Quality Assurance 344: ECSA PROJECT

Jenika Vorster
2372184

Contents

List of Figures

List Of Tables

Introduction

If used properly, data can be a very valuable resource for a business. The insights that the data may give the business may help it increase its profits. The technique of obtaining this important information from datasets is known as data analysis (EPICOR, 2022).

This report analyses data from an online retailer with a particular emphasis on delivery timeframes. Invalid data are deleted from the dataset before it is utilized for the analysis. To identify patterns, trends, and important data relating to various aspects, descriptive statistics are generated and analysed. To see how the data varies from the predetermined control bounds and evolves over time, statistical process control is used.

Part 1: Data wrangling

Data wrangling, also known as data cleaning, data remediation, or data munging, refers to a set of methods used to convert raw data into more usable representations. The specific strategies vary based on the data you're utilizing and the aim you're attempting to achieve. Steps of data wrangling include familiarising yourself with the data, structuring/ transforming the data, data cleaning, enriching your data, validating and publishing of data (Stobierski, 2021).

Valid data

Valid data should be separated from invalid data. Invalid data includes data that contain missing values. There are 180000 instances in total. There are 17 instances where there is "not a number" in the data. These should be removed.

Only the first 29 instances are shown of the 179983 instances, since it is such a large data set (exported from Excel).

t	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	Recommended
5	5	92178	76	Household	19215.01	2027	11	26	61.5	Recommended
6	6	50586	78	Gifts	4929.82	2027	4	24	14.5	Random
7	7	73419	35	Luxury	108953.5	2029	11	13	4	Recommended
8	8	32624	58	Sweets	389.62	2025	7	2	2	Recommended
9	9	51401	82	Gifts	3312.11	2025	12	18	12	Recommended
10	10	96430	24	Sweets	176.52	2027	11	4	3	Recommended
11	11	87530	33	Technology	8515.63	2026	7	15	21	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	Browsing
14	14	77795	92	Food	556.83	2025	6	3	3	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	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	Recommended
20	20	99187	54	Luxury	81620.21	2027	9	14	3	Recommended
21	21	59365	72	Gifts	3314.76	2028	4	30	13	Recommended
22	22	37221	24	Sweets	220.91	2021	3	8	3	Recommended
23	23	78120	23	Gifts	2378.31	2023	3	10	12	Recommended
24	24	65860	30	Gifts	2440.41	2021	5	11	9.5	Recommended
25	25	70953	70	Gifts	3962.67	2024	10	6	12.5	Recommended
26	26	58327	45	Luxury	83248.5	2027	1	2	4.5	Recommended
27	27	39049	60	Luxury	26681.03	2029	6	18	2	Recommended
28	28	16931	28	Technology	47135.28	2025	5	5	18.5	Browsing
29	29	74173	56	Technology	8370.39	2026	9	3	19.5	Recommended

Table 1: Valid data

Invalid data

The 17 instances of the invalid data are shown of the 179983 instances, since it is such a large data set (exported from Excel). It starts at 12345. The different features are shown like age, class, price, etc.

r	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	12345	18973	93	Gifts		2026	6	11	15.5	Website
2	16321	81959	43	Technology		2029	9	6	22	Recommended
3	19541	71169	42	Technology		2025	1	19	20.5	Recommended
4	19999	67228	89	Gifts		2026	2	4	15	Recommended
5	23456	88622	71	Food		2027	4	18	2.5	Random
6	34567	18748	48	Clothing		2021	4	9	8	Recommended
7	45678	89095	65	Sweets		2029	11	6	2	Recommended
8	54321	62209	34	Clothing		2021	3	24	9.5	Recommended
9	56789	63849	51	Gifts		2024	5	3	10.5	Website
10	65432	51904	31	Gifts		2027	7	24	14.5	Recommended
11	76543	79732	71	Food		2028	9	24	2.5	Recommended
12	87654	40983	33	Food		2024	8	27	2	Recommended
13	98765	64288	25	Clothing		2021	1	24	8.5	Browsing
14	144444	70761	70	Food		2027	9	28	2.5	Recommended
15	155555	33583	56	Gifts		2022	12	9	10	Recommended
16	166666	60188	37	Technology		2024	10	9	21.5	Website
17	177777	68698	30	Food		2023	8	14	2.5	Recommended

Table 2: Invalid data

Part 2: Descriptive statistics

Data is classified into several classes, which dictate which forms of mapping may be utilized for it. The most fundamental distinction is between continuous (or quantitative) data and categorical data, which has a significant influence on the sorts of visualizations that may be utilized.

Continuous features

Continuous data is information that can have almost any value. This covers height, weight, and any other numerical measurement. The type of information that generates continuous data is also likely to alter throughout time (iSixSigma, 2022).

Categorical features

Categorical data is statistical information that is provided based on its classification. According to the analysts' design, values are classified into predetermined categories in this model. This method of categorizing data points can be beneficial depending on the aims of the research, but it is simply one of several ways to organize statistical information (iSixSigma, 2022). In table 3, categorical features are illustrated.

Feature	Length	Missing values	Mode	Mode Frequency	Mode %	Cardinality
Class	179983	0	Gifts	39149	25.45	7
Why Bought?	179983	0	Recommended	106985	59.44	6

Table 3: Categorical features

Data instances that included NA (not available) values were removed. It was also then not used in any calculations. The invalid data can be because of mistakes in management or accidents but should be investigated.

Graphs

Delivery time vs Class of the product

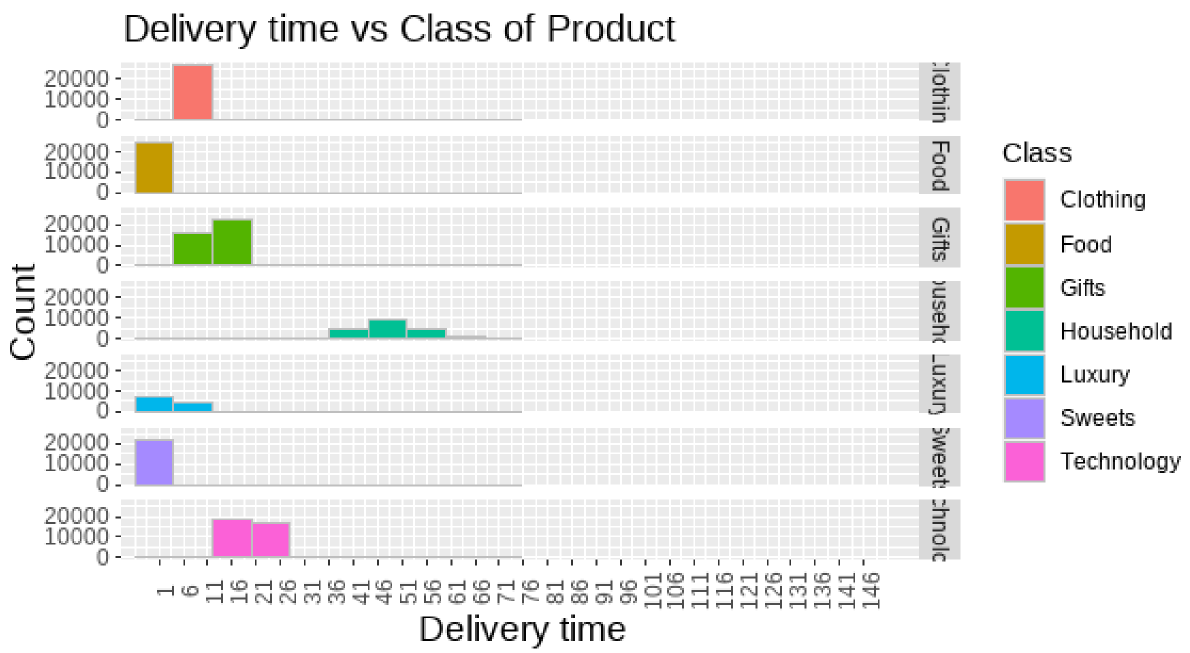


Figure 1: Delivery time vs class

The graph shows that food, luxury items, and sweets are delivered the fastest while households took the longest, since the distribution is further along the x-axis. The corporation needs to figure out why deliveries to families are taking so long. Household items are dispersed throughout delivery windows, maybe because they include both larger and heavier (which takes more time) and little things (more practical and faster to travel).

Food is often delivered quickly since they are portable, easily made, and quick to transport. Household goods typically require a while and more personnel to load onto the truck and transport to the customer's location, which may account for the lengthy delivery times.

Sale count per year

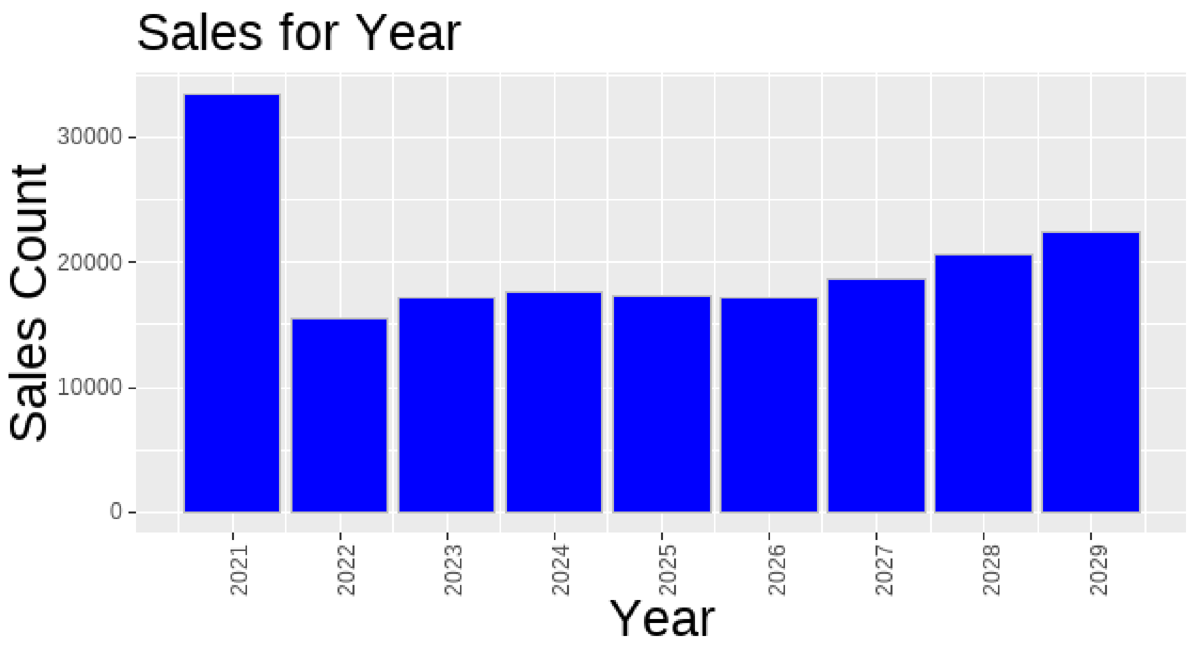


Figure 2: Sales for year

In figure 3, it is shown that there were the most sales in 2021. There is a positive trend from 2022 and onwards from 2023 to 2029. Products costs up to R320000 (max), but the average lies around R150000. Year 2022-2029 is evenly distributed.

