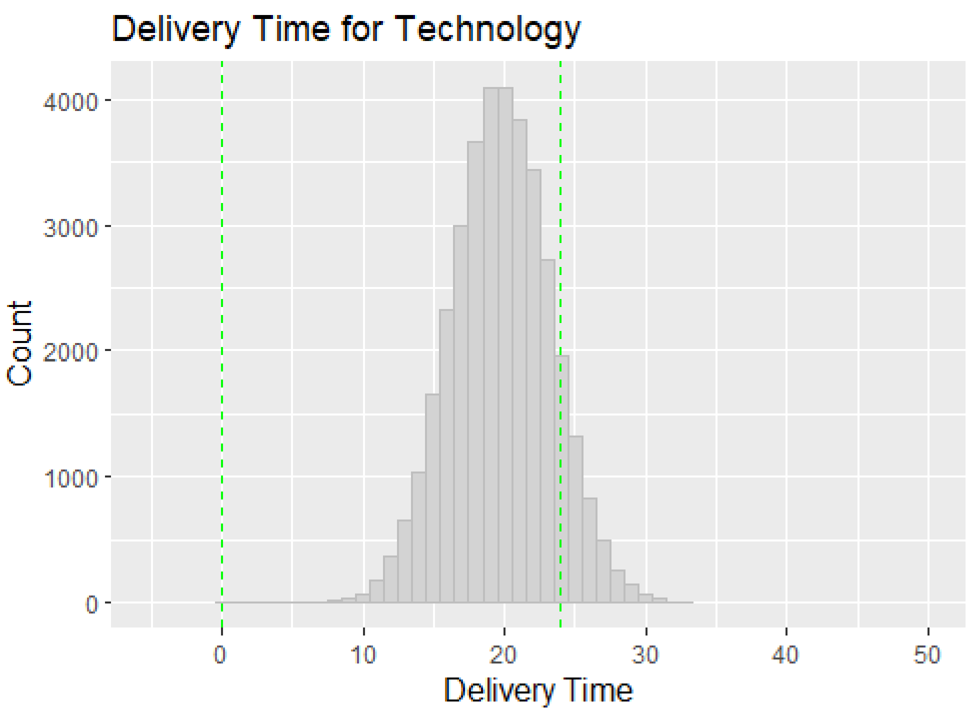


Figure 17 Process capability chart for technology delivery time

The USL of 24 and LSL of 0 are given to calculate the required process capability indices. The LSL of 0 makes sense as time cannot be a negative value.

Using USL = 24

and LSL = 0

process capability indices obtained:

CP	1.142
CPU	0.404
CPL	1.881
CPK	0.404

Figure 18 Process capability indices

Potential capability:

Cp (Process capability Ratio) is an indicator of how the distribution compares to specification width. The CP shows that the process is capable as it is more than one. Technology can be delivered within the required specifics.

The Cpk value (process capability index) is an indicator of whether there is conformance to the specifications. A low Cpk value suggests that a process can benefit from improvement, while a higher Cpk value assures a more complete process. The CPK is less than the CP which indicates that the process is not centered between the specified limits. This shows that the process could benefit from improvement by shifting the mean to the left.

The benchmark of a Cpk of 1.33 is used in many industries to analyse the process capability. As the company's Cpk value is much lower than this benchmark value, it is an indication that there should be looked at ways to improve the process of the company, reducing change.

PART 3: Statistical Process Control

The X&S chart for delivery times is plotted by using 30 samples of 15 instances of Sales each. First, the data is ordered according to date, before charts are plotted. The oldest to newest data is ordered by year, then by month and finally by day.

Values for X-chart

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22.974616	22.107892	21.241168	20.374444	19.507721	18.640997	17.774273
Clothing	9.404934	9.259956	9.114978	8.970000	8.825022	8.680044	8.535066
Household	50.248328	49.019626	47.790924	46.562222	45.333520	44.104818	42.876117
Luxury	5.493965	5.241162	4.988359	4.735556	4.482752	4.229949	3.977146
Food	2.709458	2.636305	2.563153	2.490000	2.416847	2.343695	2.270542
Gifts	9.488565	9.112747	8.736929	8.361111	7.985293	7.609475	7.233658
Sweets	2.897042	2.757287	2.617532	2.477778	2.338023	2.198269	2.058514

Figure 19 X-chart table

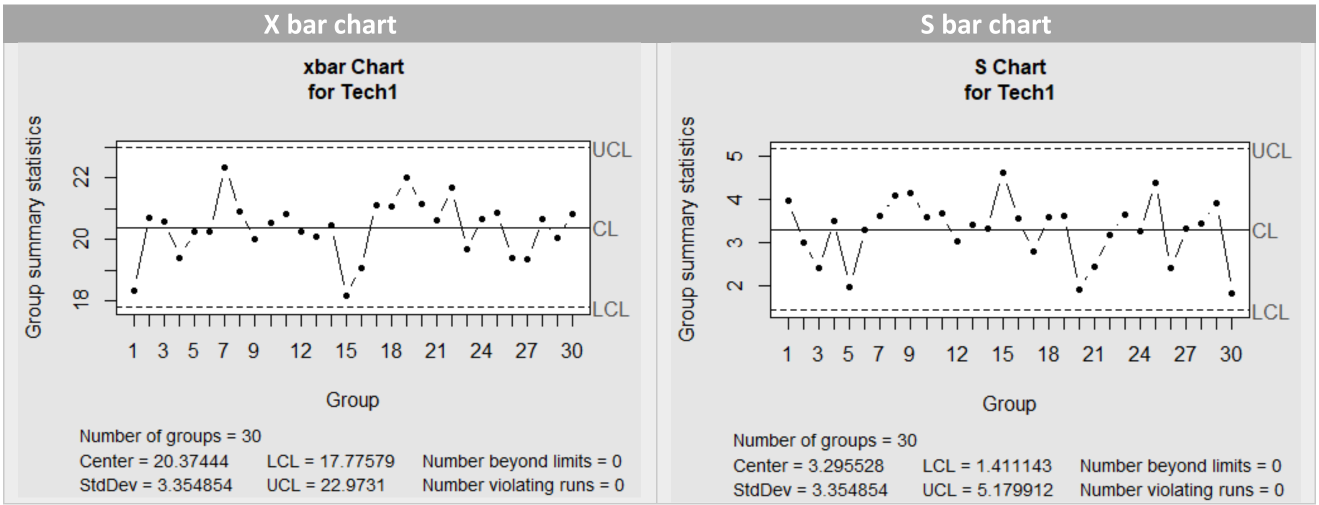
Values for S-chart

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.1805697	4.5522224	3.9238751	3.2955278	2.6671805	2.0388332	1.4104859
Clothing	0.8665596	0.7614552	0.6563509	0.5512465	0.4461422	0.3410379	0.2359335
Household	7.3441801	6.4534101	5.5626402	4.6718703	3.7811003	2.8903304	1.9995605
Luxury	1.5110518	1.3277775	1.1445032	0.9612289	0.7779546	0.5946803	0.4114060
Food	0.4372466	0.3842133	0.3311800	0.2781467	0.2251134	0.1720801	0.1190468
Gifts	2.2463333	1.9738773	1.7014213	1.4289652	1.1565092	0.8840532	0.6115971
Sweets	0.8353391	0.7340215	0.6327039	0.5313862	0.4300686	0.3287509	0.2274333

Figure 20 S-chart table

30 first samples graphs

Technology:



These first 30 samples show that the technology class is controlled as the graphs do not spike beyond the upper and lower control limits. Thus, no variation is caused in the process of ordering technology. The satisfactory S bar ensures that the X-bar chart can be evaluated.

Clothing:

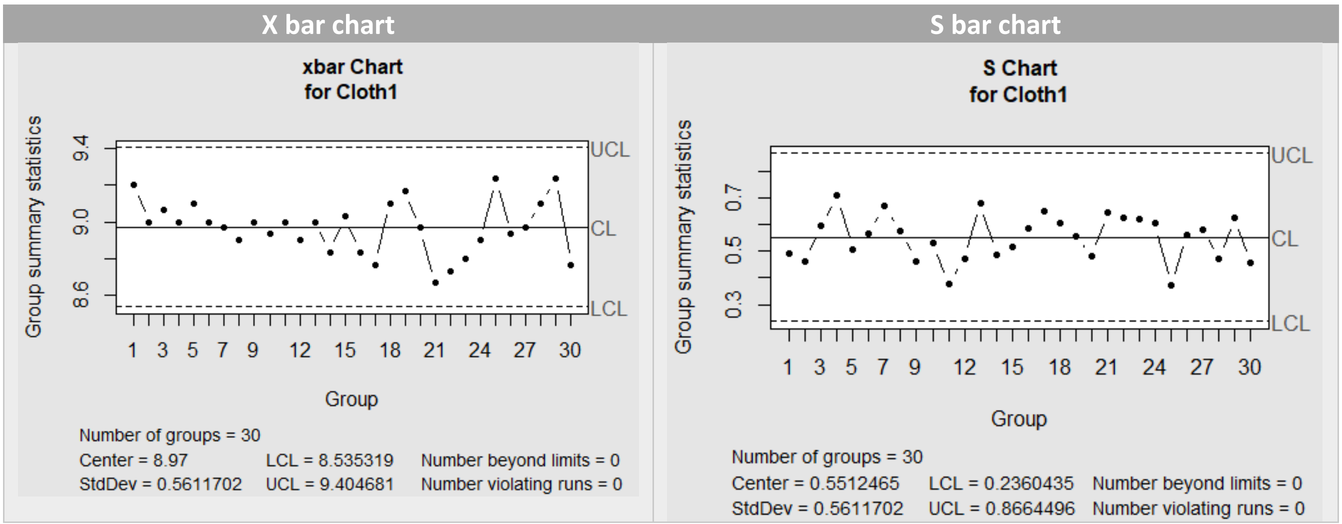


Figure 3.2: X&S Charts of Clothing

These first 30 samples show that the clothing class is controlled as the graphs do not spike beyond the upper and lower control limits. Thus, no variation is caused in the process of ordering clothing. The satisfactory S bar ensures that the X-bar chart can be evaluated.