Month count (sales)

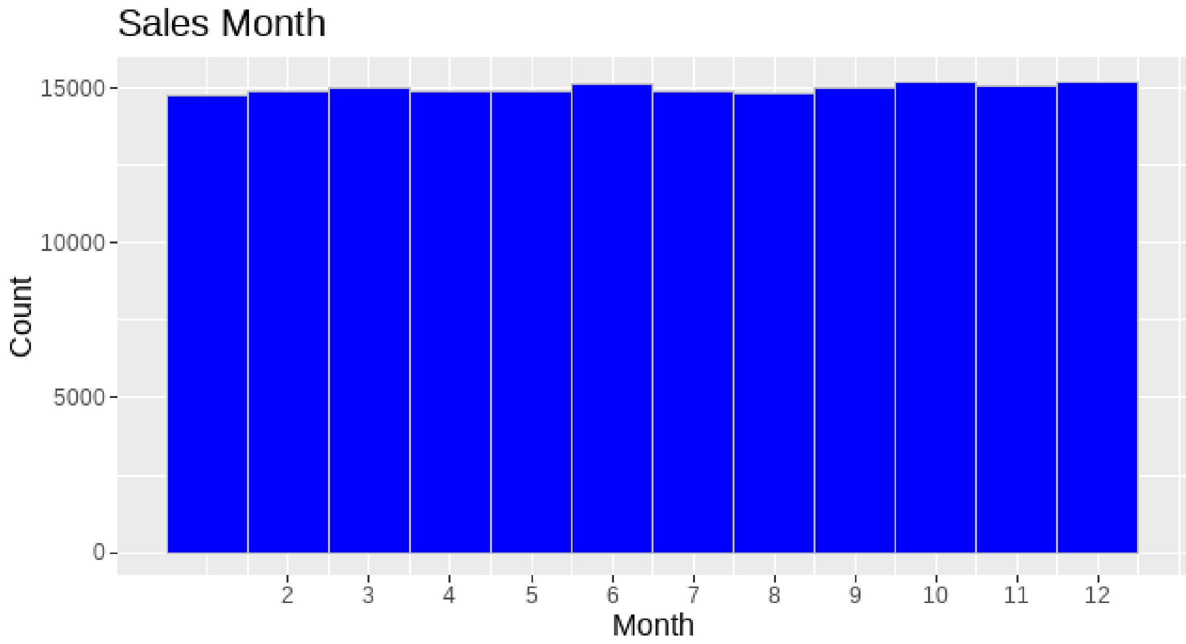


Figure 3: Sales for month

There is a uniform distribution on figure 4 that shows the sales for a month. This does not give plenty of useful information.

Day count (sales)

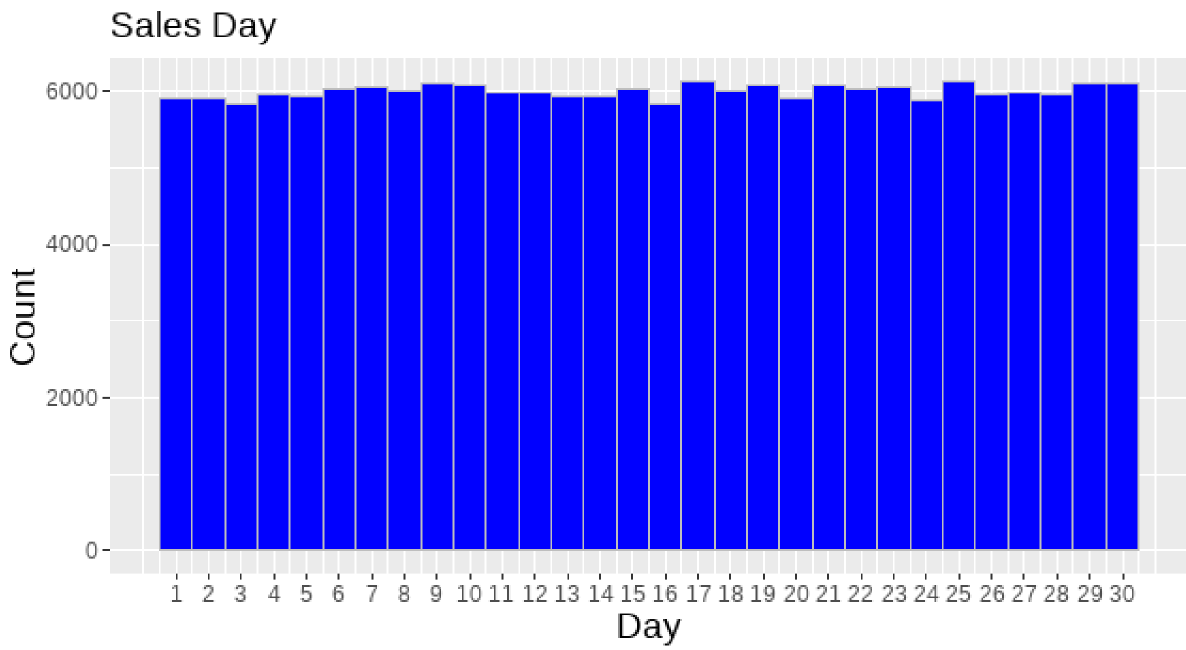


Figure 4: Sales per day

Each day's sales total is allocated equally, in other words, it in normally distributed. Each day's sales are almost the same. At the start of each month, the minimum sale is made. For each day of sales, it is impossible to detect any trends.

Age count (sales)

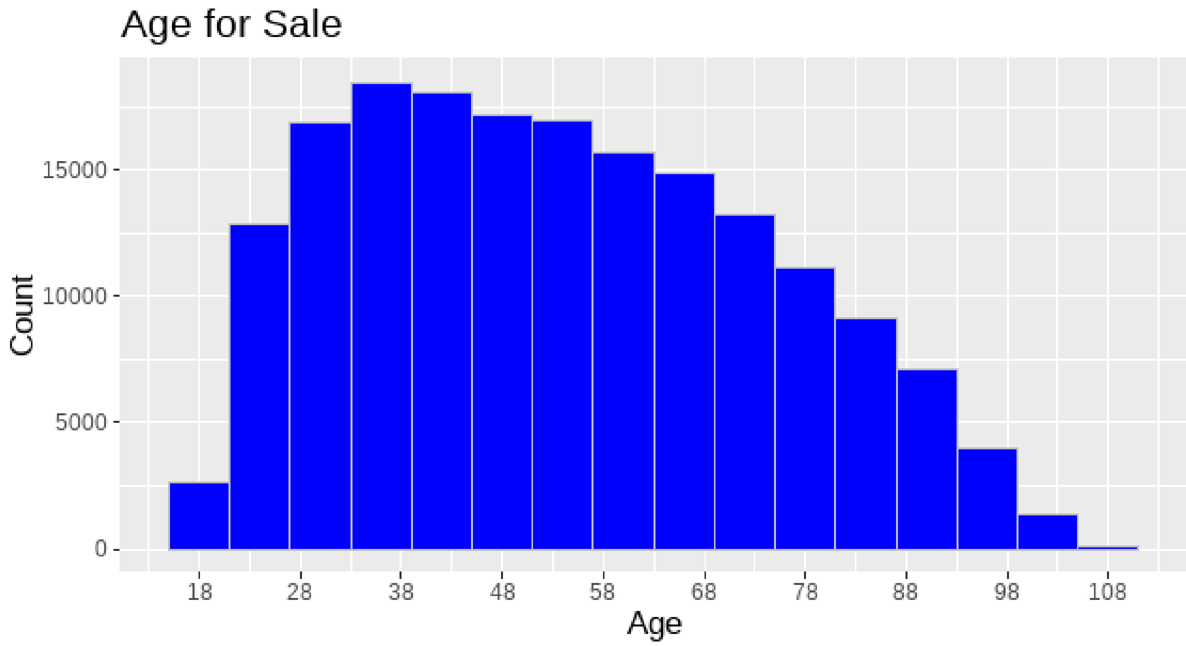


Figure 5: Age Count

It is a unimodal distribution (right tailed). Skew is just a tendency towards extremely low (right skewed) numbers. The age range between 32 and 38 is the most common. According to this graph, younger individuals are more likely than older people to use the internet platform. The sales management team must devise strategies for marketing to young people.

Delivery time count

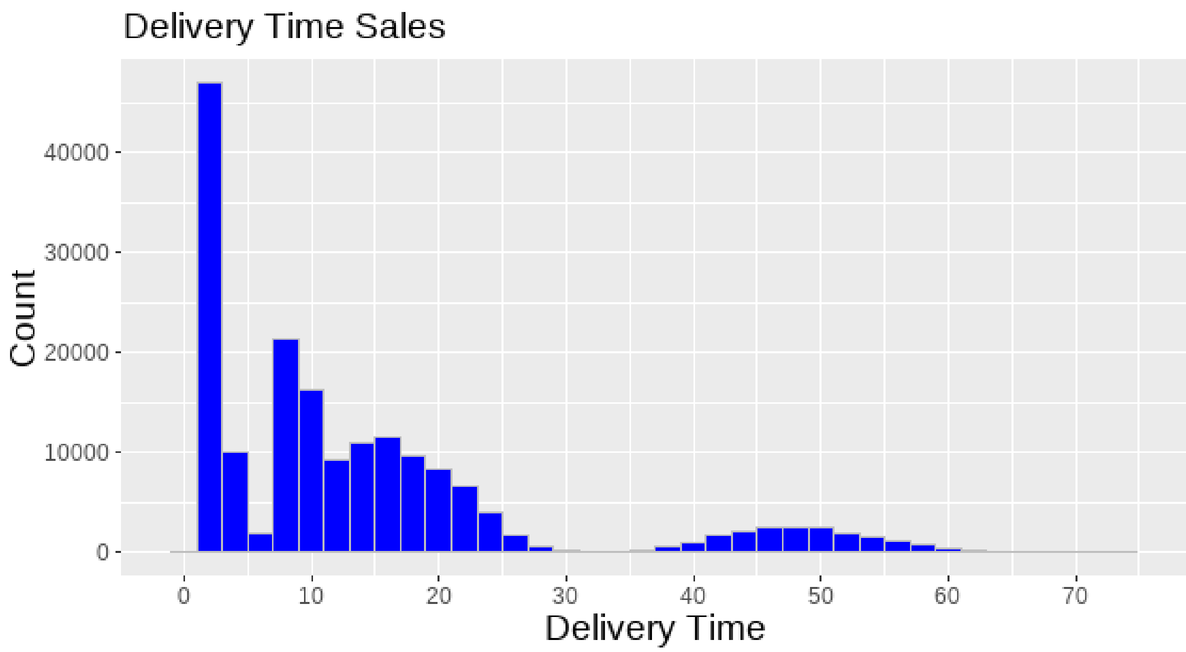


Figure 6: Delivery time

This figure 7 is quite normally distributed between 40 and 60 days. It has a mean of 48 days and the most frequent delivery times are 2-4 days. Delivery times of 2-4 days could be for everyday items like household items and food. The maximum lies around 46500.

Why bought count

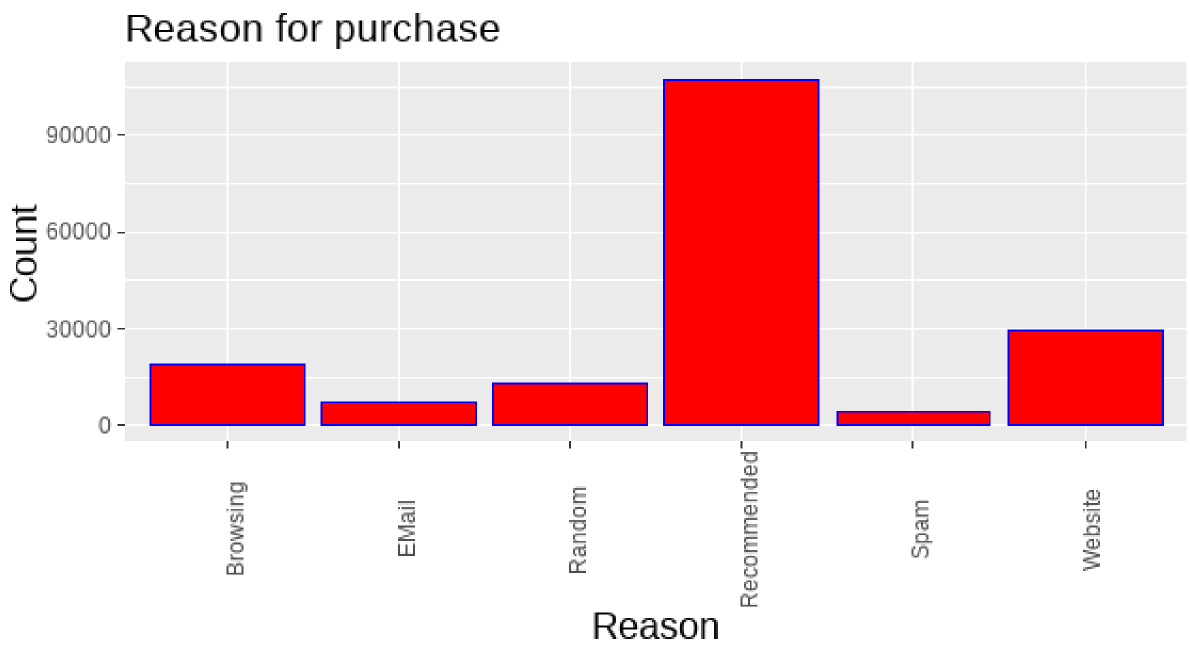


Figure 7: Reason for purchase

The most frequent reason for purchasing a product is due to a recommendation from another person. Spam is the least common reason for someone to purchase a product. Customers must receive high-quality service in order to increase word-of-mouth marketing of the items.

Class of the product

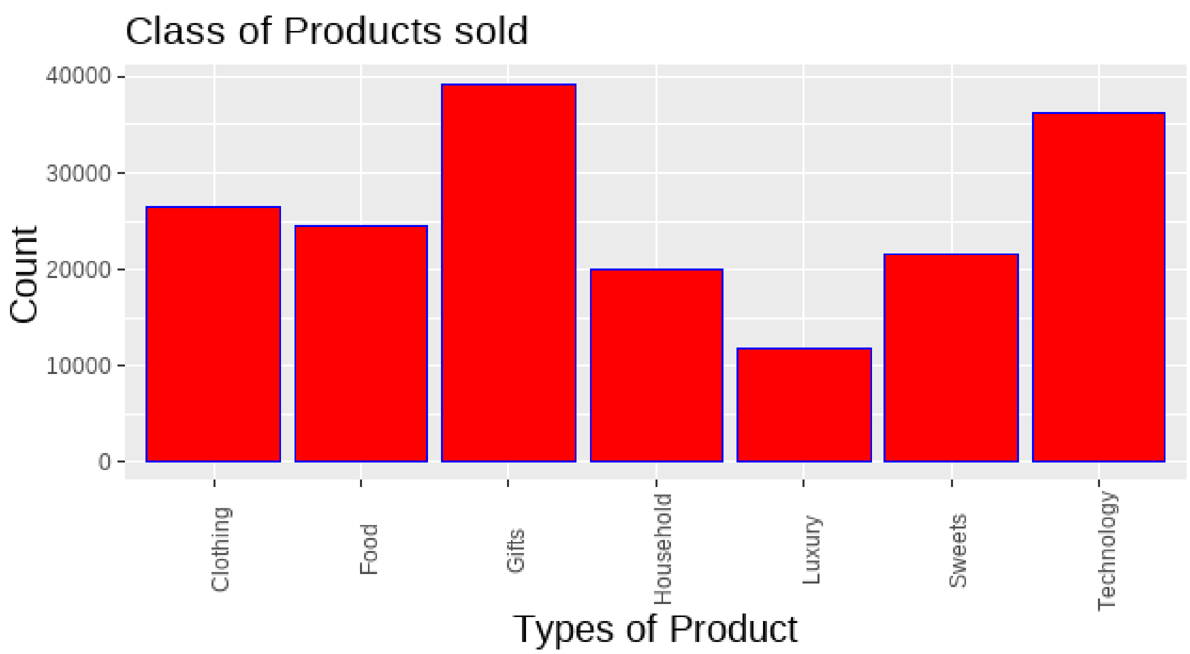


Figure 8: Class of product

Gifts are the most often purchased item, followed by technology. Luxury goods are the least often purchased things. Since gifts and technology are the most popular items purchased by consumers, they must be of the highest quality.

Age vs why bought

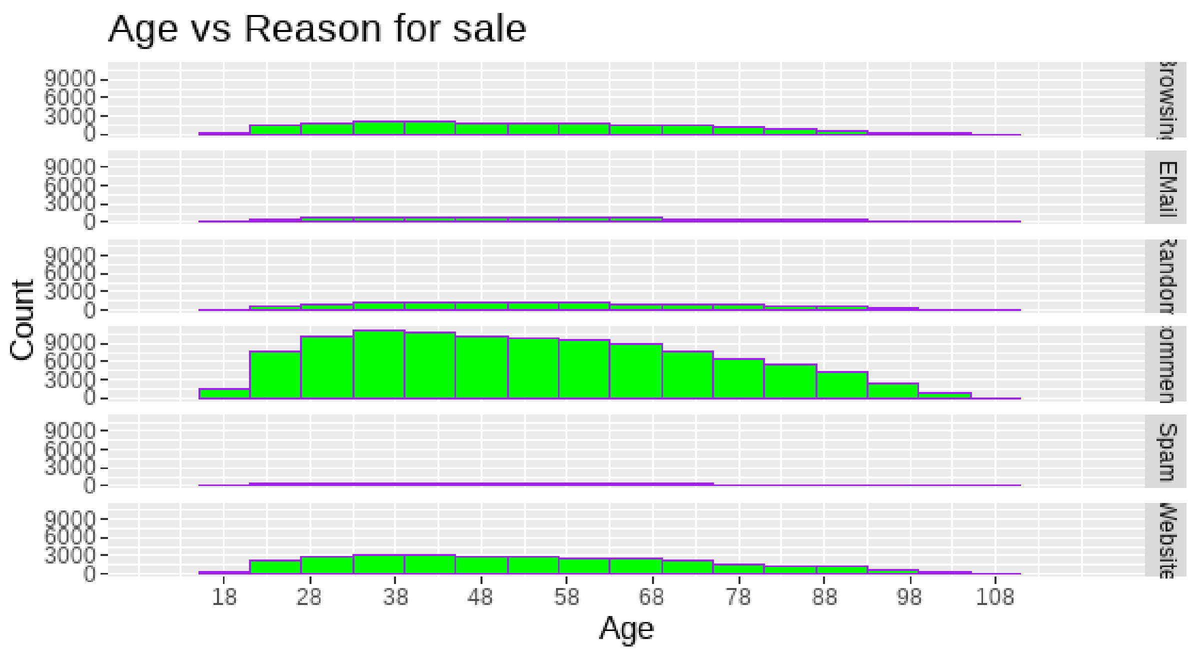


Figure 9: Age vs why bought

The suggested reason for purchasing a product is unimodal (skewed right). It seems that all ages use the same justifications for purchasing products.

*Right hand side order of labels: Browsing, email, random, recommended, spam, website (got cut off).

Month vs class

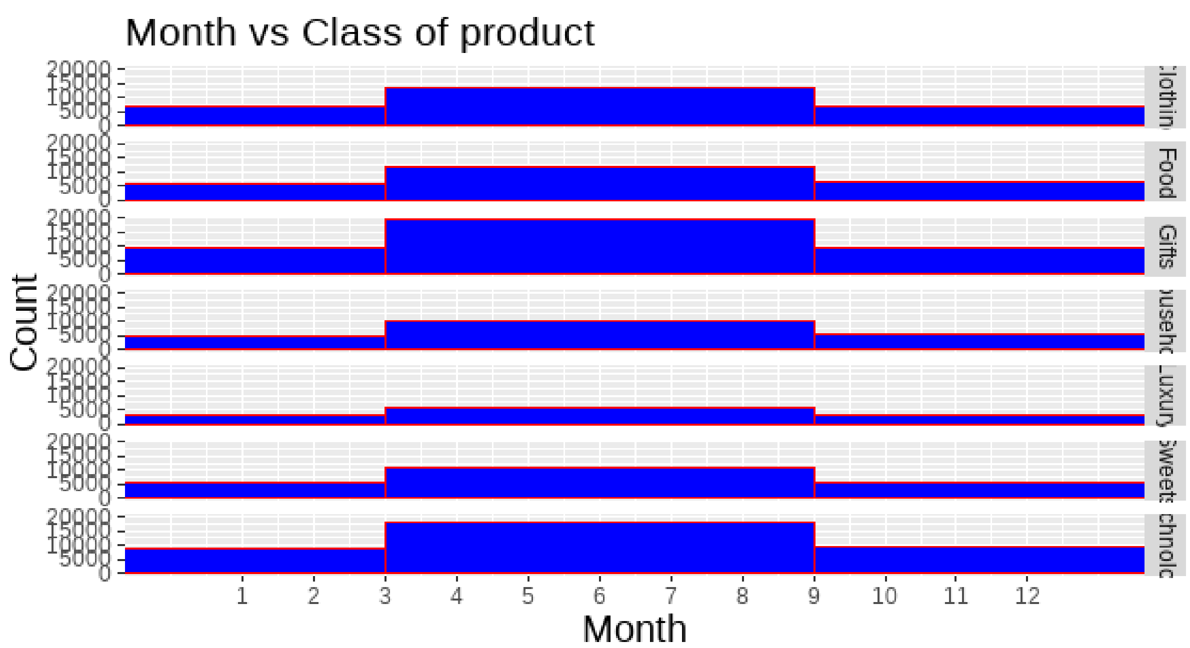


Figure 10: Month vs Class

The majority of the items are purchased between March and August according to the various distributes. Between Autumn and Winter, a seasonal pattern has been discovered. The sales team has to determine why people purchase more goods during cold weather.

*Right hand side order of labels: Clothing, food, gifts, household, luxury, sweets, technology

Age vs class

Figure 11: Age vs class

All classes include ages of 98 years and above. It is assumed that retirement communities use internet marketplaces to buy goods. The age range between 28 and 33 is the most common for purchasing clothing, which has an exponential distribution. This makes logical given that younger individuals are more likely to purchase fashionable clothing.

The most common age range for purchasing food is between 48 and 80 years old. People over 65 are more likely to buy groceries online. Gifts are widely spread across all age groups, which makes sense given that they are a very popular thing to purchase for anyone.

The distribution of households is exponential, with the average age being 30.

Technology is unimodal (right tailed), with the most common age range falling between 28 and 38. Sales managers should look for techniques to promote technology to the younger age group as it has the youngest mean age group among the other classes.

Conclusion: Younger and middle-aged adults are more likely to purchase apparel, homes, luxury goods, and technology.