Household:

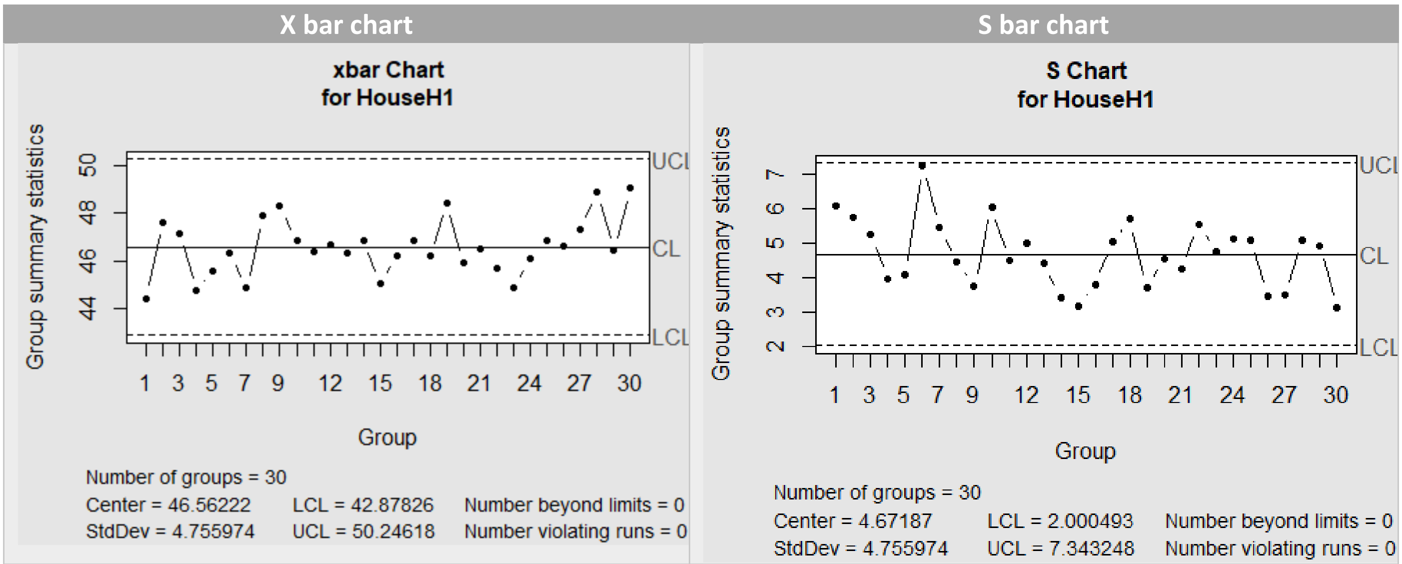


Figure 3.3: X&S Charts of Household

These first 30 samples show that the household class is controlled as the graphs do not spike beyond the upper and lower control limits. Thus, no variation is caused in the process of ordering household items. The satisfactory S bar ensures that the X-bar chart can be evaluated.

Luxury:

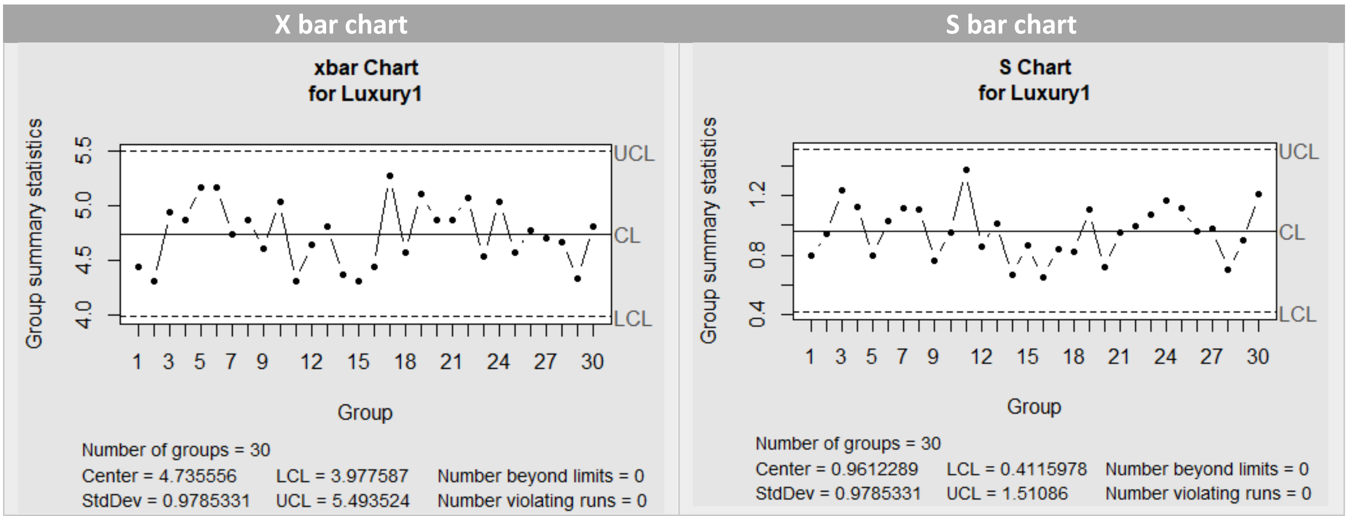


Figure 3.4: X&S Charts of Luxury

These first 30 samples show that the luxury class is controlled as the graphs do not spike beyond the upper and lower control limits. Thus, no variation is caused in the process of ordering luxury items. The satisfactory S bar ensures that the X-bar chart can be evaluated.

Food:

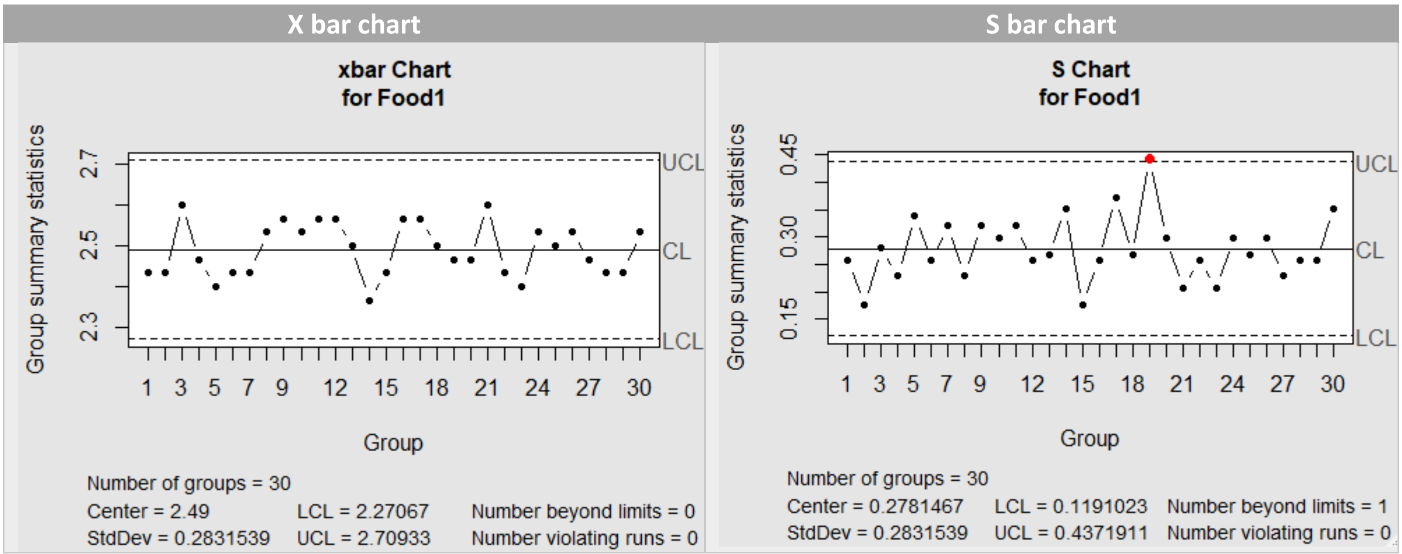


Figure 3.5: X&S Charts of Food

These first 30 samples show that the Food class is controlled as the graphs does not spike beyond the upper and lower control limits. Except for sample 19, the sample's standard deviation spikes beyond the upper control limit. This is an indication that this sample needs to be removed.

Gifts:

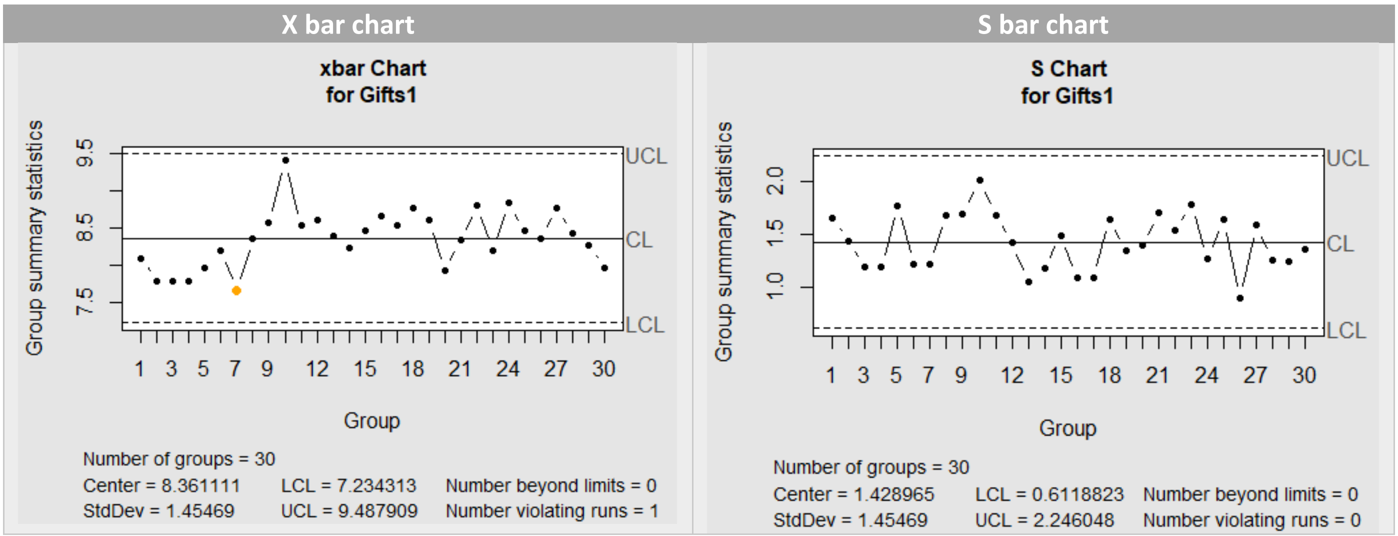
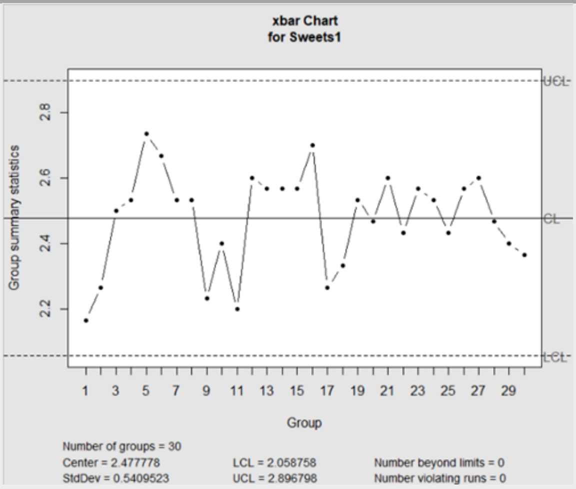


Figure 3.6: X&S Charts of Gifts

These first 30 samples show that the gift class is controlled as the graphs does not spike beyond the upper and lower control limits. Thus, no variation is caused in the process of ordering gifts. The satisfactory S bar ensures that the X-bar chart can be evaluated.

Sweets:

X bar chart



S bar chart

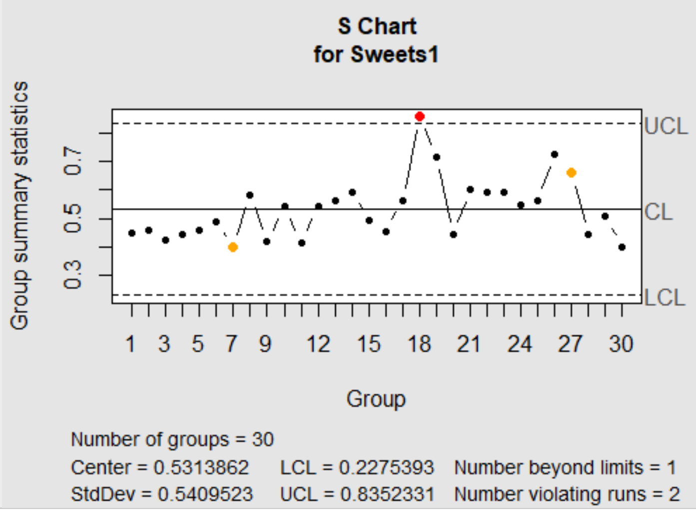


Figure 3.7: X&S Charts of Sweets

These first 30 samples show that the Sweets class is controlled as the graphs does not spike beyond the upper and lower control limits. Except for sample 18, the sample's standard deviation spikes beyond the upper control limit. This is an indication that this sample needs to be removed.

3.2 GRAPHS FOR ALL SAMPLES:

Technology:

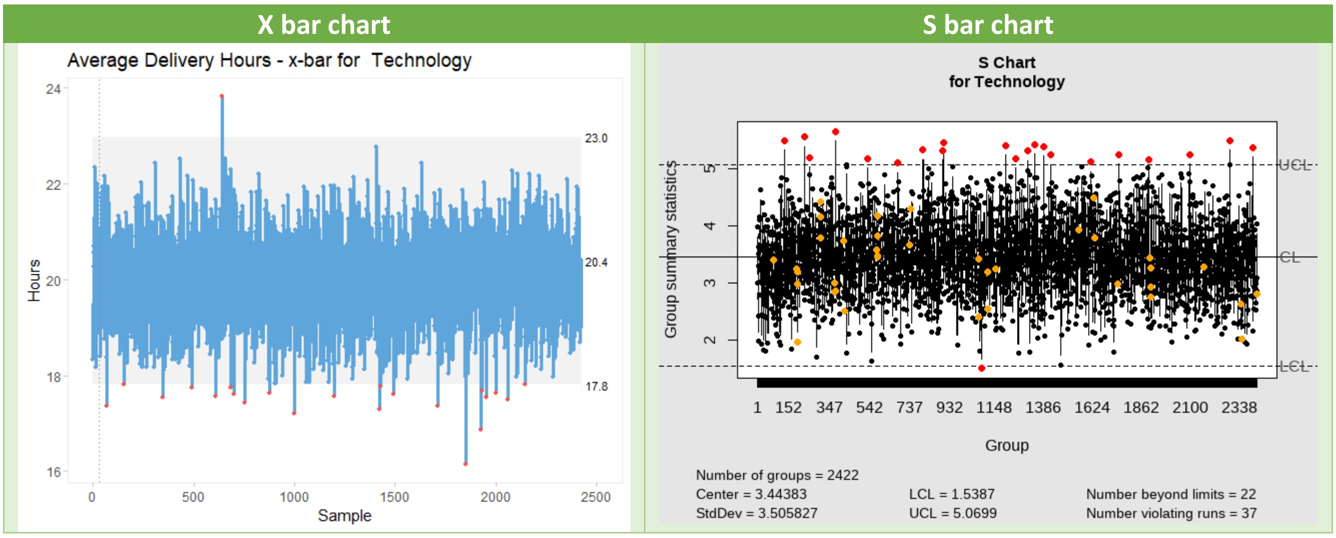


Figure 3.8: X&S bar chart for technology (samples)

Most samples are within control limits. The Technology class seems to be controlled. The S-bar chart is under control (only 22 samples out of control limits), so the conclusion is thus appropriate.

Clothing:

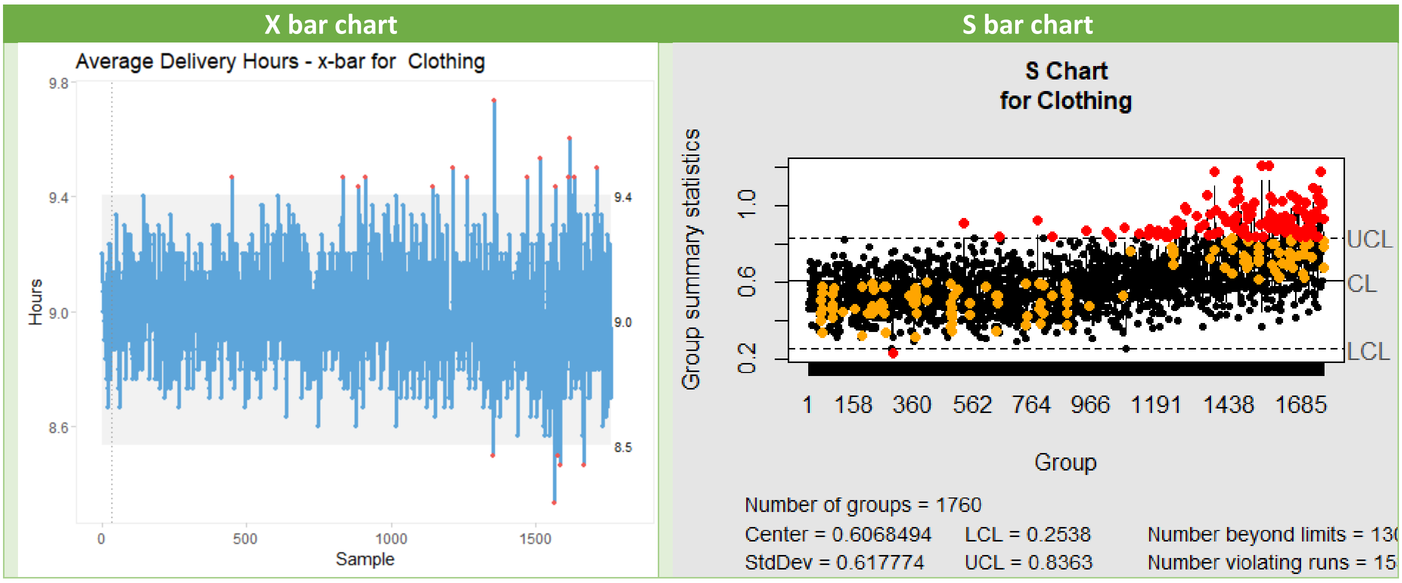


Figure 3.9: X&S bar chart for clothing (samples)

Most samples are within control limits. The Clothing class seems to be controlled, with a few odd occurrences where the samples exceed limits. This could be caused due to seasonal changes. There are quite a few samples beyond control limits for the S-bar chart, but even when these samples would be removed, the results for the X-bar chart would remain the same. Therefore, the conclusion for the X-bar chart is accepted.

.

Household:

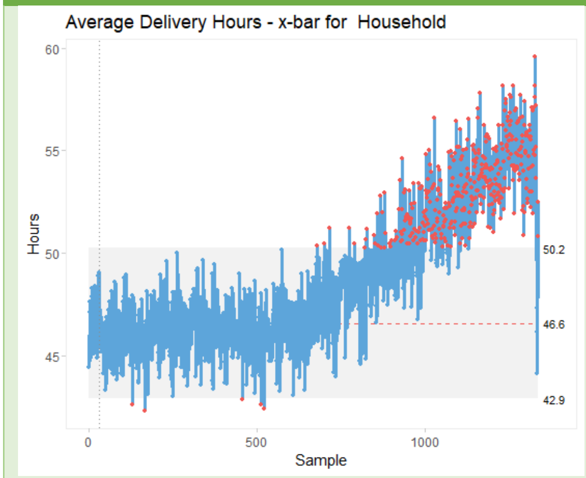
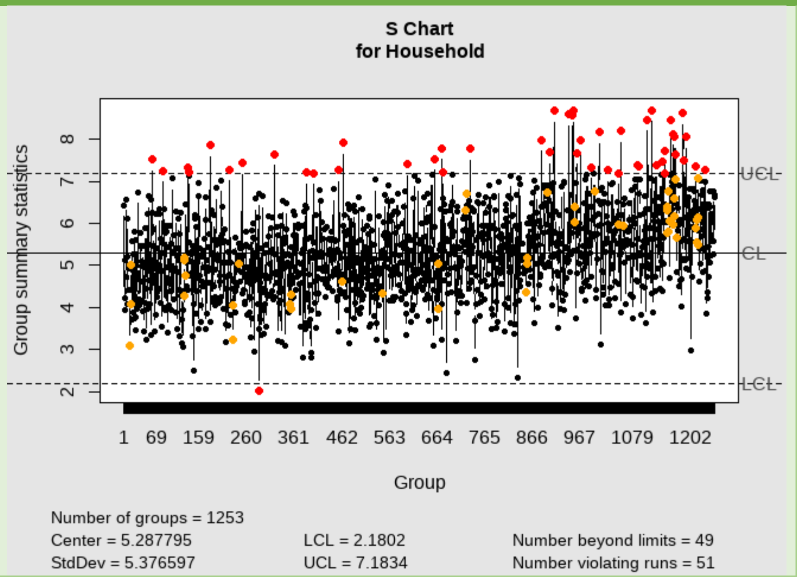
X bar chart	S bar chart
	

Figure 3.10: X&S bar chart for Household (samples)

The delivery time for household products increased. The reason for this increase needs to be investigated. The delivery time for household products is uncontrolled and unstable. The positive trend seems to be increasing continuously after the 579th sample. This increase could be due to increased sales of larger household products which require more handling work.

Luxury:

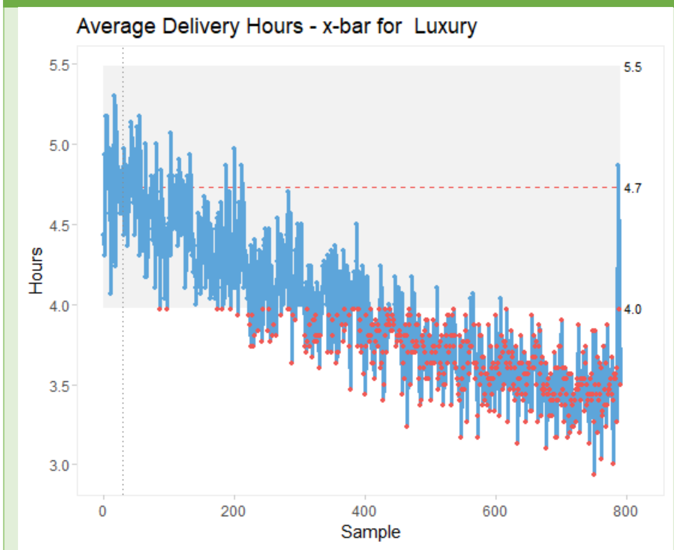
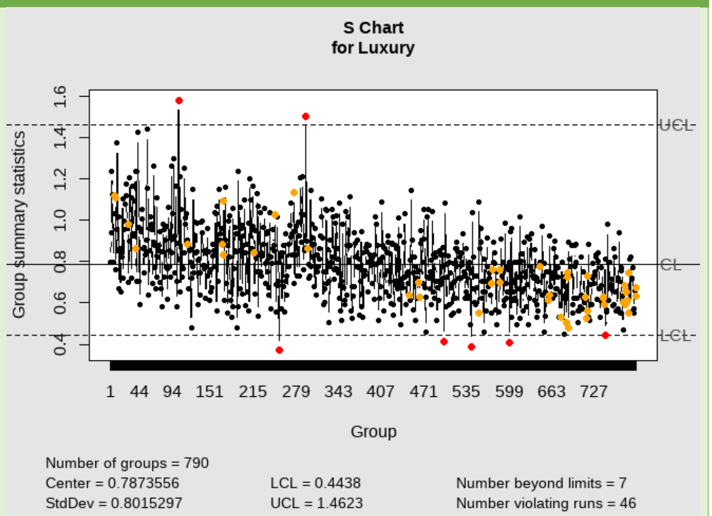
X bar chart	S bar chart
	

Figure 3.11: X&S bar chart for luxury (samples)

Luxury delivery time decreased. This could be because luxury high-value product class and thus needs to deliver fast to ensure a high revenue of luxury items. Luxury seems to continuously decrease out of the control limits after the 191st sample. A professional in the sales department needs to investigate the reason for this decrease. The decrease could indicate that the company has put more emphasis on delivering luxury items because luxury items are the highest-valued products and revenue will increase if the items are delivered faster (the customers will be more satisfied and buy more luxury products). The S-bar chart is under control (only 2 samples out of control limits), therefor the conclusion of the X-bar chart is appropriate.