*Right hand side order of labels: Clothing, food, gifts, household, luxury, sweets, technology

Price vs class

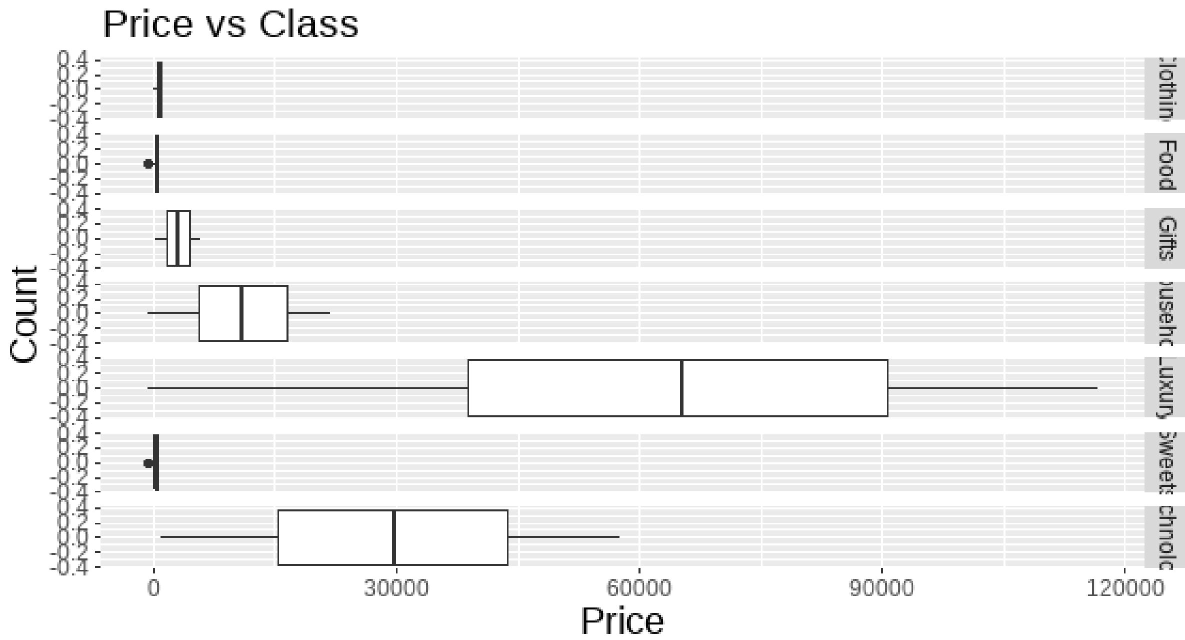


Figure 12: price vs class

A conclusion drawn from the boxplots is that luxury goods are the most expensive, followed by technology. Clothing, food, and sweets are the least expensive goods. Given that luxury and technological goods are more expensive than food and sweets, this makes sense. Since luxury goods generate the largest income among all classes, emphasis should be put on advertising them. Additionally, the price of luxury goods is more evenly spread, which suggests that their prices vary considerably more than those of other goods.

*Right hand side order of labels: Clothing, food, gifts, household, luxury, sweets, technology

Why bought vs delivery time

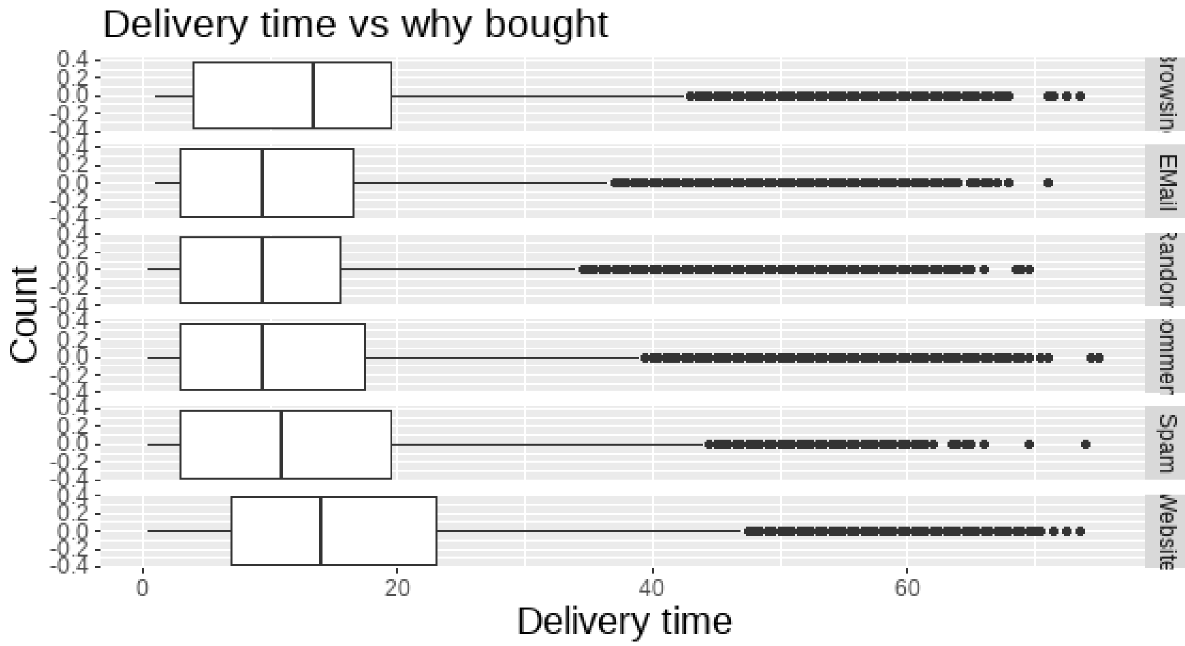


Figure 13: why bought vs delivery time

The bulk of box plots cross over one another, however internet purchases take longer to arrive than other types of purchases. The box plots' right side contains an indication of random fluctuation.

Correlation plots

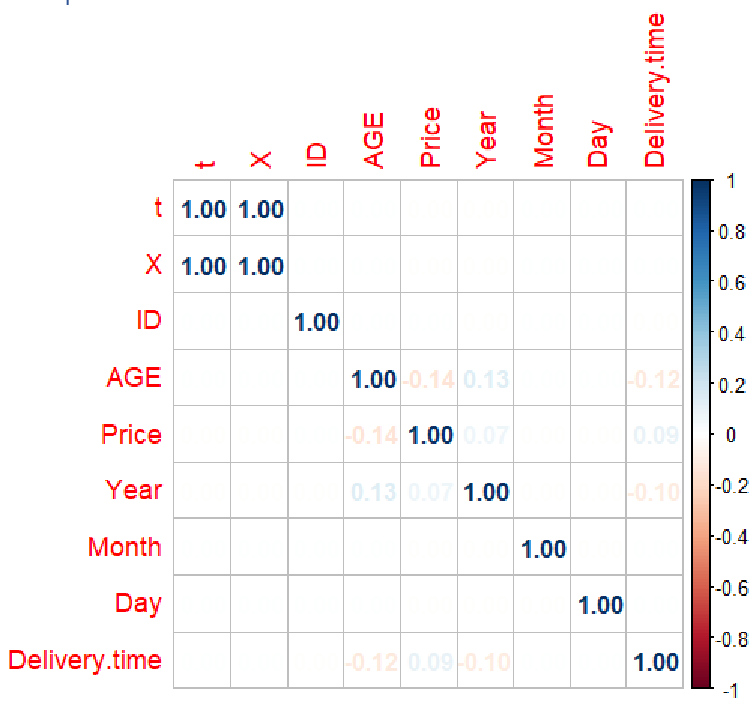


Figure 14: Correlation plot of features

On the figure above, it can be seen that where the “1s” are, the features are just compared to itself, and thus the correlation of 1. On the plot, there is a positive correlation between the features year and age, year and price & price and delivery time. There are negative correlations between price and age, delivery time and age & year and delivery time.

Process capabilities indices

Figure 15: Distribution of delivery time of clothing

So calculated from using a USL = 24 (hours) and a LSL = 0 (Since all of the data are integers with the lowest value beginning at 0, the LSL of 0 is reasonable. As a result, the Lowest beginning limit must be equal to zero.

CP= 1.142207
CPU= 0.3796933
CPL= 1.90472
CPK= 0.3796933

Table 4: Process capabilities values

Potential capability

Since CP >1, the process is capable. Since CPK<CP, it means that the process is not completely centered between the 2 limits. The improving of the process can include moving the mean to the left (NIST, 2022).

Part 3: Statistical Process Control

An X&s chart showing delivery times is created with 30 samples, each with 15 Sales. The data must first be arranged chronologically before being used to compute and build the charts. The year, month, and day are used to sort the data from oldest to newest. (Hessing, 2022).

Control charts are designed to divide process variation into common and special causes, in order for these factors to be addressed differently. The initial 30 samples are used to establish the limit control. The control charts indicate when a process should be left alone and when it should be monitored or when it should be tweaked. Control charts, when implemented effectively, may assist drive process improvement.

This was filled in manually from the values gotten on the graphs shown below.

X-Bar chart

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	**22.9731**	43.9367	42.2813	**20.3744**	38.981	37.3329	**17.7768**
Clothing	**9.4046**	18.5012	18.2204	**8.97**	17.5874	17.379	**8.5353**
Household	**50.246**	97.876	95.5009	**46.5622**	90.824	88.444	**42.2462**
Luxury	**5.49**	10.433	9.9546	**4.7355**	9.9555	8.4876	**3.9776**
Food	**2.709**	5.2591	5.11914	**2.49**	4.8393	4.622	**2.27067**
Gifts	**9.4879**	18.267	17.4863	**8.3611**	16.0082	15.2691	**7.234**
Sweets	**2.897**	5.491719	5.2219	**2.4778**	4.6825	4.4147	**2.0585**

Table 5: S chart

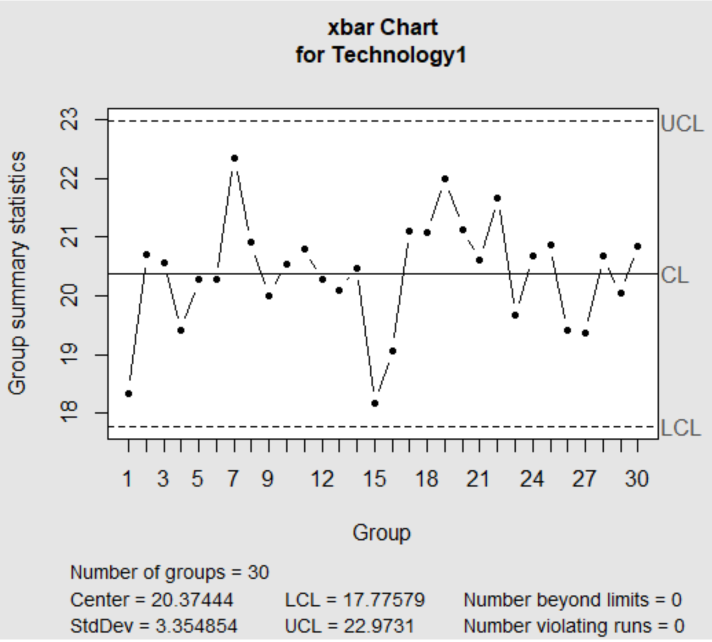
S-chart

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	**5.1799**	8.876333	7.67869	**3.2955**	5.2916	4.09817	**1.4111**
Clothing	**0.8664**	1.511667	1.300333	**0.5512**	0.9797	0.6544	**0.2360**
Household	**7.344**	12.127067	10.93133	**4.67**	7.533	5.8667	**2.00**
Luxury	**1.5108**	2.593333	2.24667	**0.96122**	1.5133	1.198267	**0.41159**
Food	**0.437**	0.75363	0.4495	**0.278**	0.43075	0.34679	**0.119**
Gifts	**2.246**	3.977633	3.43267	**1.4289**	2.3933	1.837167	**0.6118**
Sweets	**0.8352**	1.457333	1.2667	**0.5313**	0.8433	0.670767	**0.2275**