Food:

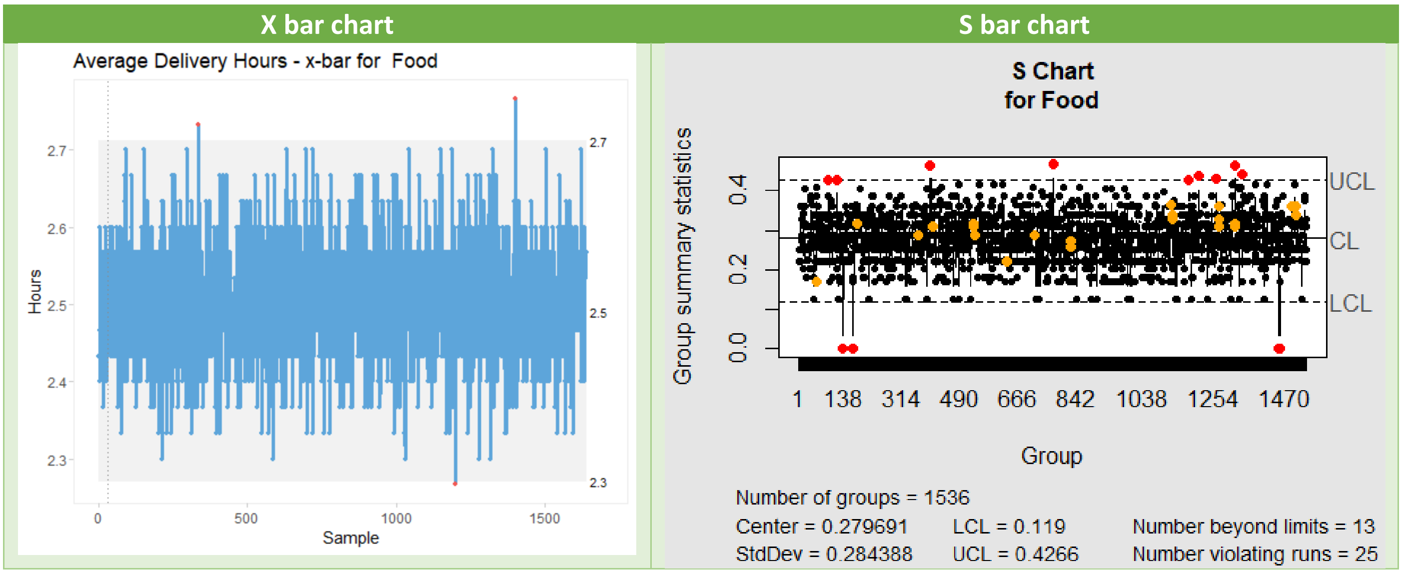


Figure 3.12: X&S bar chart for food (samples)

The Food class seems to be controlled. A few instances exceed the control limits, but the chart is stable for now since the last limit exceeding the sample is at the 1111th sample. The S-bar chart is controlled (only 13 samples out of control limits), therefor the conclusion of the X-bar chart is appropriate.

Gifts:

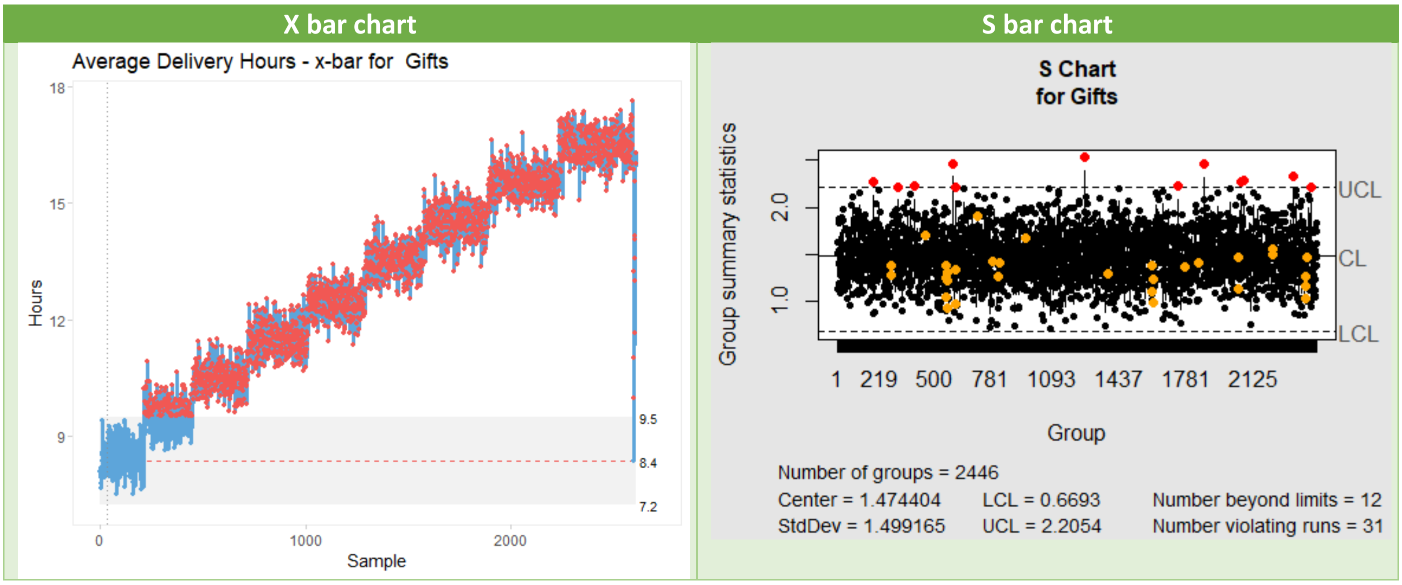


Figure 3.13: X&S bar chart for gifts (samples)

There is an increase in delivery time for gifts. It could be useful to investigate the reason for this. There is an indication that the delivery times for gifts is uncontrolled and unstable. This could be due to the large total amount of products that need to be shipped at certain delivery times. The company might not be able to handle these amounts. It could also be caused by an unfixed logistics problem or an increased demand for gifts. The S-bar chart is controlled (only 12 samples exceed control limits), therefor the conclusion of the X-bar chart is appropriate.

Sweets:

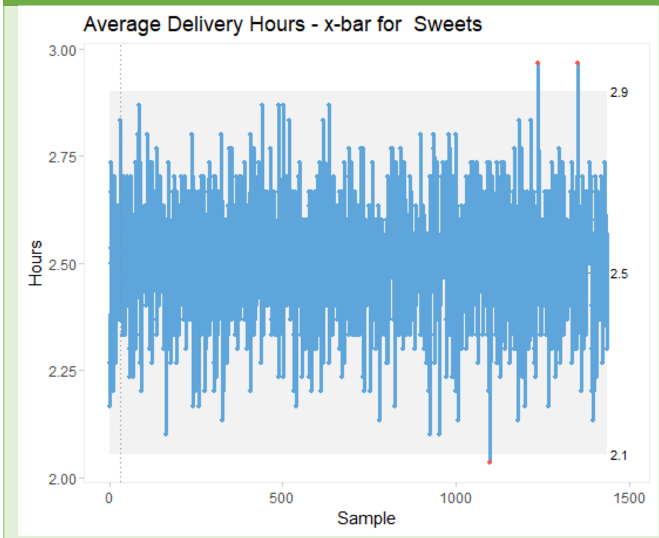
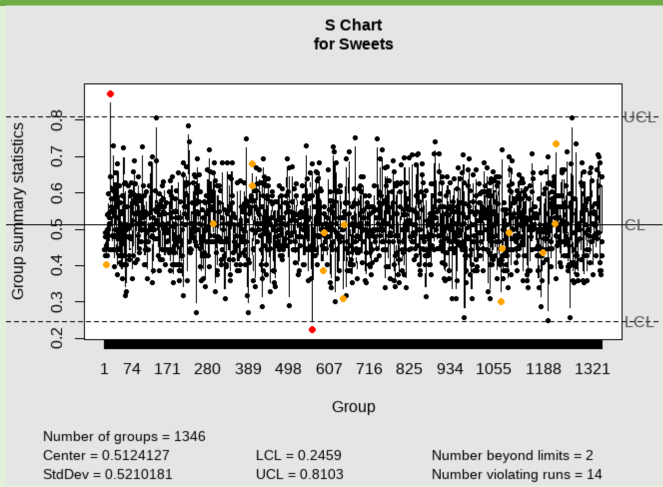
X bar chart	S bar chart
	

Figure 3.14: X&S bar chart for sweets (samples)

The Sweets class seems to be controlled. Eight samples exceed the limits, with some occurring rather recently. This occurrence needs investigation. The S-bar chart is controlled (with only 2 samples outside the control limits), thus the conclusion of the X-bar chart is appropriate.

PART 4: Optimizing delivery processes

Xbar sample mean outside of the outer control limit (using control limits calculated in 3.1)

Samples beyond control limits:

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

Figure 21 table of samples control limits

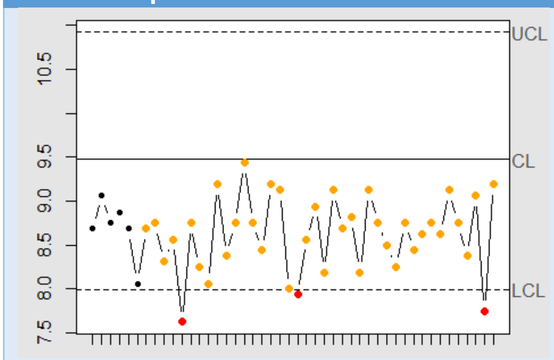
Clothing, Food, Technology and Sweets classes are controlled , because samples that go beyond the control limits. Household, Luxury and Gift items are out of control because of the high number of samples that is not within the upper and lower control limits. Further investigation is recommended to determine why these items' delivery times are not controlled.

Plots of the first 3 and last 3 samples out of control limits

Only Household, Gifts and Luxury are plotted as these classes have the highest occurrences of samples outside the control limits. This is an indication that almost all deliveries will be expected over all the samples for classes technology, clothing, food and sweets. Deliveries will not be expected on time for the luxury, household and gift classes. The following plot shows the first three as well as the last three samples that did not meet control specifications.

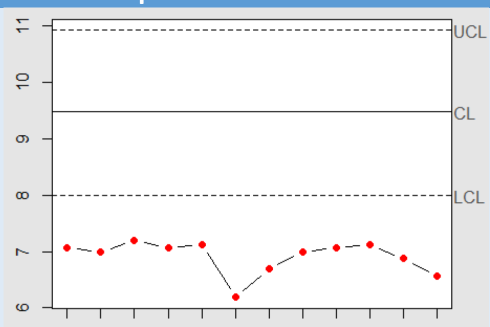
Luxury:

First 3 samples



Showing samples between 150 and 195

Last three samples

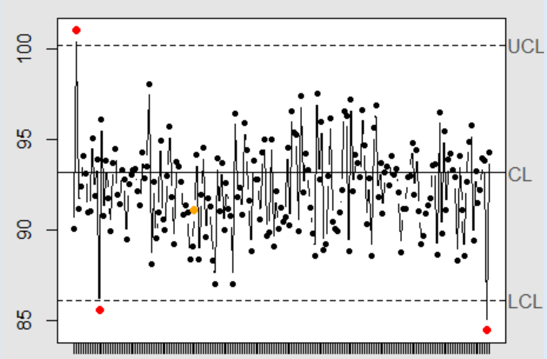


Showing samples between 730 and 741

Figure 4.1: Luxury's first 3 and last 3 out of the control limits

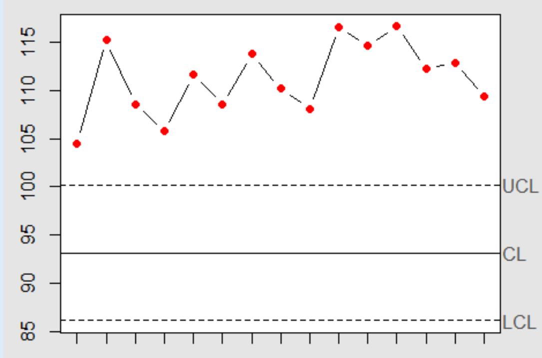
Household:

First 3 samples



Showing samples between 421 and 604

Last three samples

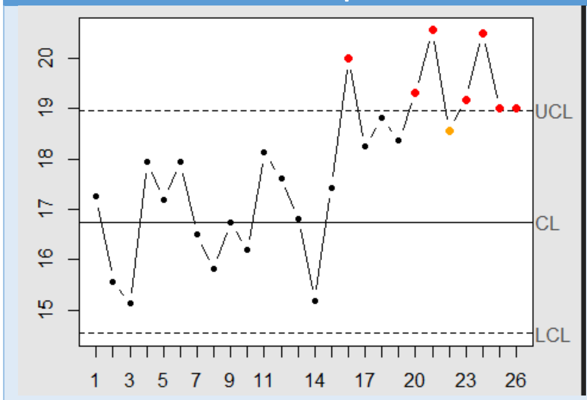


Showing samples between 1240 and 1254

Figure 4.2: Household's first 3 and last 3 out of the control limits

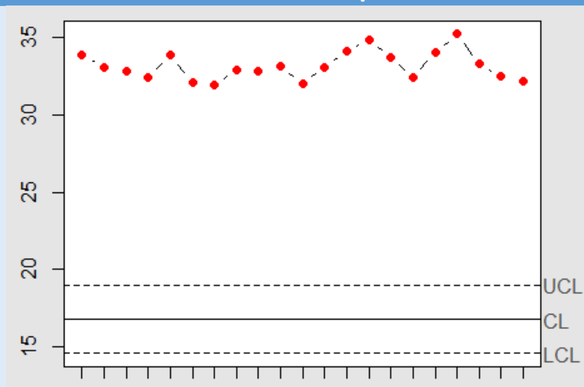
Gifts:

First 3 samples



Showing samples between 185 and 210

Last three samples



Showing samples between 2426 and 2446

Figure 4.3: Gifts first 3 and last 3 out of the control limits

The samples outside the control limits happened recently due to these samples being the last samples of the classes.

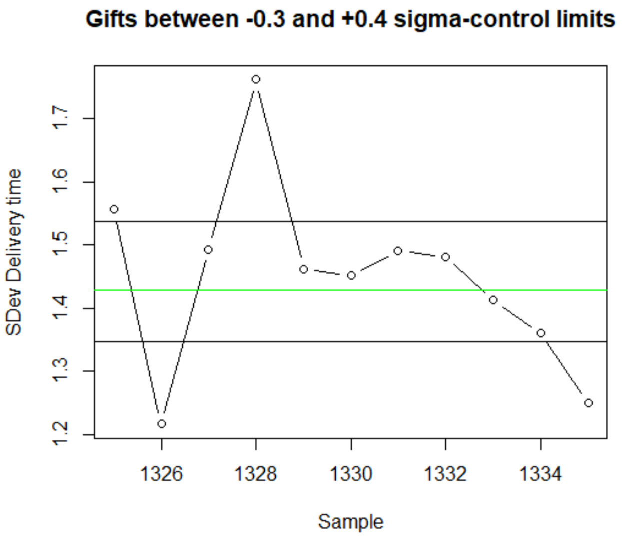
Most consecutive samples of "s-bar" between -0.3 and 0.4 sigma control.

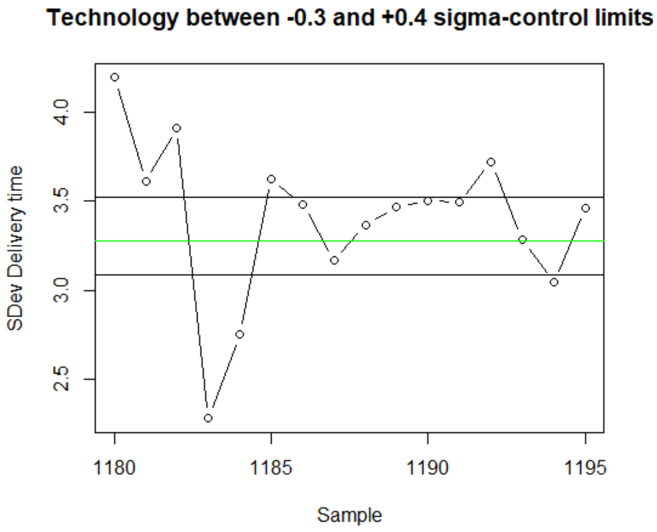
Class	maximum between -0.3 & 0.4 sigma	Position of first	Last Sample position
Clothing	5	665	665
Household	4	253	761
Food	5	752	905
Technology	6	1191	1191
Sweets	5	692	692
Gifts	6	1334	1334
Luxury	3	230	230

The maximum number of f samples between the limits for the S-bar is equal to 6. This value is rather low, indicating that plenty of samples for all relevant classes are beyond the 0.4 and -0.3 sigma control limits.

The Gifts and Technology classes have the highest number of consecutive samples between -0.3 and 0.4 sigma control limits. This implies that these classes will be more stable within the specified limits compared to the rest of the classes.

S-chart mean





Estimate the likelihood of making type 1 error for A and B

To calculate the type I error, the following Null Hypothesis assumption is made:

- **H0:** The process is in control and centred on the centre line (mean within control limits)
- **H1:** The process is out of control, is not cantered on the centre line with increased/decreased variation (mean not within control limits)

	Process is fine	*The process is not fine*
SPC indicated the process is not fine	Type 1 error or Manufacturer's error	Correct to fix the process
SPC indicated the process is fine	Correct to do nothing	Type 2 error or Consumer's error.

Table 3.3: Difference between type I and type II error

Question	*Probability of performing type 1 error*
A	The probability of making a mistake with A is 0.002699796 (0.27%) . This is an indication that the probability of mistakenly assuming that products are not delivered on time, when the products are delivered on time.
B	The probability of making a mistake with B is 0.13165941 (13.17%).

Table 3.4: Probability of making type 1 error for questions A and B

It can be concluded that a Type I error has been made. The probabilities of making a type l error for A and B are shown in the figure below.

Minimizing delivery cost

To determine the minimum delivery cost associated with the technology class, it is necessary to compare the costs of all relevant delivery times (in hours) to find the exact hour associated with the least cost. This result will be given by plotting all the delivery times and their associated costs and finding the global minima.

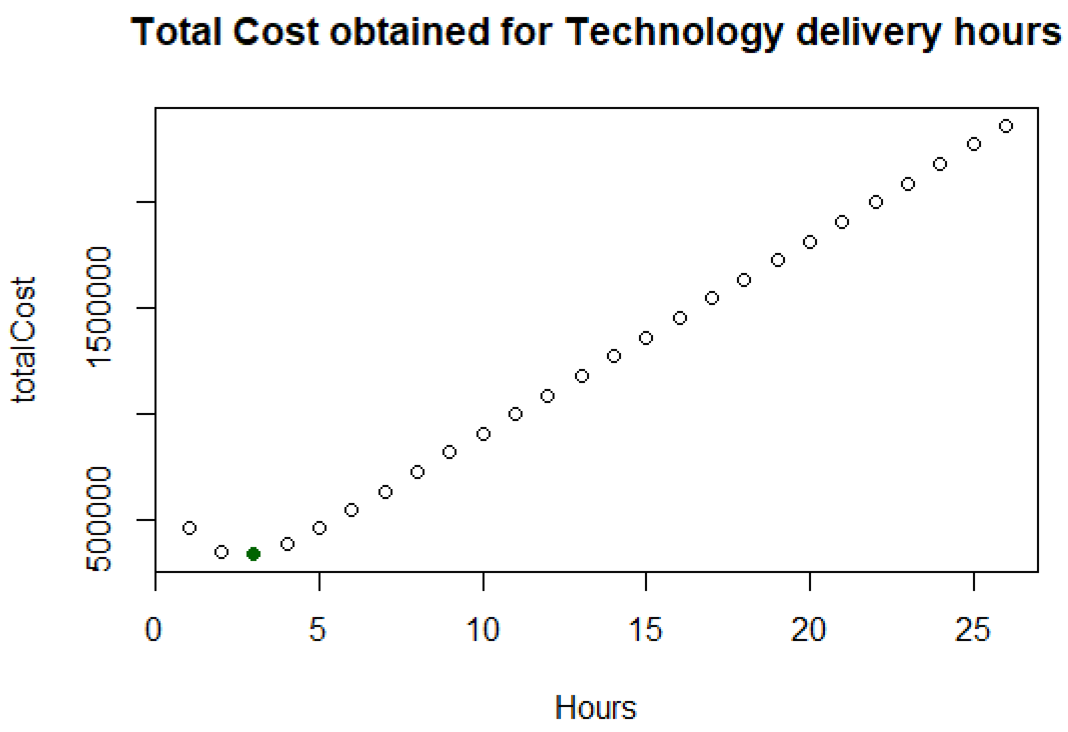


Figure 22 Graph of total cost for delivery times of technology

.

The figure above shows that the best delivery time center is three hours. A cost of R340870 will be added when reducing the delivery time by three hours. The weighted average will decrease from 20 to 17 hours. Delivery times exceeding 26 hours will be more costly when compared to the price of reducing delivery time by three hours

This is like the Taguchi Loss as Taguchi claimed that loss is even possible when the specifications are not exceeded. The consumer is at its happiest when the product is perfectly on target (on time in this case), thus any deviation will result in a growing loss. The loss is not a sudden drop but rather starts dropping the moment the product deviates from the promised optimal delivery time.

As seen in the graph above, the parabolic curve mimics the parabolic curve resulting from the Taguchi Loss function. The global minimum on the parabolic curve represents the minimized loss and thus minimized cost.