Table 6: X-Bar chart

Graphs from first 30 samples:

Technology



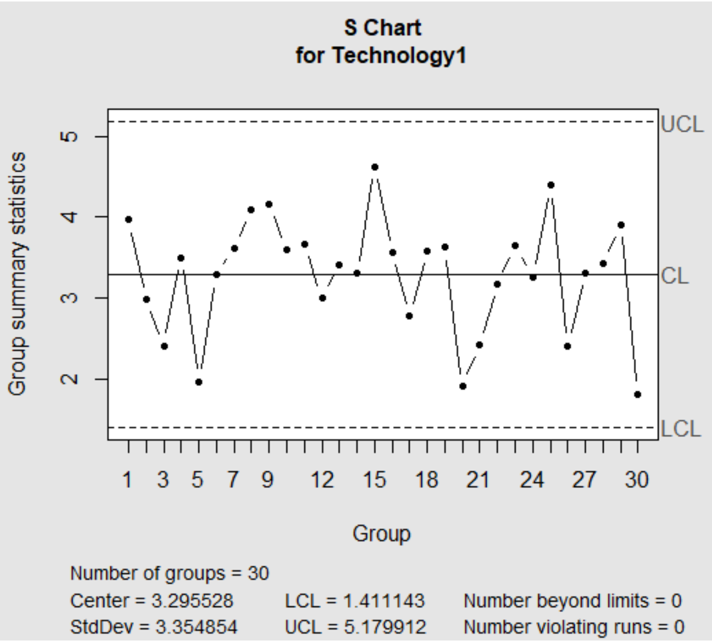


Figure 16: Xbar and S chart for Technology

The first 30 samples show that there is no variance in the technology order process and that the technology class is in charge. Due to the favourable S-chart, the X-bar chart may be examined.

Sweets

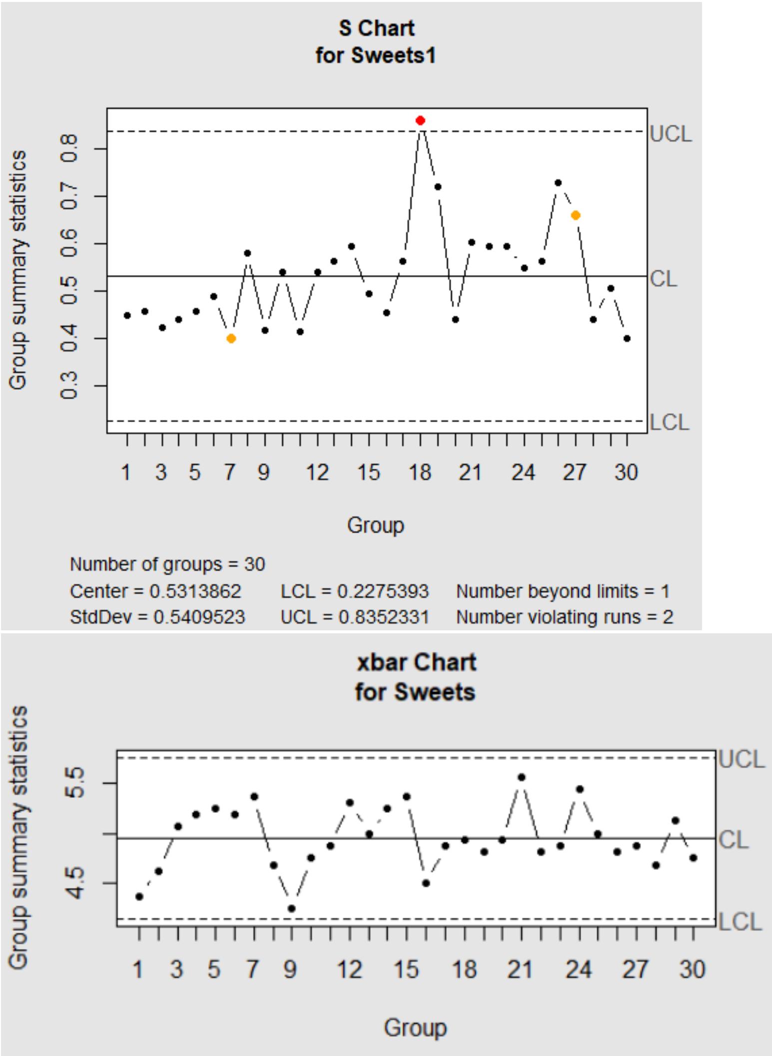
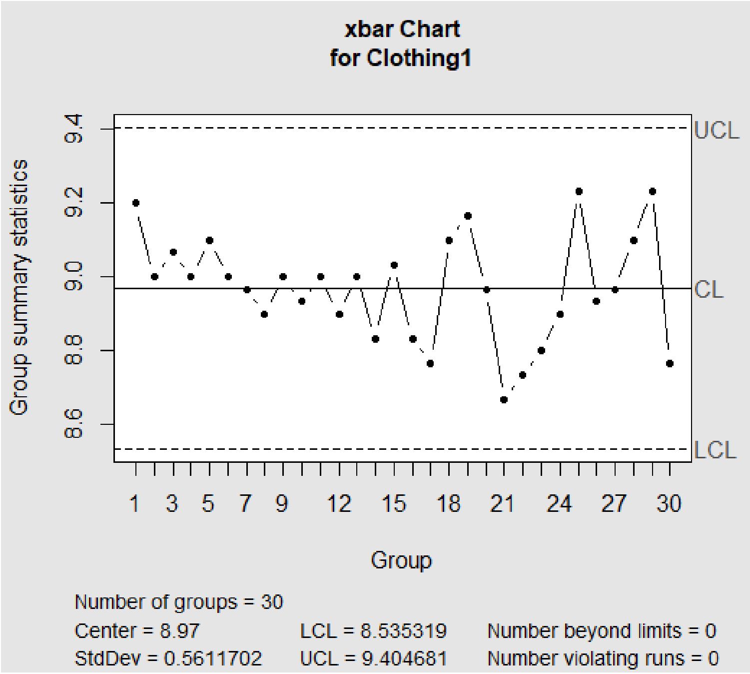


Figure 17: Xbar and S chart for Sweets

With the exception of sample 17, whose standard deviation is outside the acceptable range (exceeding the UCL), the first 30 samples show that the sweets class is under control. This suggests that sample 17 must be eliminated.

Clothing



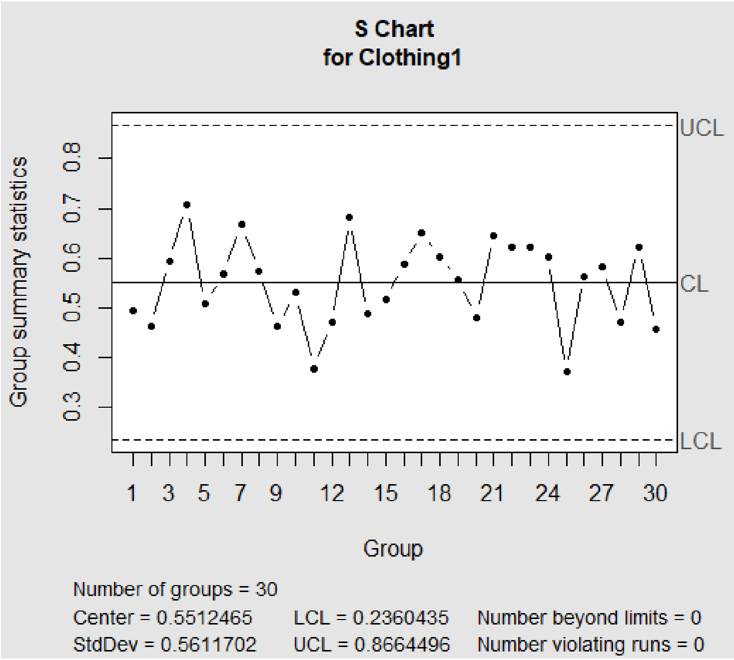
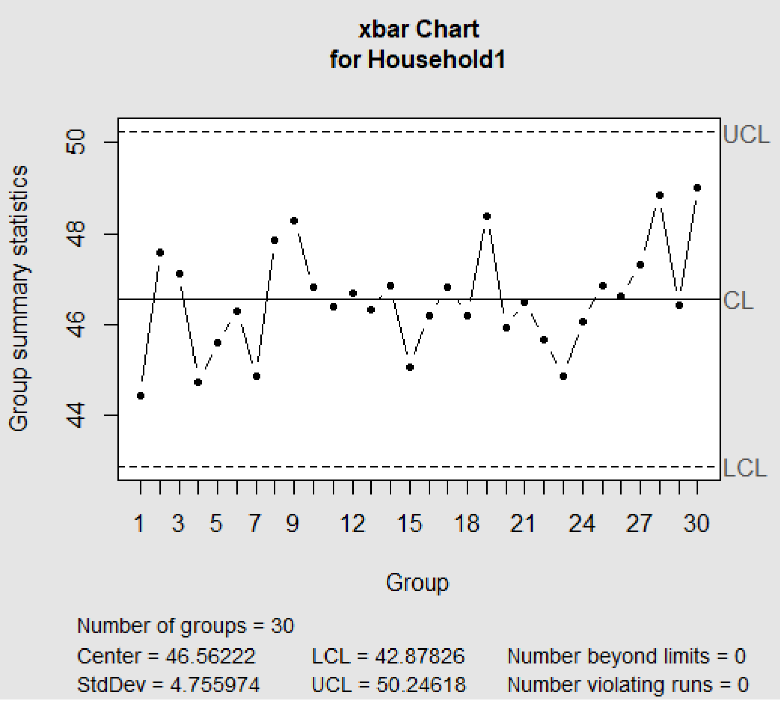


Figure 18:Xbar and S chart for Clothing

The first 30 examples show that the clothes class is in control and that there is no need for difference in the way clothing orders are processed. Due to the favourable S-chart, the X-bar chart may be examined.