Estimate the likelihood of making a type II error for A

Used LCL and UCL of first 30 sample limits

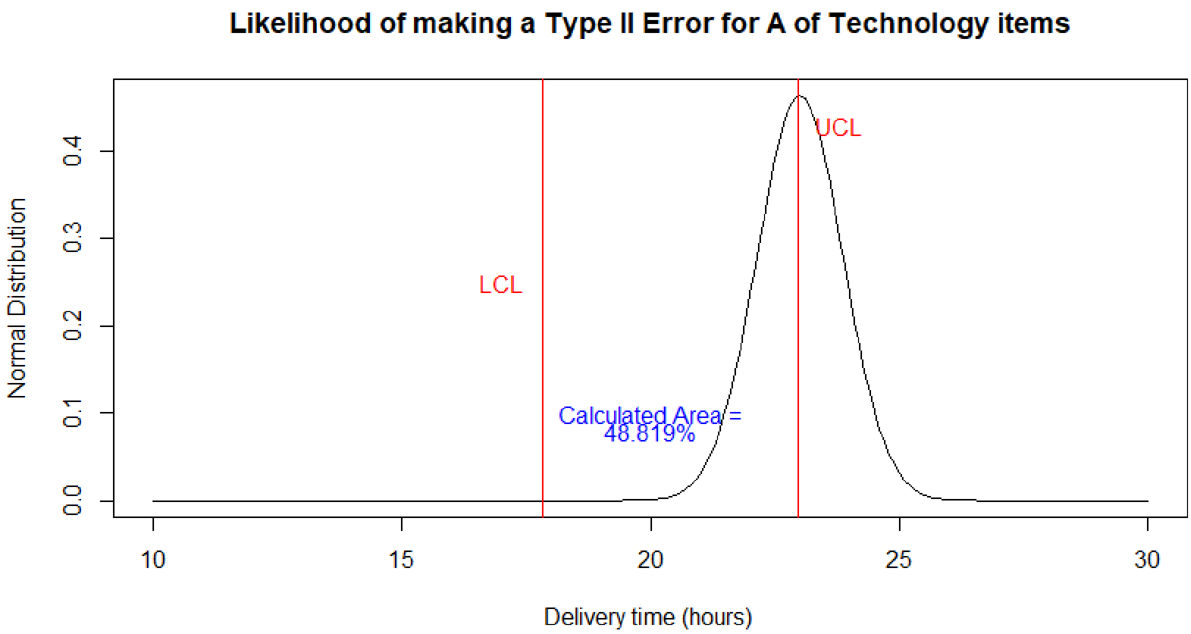


Figure 23 likelihood of Type II error

A type II error, or a false negative, is known as the probability of mistakenly failing to reject the null hypothesis even when it does not apply to the whole population. The case for delivery times occurs when the product is assumed to be delivered on time, when it is not delivered on time.

The probability of making a type II error for the delivery time of the Technology class when the delivery process average moves to 23 hours, is 0.48819. this indicates that there is a 48.82% chance of mistakingly thinking that products of the technology class was delivered on time, while in reality, it was late.

This error can have an implication on customer satisfaction as the products will not be delivered on time, making the company less reliable. The probability of making this mistake is rather high and the company must take action to ensure that the products are delivered on time, rather than just assuming that the products arrive at the customers on time.

PART 5: DOE and MANOVA test

Hypothesis 1:

H0: The class does not influence the price, age and delivery time of the product.

H1: The class of a product influences the price, age or delivery time of the product.

Independent variable: Class.

Dependent variables: Price, Age and Delivery time.

P value: 0.05 (most popular/universal P value)

Manova test: test whether there is a feature that has an influence

P value	<2.2e-16, which is smaller than 0.05. Thus, Reject Null Hypotheses. At least one dependent variable's average is different. Either Price, Delivery times or age has a significant difference among Classes.

Each dependent variable and class:

```
 Response Delivery.time :
                Df   Sum Sq Mean Sq F value    Pr(>F)
Class            6 33461034 5576839  629489 < 2.2e-16 ***
Residuals   179976  1594464       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 5.7165e+13 9.5275e+12   80238 < 2.2e-16 ***
Residuals   179976 2.1370e+13 1.1874e+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  8423110 1403852  3805.2 < 2.2e-16 ***
Residuals   179976 66397874     369
```

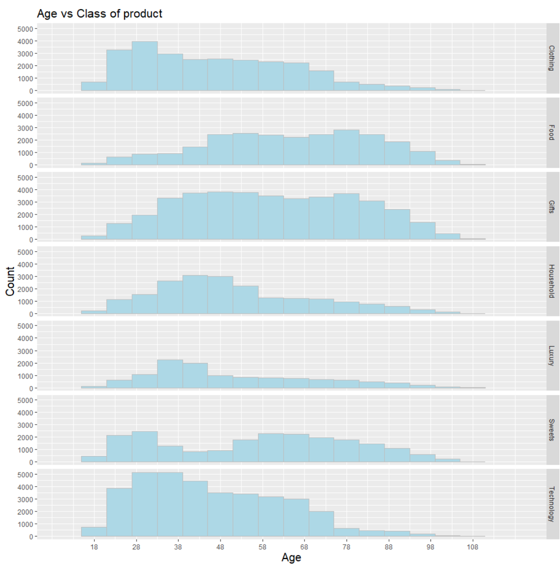
Figure 24 dependant variables p-values Manova 1

Dependent variable	P value	Analyses
Price	2.2e-16	P value < 0.05. This means Price is influenced depending what class the product is.
Delivery times	2.2e-16	P value < 0.05. This means Delivery times are influenced by the class of the product.
Age	2.2e-16	P value < 0.05. This means Age is influenced by class of the product.

Visualization:

Figure 25 Price vs Class

From the boxplot above it is clear that the class of product influences the price of the product. For luxury items the price tends to be the highest, with technology following the second highest prices. Food, clothing and sweets have the lowest prices.



From the graph above the class of product will influence which age of buyers will purchase the products. Clothing and technology are purchased by younger age groups, it could be useful to make sure clothing and technology advertisements reach these age groups, as well as make sure to be up to date with clothing and technology trends for younger people to ensure the right products are on the market. Food and household items are bought by middle-aged people who might have families they buy food and household item for. Food and Household items' advertisements should reach these age groups. Gifts are relatively uniformly distributed as all age groups have the need to buy gifts from time to time, with their age irrelevant (older people seem to buy fewer gifts, this may be due to having fewer friends and family and fewer finances).

Figure 26 Delivery time vs Class

From the graphs above, the delivery time for each class differs depending on the class of the product.

Clothing, sweets,food and luxury items are delivered relatively quickly compared to Household items. A reason for this could be that household items tend to be larger and more time-consuming to deliver. Service delivery reliability decreases.

Luxury items are more valuable products and it can be seen from the buying pattern that these products are more expensive. To increase revenue, the company needs to ensure high reliability and customer satisfaction for these class items. This could be a reason for the shorter delivery times seen on the graph above.

Conclusion:

The null hypothesis is rejected as the class influences the delivery time, the price and the age group buying the product.

Hypothesis 2:

H0: The reason why a product is bought is not influenced by the day, month, year in which product is bought.

H1: at least one of the features (Day, month, year) has an influence on patterns in the reason for the purchase of the product.

Independent variable: Why bought

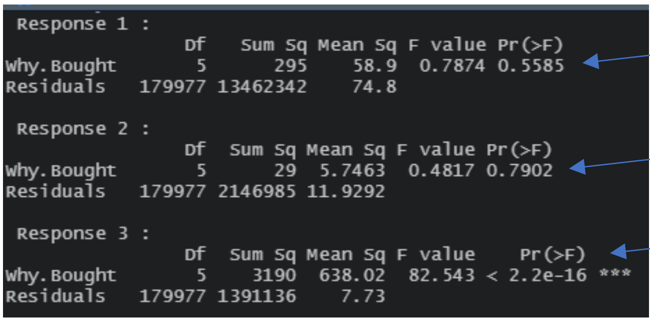
Dependent variables: Day, month, year

P value: 0.05 (most popular/universal P value)

Manova test

P value	<2.2e-16 which is less than 0.05. Reject Null Hypotheses. At least one dependent variable has a different average. Either day, month or year in which a product is purchased has a significant difference in the reason for purchase.

Each dependent variable and class:



```
 Response 1 :
               Df    Sum Sq Mean Sq F value Pr(>F)
Why.Bought      5       295    58.9  0.7874 0.5585
Residuals  179977 13462342    74.8

 Response 2 :
               Df   Sum Sq Mean Sq F value Pr(>F)
Why.Bought      5       29  5.7463  0.4817 0.7902
Residuals  179977 2146985 11.9292

 Response 3 :
               Df   Sum Sq Mean Sq F value    Pr(>F)
Why.Bought      5     3190  638.02  82.543 < 2.2e-16 ***
Residuals  179977 1391136    7.73
```

Figure 27 each dependent variables P-value for Manova 2

Dependent variable	P value	Analyses
Day	0.5585	P value > 0.05. Reason for purchase is not influenced by the day on which it is bought.
Month	0.7902	P value > 0.05. Reason for purchase is not influenced by the month on which it is bought.
Year	2.2e-16	P value < 0.05. This means Year has an influence on the reason for purchase.

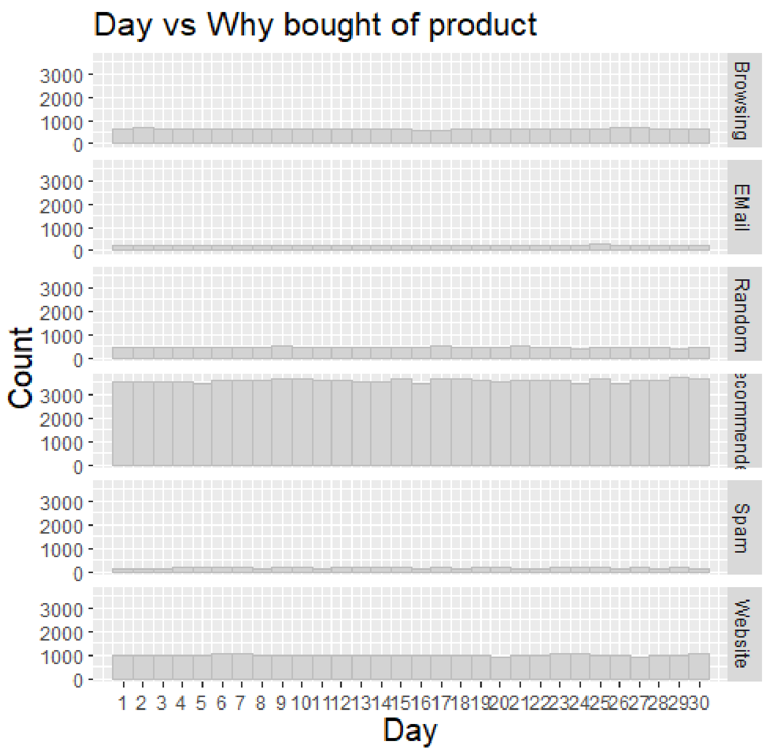


Figure 28 Day vs reason for purchase

As seen in the graph above, the count for reason for purchase for each day is uniformly distributed, thus the day does not influence the reason why a product is bought.