Household



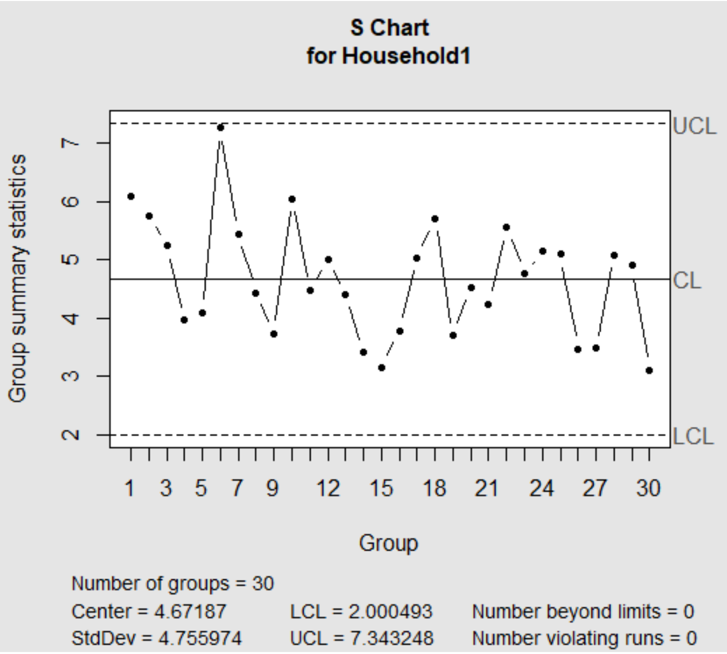
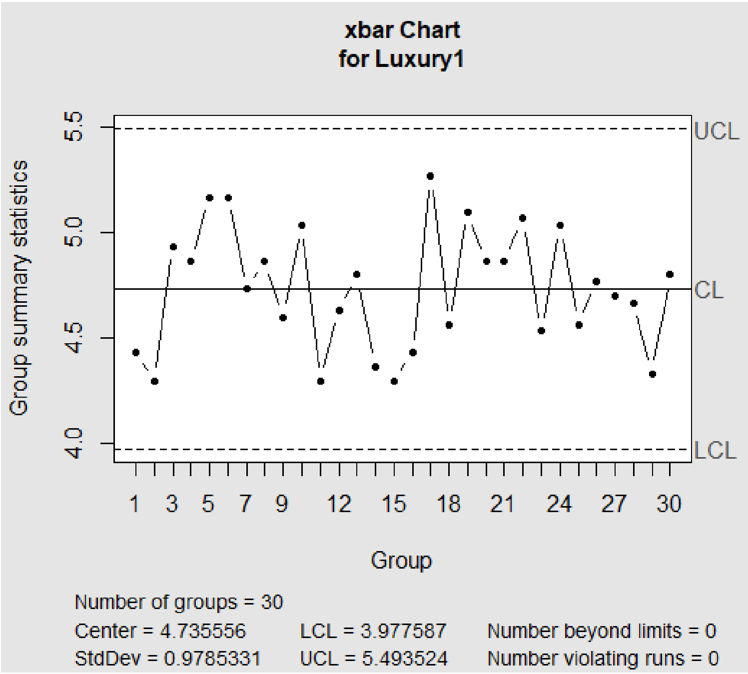


Figure 19: Xbar and S chart for Households

The first 30 examples show that the household charts are in control and that there is no need for difference in the way household orders are processed. Due to the favourable S-chart, the X-bar chart may be examined.

Luxury



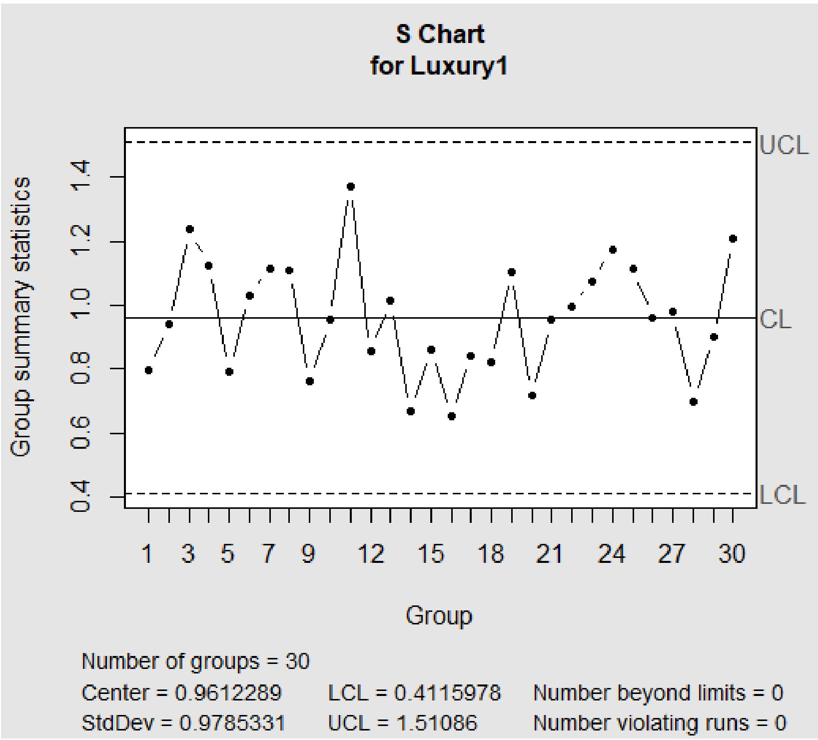
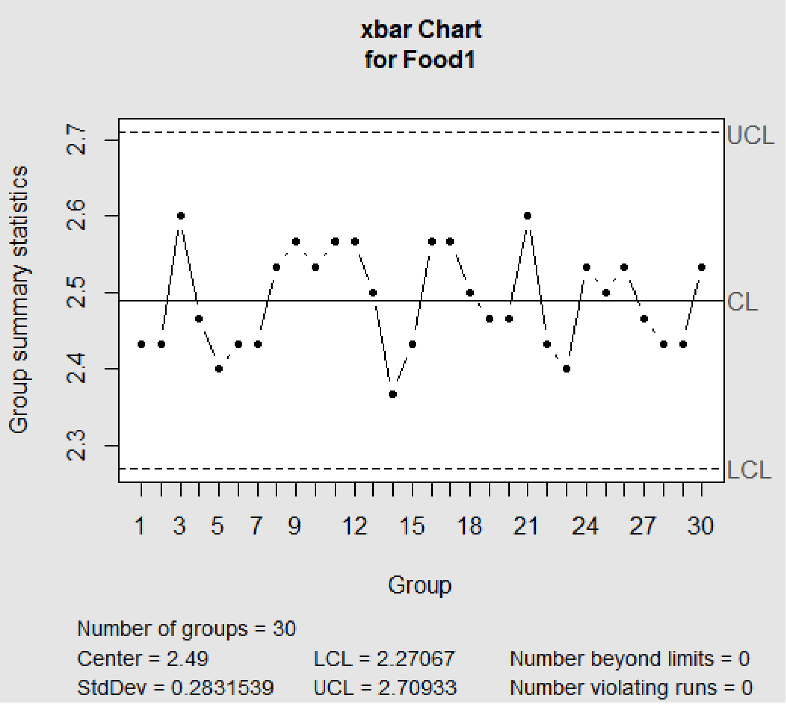


Figure 20: Xbar and S chart for Luxury

The first 30 examples show that the luxury class are in control and that there is no need for difference in the way luxury items are ordered. Due to the favourable Sbar-chart, the X-bar chart may be examined.

Food



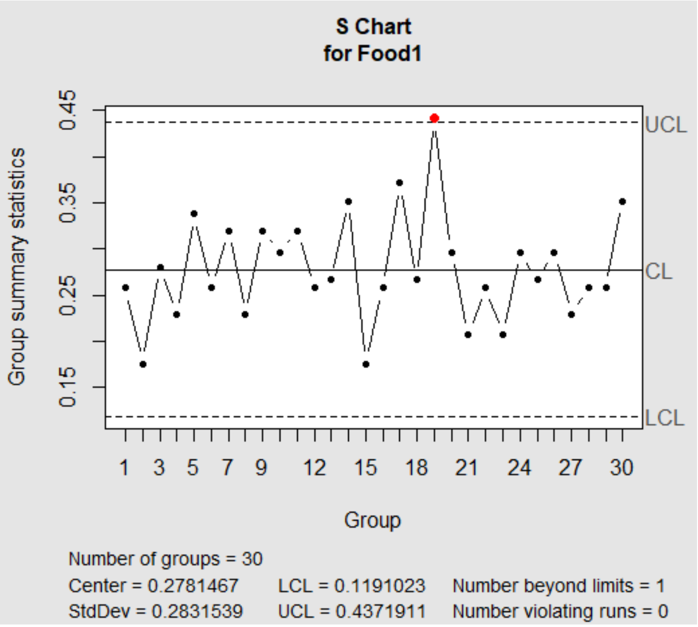
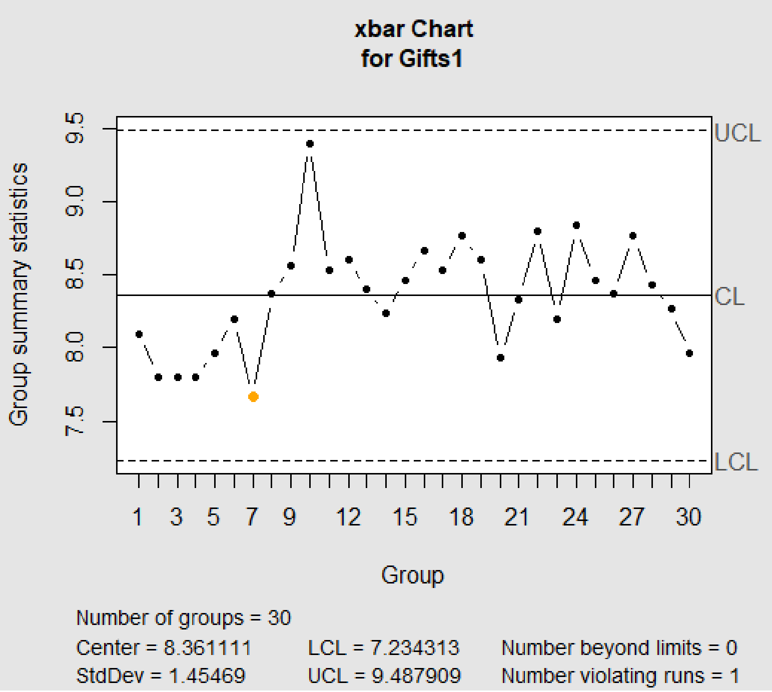


Figure 21: Xbar and S chart for Food

The first 30 examples show that the food class is in control and that there is no need for difference in the way food are ordered. Due to the favourable S-chart, the X-bar chart may be examined. Sample 19 could be eliminated since it is out of the control limits.

Gifts



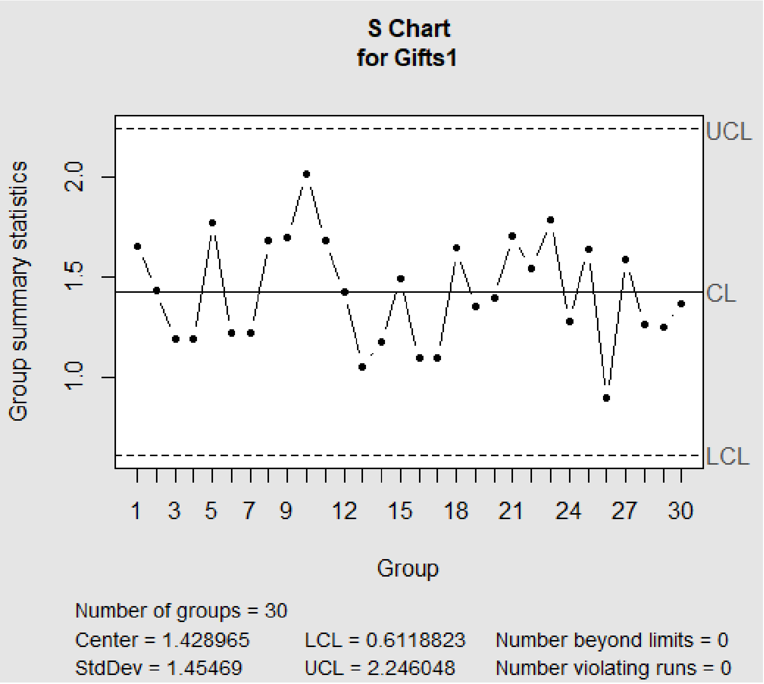
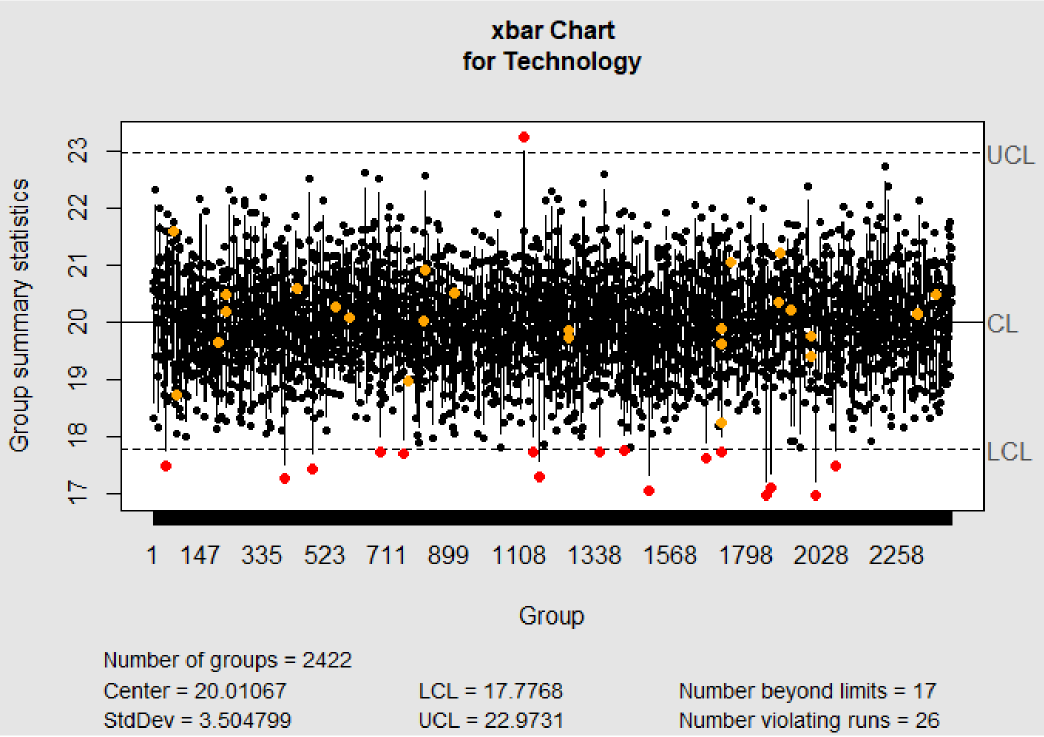


Figure 22: Xbar and S chart for Gifts

The first 30 examples show that the gifts class is in control and that there is no need for difference in the way gift orders are processed. Due to the favourable S-chart, the X-bar chart may be examined.

Graphs for all samples of few products

Technology



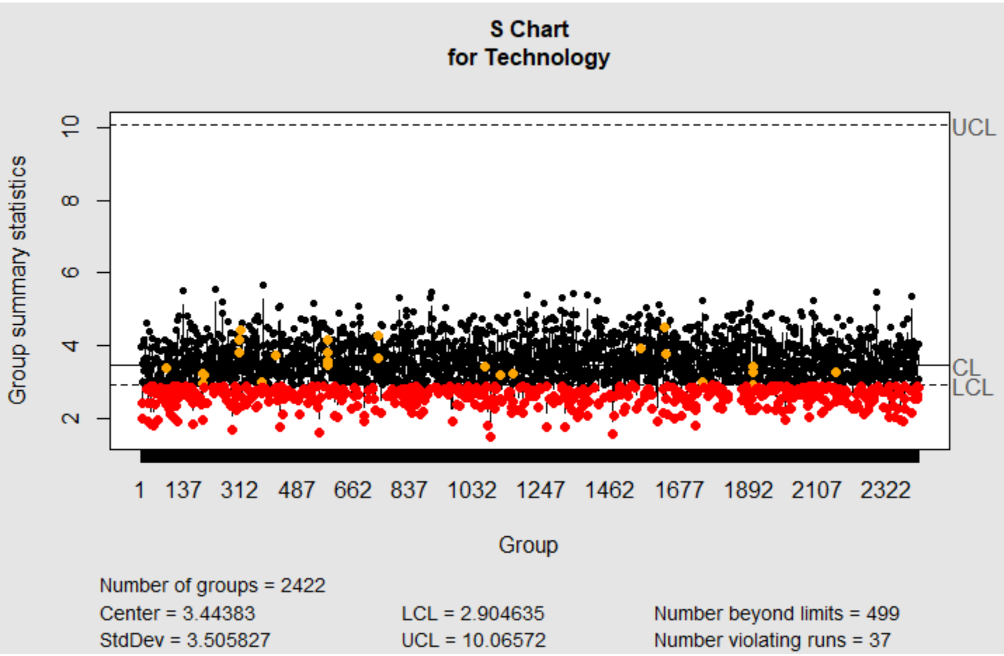
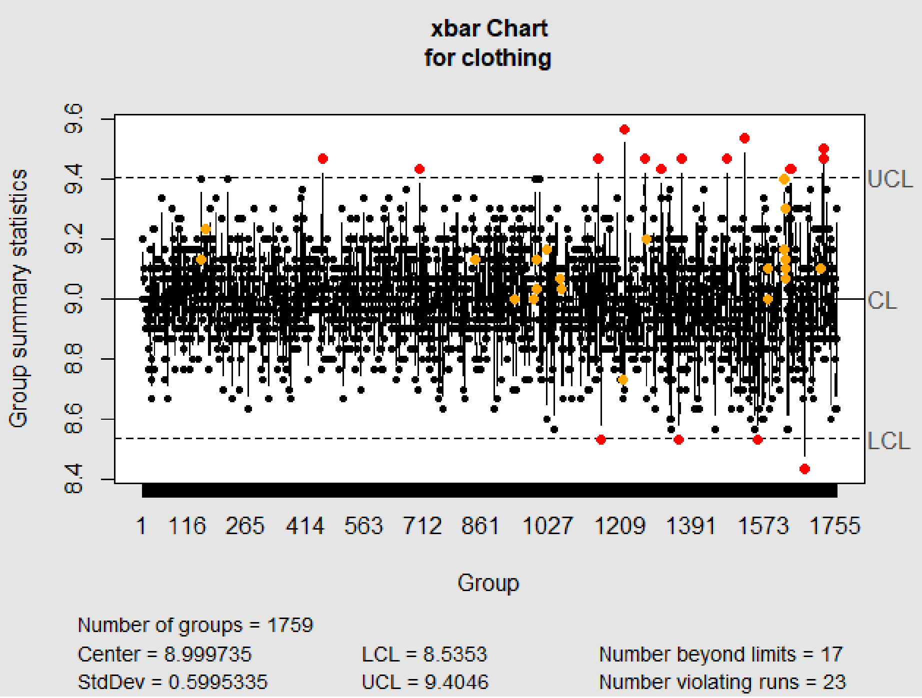


Figure 23: Xbar chart and S-chart for all samples of Technology

Most samples fall within the control limits. Technology appears to be in check. The S-bar chart is under control (there are only 499 samples that are outside of control limits), hence the X-bar chart's conclusion is valid.

Clothing



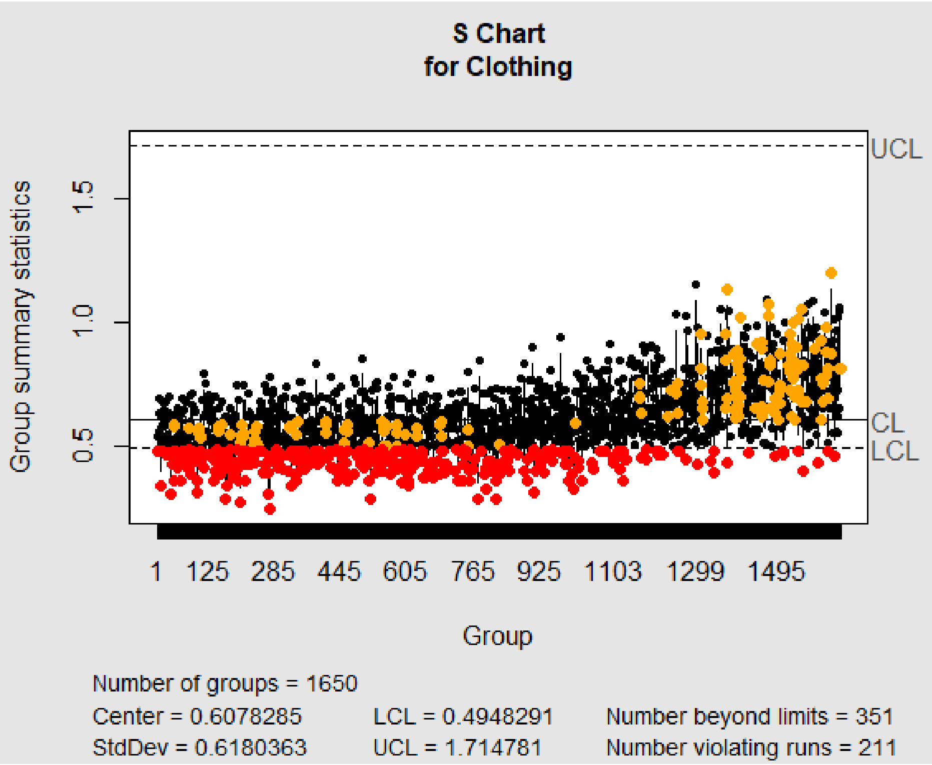


Figure 24: xbar and S Chart for clothing

The vast majority of samples are inside the control limits. Clothing appears to be under control, however there are occasional unusual instances when samples exceed the restrictions. This might be related to seasonal fluctuations. There are some samples that are beyond the S-bar chart's control boundaries, but removing those samples yields the same findings as the S-bar chart.

As a result, the conclusion of the X-bar chart is appropriate.