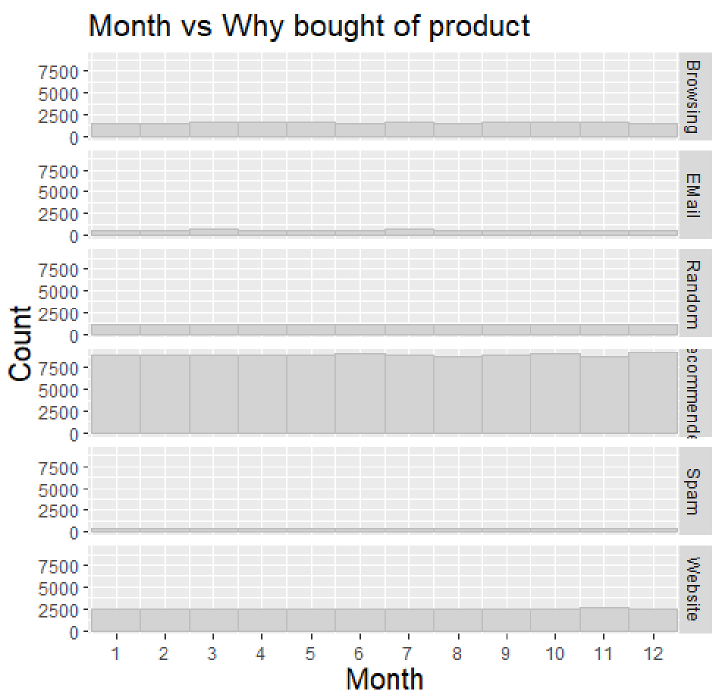


Figure 29 month vs reason for purchase

As seen in the graph above, the count for reason for purchase for each month is uniformly distributed, thus the month does not influence the reason why a product is bought.

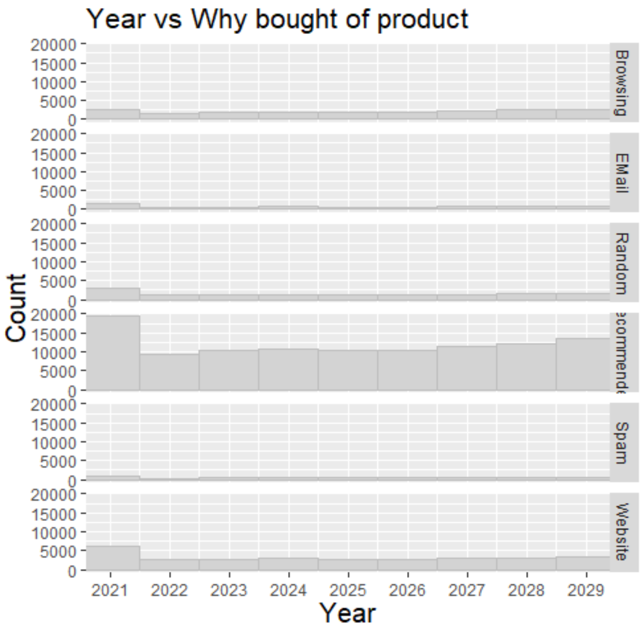


Figure 30 Year vs Why Bought

As seen on the graph above the year had an influence on the count of reasons for purchase. The variation in the count can be seen in the Recommendation and Website reasons for purchase.

Conclusion:

The class for the products does not influence the day and month on which sales take place. The year, however, influences when a certain class product is bought more. The reason why a customer bought a product for "recommend" and "website" decreased from the year 2021, but started to pick up again after the year 2026. Although the decrease is not that significant, it could still be useful to investigate this decrease as well as the reason for the slight increase.

Hypothesis 3:

H0: The class of a product that is bought is not influenced by the day, month, year in which product is bought.

H1: at least one of the features (Day, month, year) has an influence on patterns in a class of a product.

Independent variable: Class

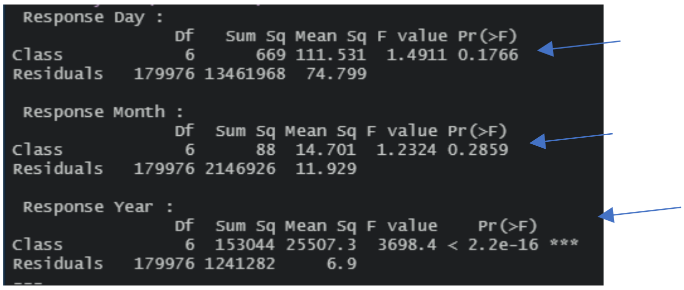
Dependent variables: Day, month, year

P value: 0.05 (most popular/universal P value)

Manova test:

P value	<2.2e-16 is less than 0.05 Reject Null Hypotheses. At least one dependent variable has a different average. Either day, month or year in which a product is purchased has a significant difference among the class of product.

Each dependent variable and class:



```
 Response Day :
                Df   Sum Sq Mean Sq F value Pr(>F)
Class            6      669 111.531  1.4911 0.1766
Residuals   179976 13461968  74.799

 Response Month :
                Df  Sum Sq Mean Sq F value Pr(>F)
Class            6      88  14.701  1.2324 0.2859
Residuals   179976 2146926  11.929

 Response Year :
                Df  Sum Sq Mean Sq F value    Pr(>F)
Class            6  153044 25507.3  3698.4 < 2.2e-16 ***
Residuals   179976 1241282     6.9
```

Figure 31 dependant variables P-values Manova 3

Dependent variable	P value	Analyses
Day	0.1766	P value > 0.05. Class of the product bought is not influence by the day on which it is bought.
Month	0.2859	P value > 0.05. Class of the product bought is not influence by the month on which it is bought.
Year	2.2e-16	P value < 0.05. This means Year in which a product is bought will influence the class of the product which is bought.

Visualization:

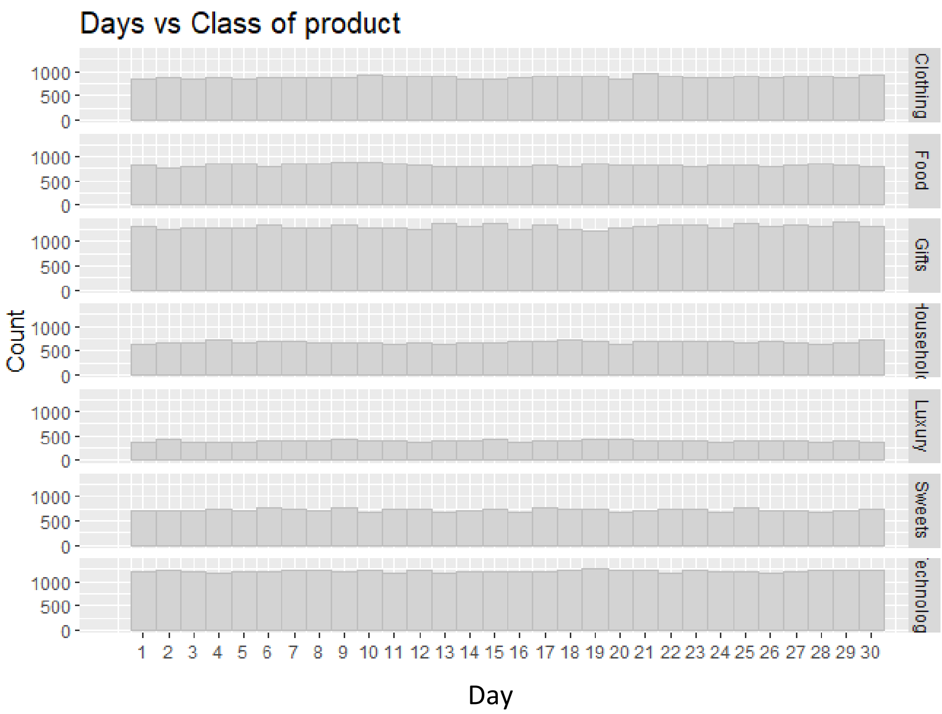


Figure 32 Days vs Class

As seen on the graph above, the sales for each class are uniformly distributed over each day. Thus, days do not have an influence on the sales of different classes

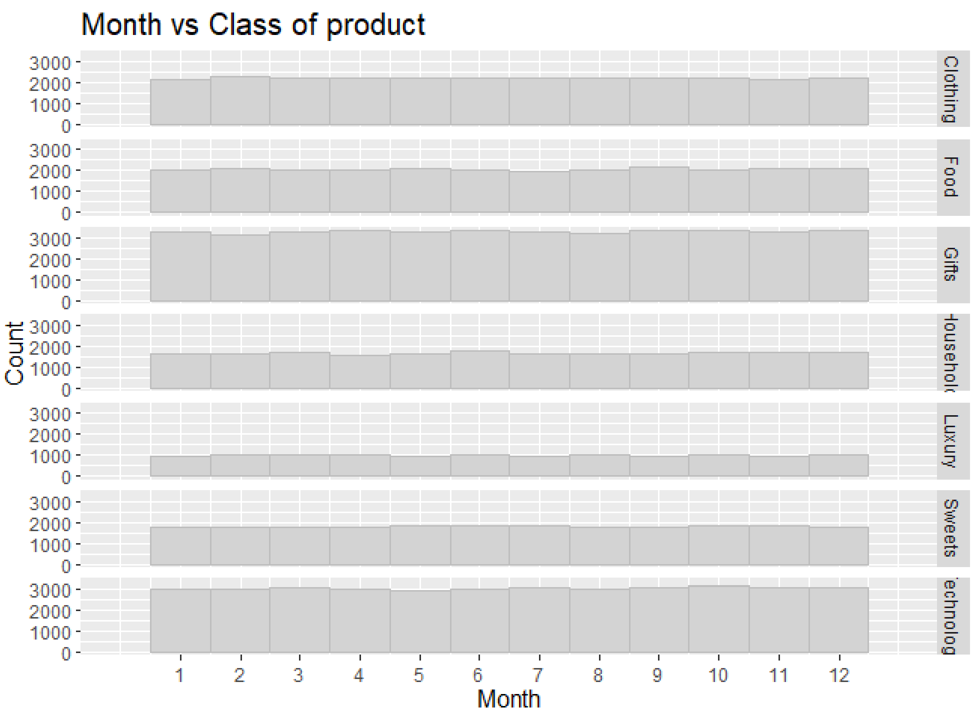


Figure 33 Month vs Class

As seen on the plot above, the sales for each class are uniformly distributed over each month. Thus, months do not have an influence on the sales of different classes.

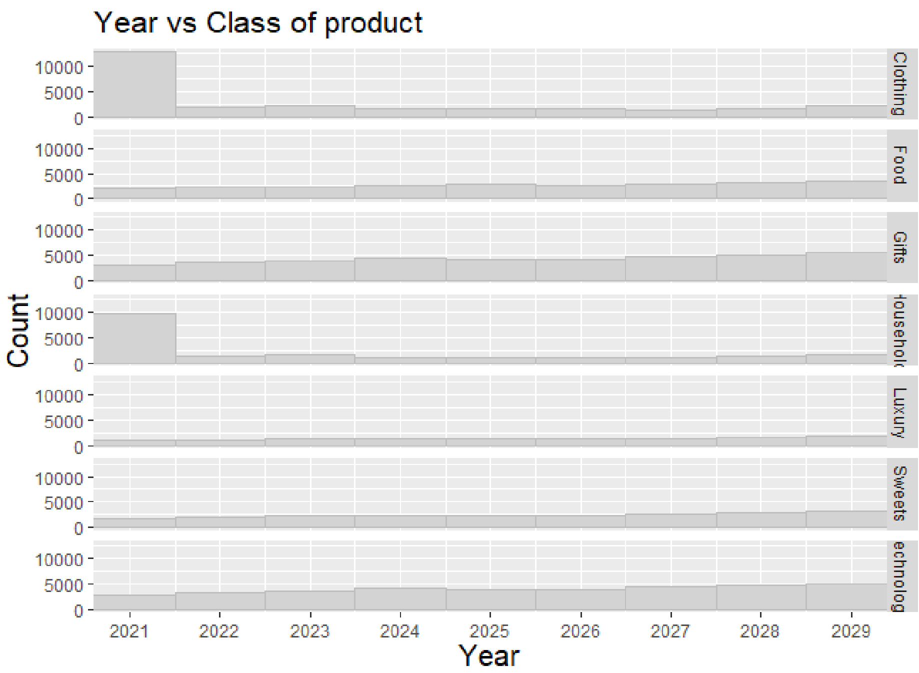


Figure 34 Year vs Class

As seen on the plot above, the demand for household products as well as clothing decreased after 2021. It is the sales department's responsibility to investigate this decrease and consider the reasons for this. It could be that the reliability decreased because of a decrease in the service and quality. The sales are distributed over the years for the classes.

Thus, the null hypothesis is rejected and the year (dependent variable) has an influence on the classes purchased (Independent variable).

PART 6: Reliability of the service and products

Question 6.1: Problem 6: Taguchi Loss

Taguchi Loss function	$L = k(y\text{-}m)^2$

Target (t): 0.06
Deviation/tolerance (D): 0.04
Loss (scrap value) (L): 45
Constant (k) : $L/(D^2)$ = $45/(0.04^2)$

Calculate the constant:

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

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

$$\mathrm{k} = 45/(0.04)^2$$

$$\mathrm{k} = 28125$$

Calculate loss function

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

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

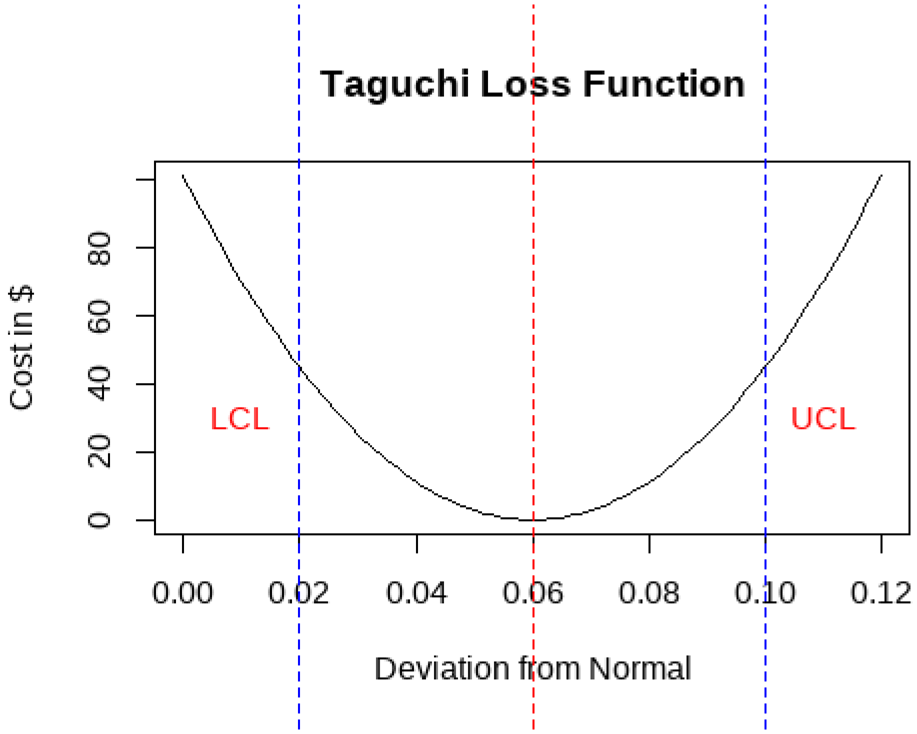


Figure 6.1: Taguchi loss function

The bigger the deviation from the target value (0.06 in this case), the worse the product's quality is. The worse the quality is, the higher the cost of the company will be when they must manufacture more products to meet the specific requirements and the more waste they will have. This can be seen on the figure above. Unreliable products with characteristics deviating from specifications will cause the service efficiency to decrease at the expense of the company.

When the thickness of the refrigerator part is within the range of the lower and upper limits, 0.02cm and 0.1cm, the customers will be satisfied.

When the thickness is less than 0.02cm or more than 0.1cm the customers will be dissatisfied. It will cost $45 per part to scrap parts that do not conform to the specifications and thus lead to dissatisfied customers.

Question 6.2: Problem 7: Taguchi Loss

a) Taguchi loss function

1. Calculate constant:

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

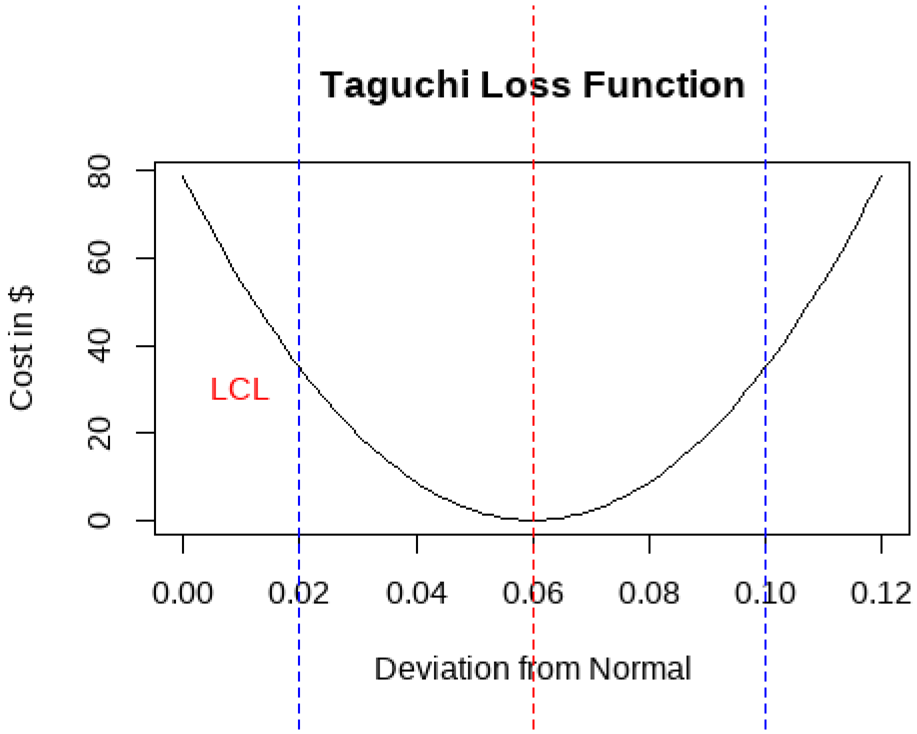
$35 = \mathrm{k}(0.04)^2$

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

2. Calculate the loss function

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

$L(x)\ 21875\ (x - 0.06\ )^2$



The bigger the deviation from the target value of 0.06, the worse the quality of the product is. Lower service efficiency and more unreliable products will result in a bigger deviation. This will increase the company's costs and the company will face larger losses.

When the thickness of the refrigerator part is within the range of the lower and upper limits, 0.02cm and 0.1cm, the customers will be satisfied.

When the thickness is less than 0.02cm or more than 0.1cm the customers will be dissatisfied. It will cost $35 (per part) to scrap parts that do not conform to the specifications and thus lead to dissatisfied customers.

b) Loss reduced to 0.027

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

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

A loss of $15.95 is made per item when the process deviation is reduced to 0.027 cm from the target. Reducing the quality of service provided by the company.

Problem 27: System Reliability

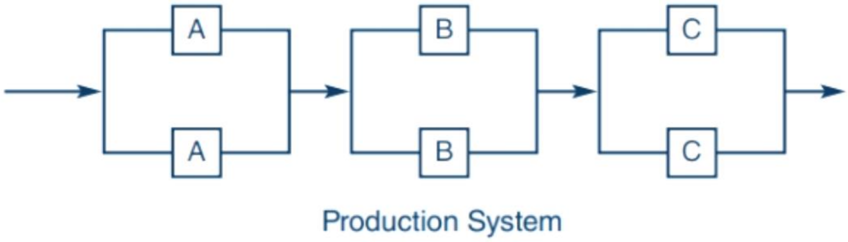


Figure 35 Reliability block diagram for Magnaplex

Machine	Reliability
A	0.85
B	0.92
C	0.90

Figure 36 Reliabilities for each machine

a) Reliability if only one machine in A, B and C is working

$\boldsymbol{Reliability}$ = $R_A*R_B*R_c$

$\boldsymbol{Reliability}$ = 0.85 * 0.92 * 0.90 = 0.7038

b) Improved reliability by using two machines per stage

When two machines of each machines A, B and C are working it will result in the first process reliability being higher.

When A components are connected parallel, the combined reliability will be:

1 – probability (both fail)

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

Repeating for B and C:

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

$R_{CC} = 1 - (1\text{-}0.9)^2 = 0.99$

The reliability then simply equates to:

$R_{AA} * R_{BB} * R_{cc}$

0.9775 * 0.9936 * 0.99

= 0.9615

The percentage improvement can then be calculated as the difference between the new and old reliability divided by the new reliability:

(0.9615 – 0.70380) / 0.9615

= 26% improvement.

Therefore, having two identical machines in parallel with each other will result in a 26% improved reliability. The reason for this is that when one of the machines breaks, the other identical machine can continue to operate during the breakdown. This will improve the reliability of the company when running the same tp machines simultaneously and would be highly recommended.

Question 6.3 Using a Binomial distribution

The required calculations needed for the vehicle and driver reliability were done by using the constants given and the dbinom() function in R

Case 1: 20 vehicles available

Results:

R(V) = ProbabilityReliableNrVehicles = 0.990

R(D) = ProbabilityReliableNrDrivers = 0.998

TotRel = R(v) * R(D) * 365 = 0.98834

TotRel = 360.7449

Case 2: 21 vehicles available

New results:

R(v) = ProbabilityReliableNrVehicles = 0.999

R(D) = ProbabilityReliableNrDrivers = 0.998

TotRel = R(v) * R(D) * 365 = 0.99788

TotRel = 364.229

Conclusion:

The addition of one extra driving vehicle will result in an additional 3.48 days available for deliveries.

Conclusion

The valid data set, obtained from cleaning and sorting the original data set of the online store, is used to gain a good understanding of the data through the construction and analysis of tables and charts. The sales of the company can be statistically analysed.

The Control charts constructed are useful in giving the state of deliveries for different classes. It can be conducted that gifts, luxury items and household products are not controlled, or stable. This is an indication that there is a serious need for investigation regarding this instability. Either the company must negotiate with its current logistic partner to obtain shorter delivery times or decide to contract a logistics partner.

The same results are obtained from the MANOVA test. The demand for clothing and household items decreased over the year 2021 and started increasing again after 2026. It can be useful for the company to take note of this decrease and slight increase and investigate whether the reason could have been because of a lack in quality and service of these products and to evaluate if trends or quality increased again to ensure an upward trend in sales.

It is less probable that a type I error is made compared to the probability that a type II error could occur. Thus, the company needs to ensure that products are delivered at the right time instead of assuming that the delivery time is accurate.

The significance of explorative analysis is brought to light and how the company can benefit from this is made understood.

Reference

Gimenez, L., 2018. data-cleaning. Retrieved from geotab.com: https://www.geotab.com/blog/data-cleaning/

Hessing, T., 2014. process-capability-cp-cpk. Retrieved from sixsigmastudyguide.com: https://sixsigmastudyguide.com/process-capability-cp-cpk/

Joseph M., 2003. Six Sigma Education and Using the Existing Quality Methods and Procedures. Science direct.

Smith B, 2020. 45 Ecommerce Statistics You Need to Know in 2019, Available at: https://www.wordstream.com/blog/ws/2019/04/04/ecommerce-statistics

Yau C., 2011. Binomial distribution. R Tutorial.