Luxury

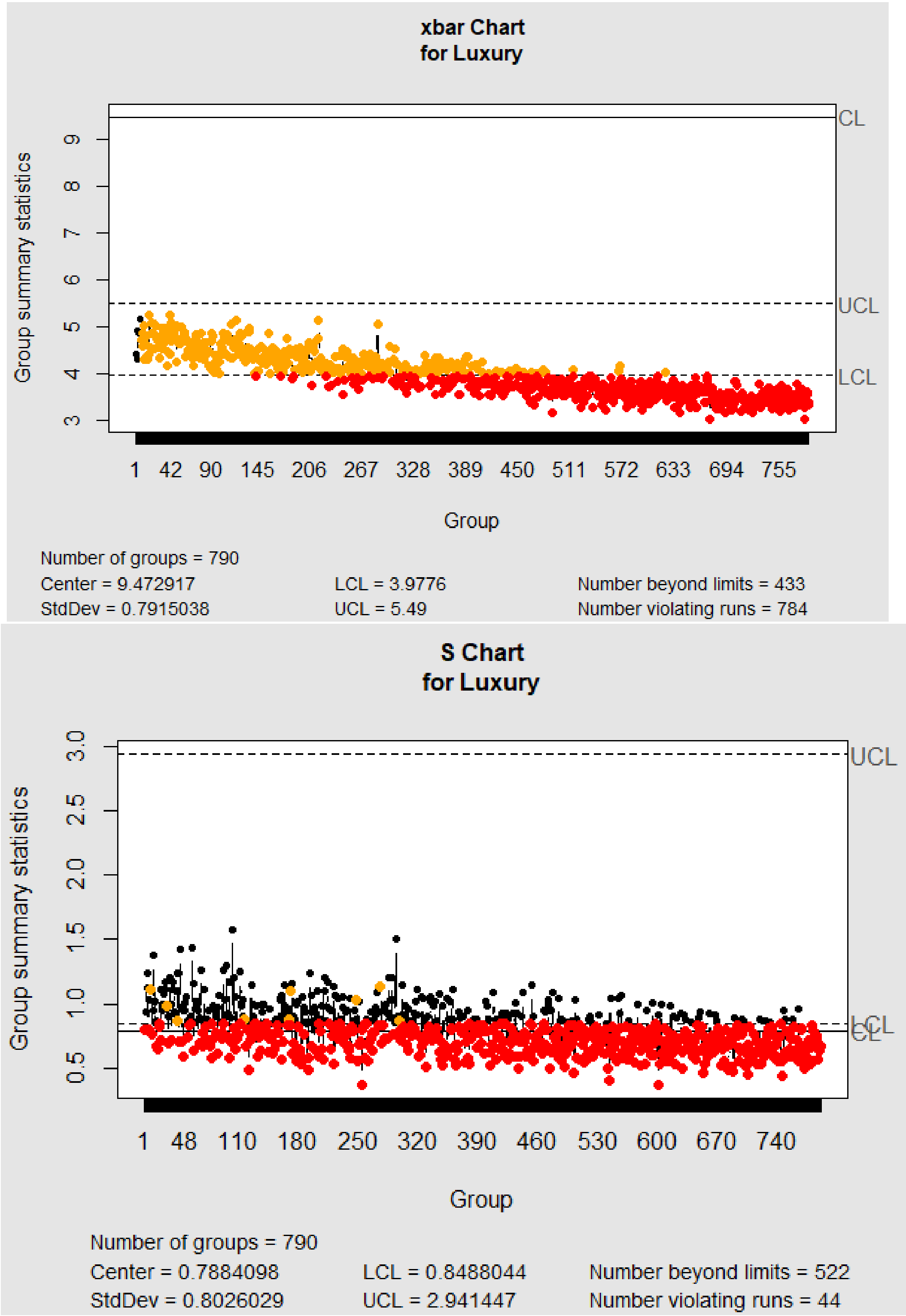


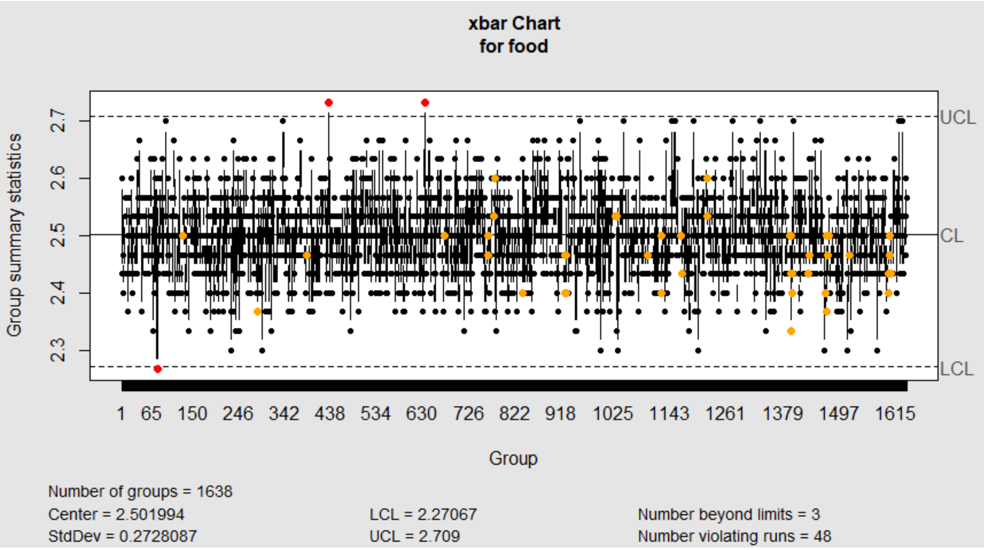
Figure 25:All samples for luxury, 2 plots

Time for luxury deliveries decreased. Perhaps this is due to the high value of the luxury goods class and the necessity for quick delivery to maintain strong luxury item sales.

After the 48th sample, luxury appears to be consistently decreasing outside of the control boundaries. The cause of this decline has to be looked at by a sales department expert. Because luxury goods are the most expensive products and faster delivery would enhance income, the drop may be a sign that the corporation has focused more on providing them (the customers will be more satisfied and buy more luxury products).

The S-bar chart is under control (just 522 samples deviate from control limits), hence the X-bar chart's conclusion is valid.

Food



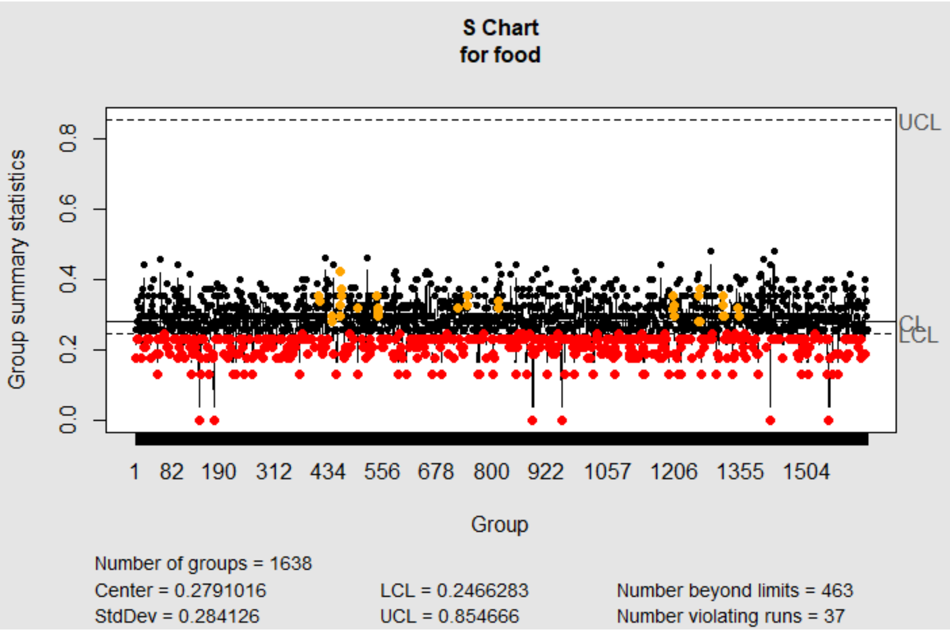


Figure 26:Xbar and S chart for all samples of food

Food appears to be in check; thus, things are steady for the time being despite a few occurrences that appear to be outside of the control limits.

The S-bar chart is under control (with just 463 samples beyond the control boundaries), hence the X-bar chart's conclusion is reasonable.

Part 4: Optimising the delivery processes

From the X-bar and s-chart, information can be seen about the classes based on the out-of-control instances and is shown below in the following two tables.

Class	Total found	First	Second	Third	Last3	Last2	Last1
Clothing	11	135	423	781	1472	1476	1546
Household	372	418	428	599	1248	1249	1250
Food	3	313	1073	1078	NA	NA	NA
Technology	16	369	453	599	1751	1808	2019
Sweets	3	1031	1269	1311	NA	NA	NA
Gifts	2144	196	200	201	2440	2441	2442
Luxury	432	156	168	169	735	736	737

Table 7: Outliers

From the table above, there can be seen clearly that the first class, clothing and food (3rd) and sweets (5th) are in control since there are only a few examples outside of the control limits. Household and gift- classes are not in control, as there as many instances out of the control limits. A further investigation could be done to see if there is a core issue. As for the luxury and household, there are only a few instances out of control, but inspection could still be taken.

The table below indicates the consecutive samples out of the control limits of -0.3 and +0.4 sigma.

Class	Maximum Pattern Length	Last Sample position of first	Last Sample position of last
Clothing	4	202	202
Household	4	483	491
Food	3	118	1285
Technology	5	725	1344
Sweets	4	693	693
Gifts	3	53	53
Luxury	5	311	388

Table 8: Consecutiveness of outliers

Luxury and technology have the most consecutive samples (5) out of the control limits, which can show a policy that needs to be changed etc. It also shows that luxury and technology class have the most variability. There should be inspection done on these classes. Recalculations can also be done again to ensure accurate results, like on luxury or even technology.

4.2 Type 1 error: A&B

Estimating the probability of making a type 1 error will be discussed.

Null hypothesis: H0: The process is centred on centreline (mean is in the control limits) and the process is in control.

H1: The process is not centred on centreline (mean is not in the control limits) and the process is not in control. The process could have increased or decreased in variation.

	Process is fine	Process is not fine
SPC indicates the process is not fine	Type 1 error/ Manufacturer's error	Correct to fix the process.
SPC indicates the process is fine	Correct if do nothing	Type 2 error/ Consumer's error.

Table 9: Type 1 and 2 error

__Question A:__

Probability of performing type 1 error: 0.27%

Type I errors occur when a process appears to be out of control when it is actually under control. The likelihood of a type I mistake occurring is 0.00269 [0.27%]. If the process is under control, the type 1 mistake will result in an inaccurate conclusion. A type 1 mistake has an extremely low probability, at 0.27% thus this is not likely to occur.

```
> pnorm(-3)*2
[1] 0.002699796
> #B
> (1-pnorm(0))
[1] 0.5
> 
```

Figure 27: Calculation code of A and B

__Question B:__

Probability of performing type 1 error:

Type I errors occur when a process seems to be beyond the -0.3 and +0.4 sigma-control limits (as specified) when it is actually within the control limits. The probability of a type 1 mistake occurring is 0.5 (as seen above in the calculation). If the operation is actually inside the control boundaries, the type 1 mistake will result in an inaccurate conclusion.

4.3 Optimal number of delivery days

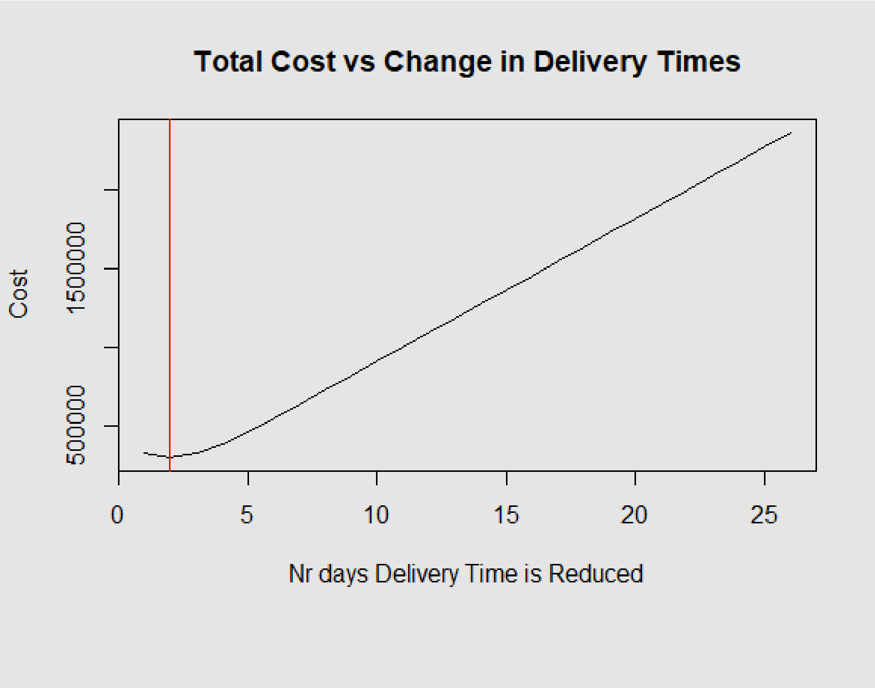
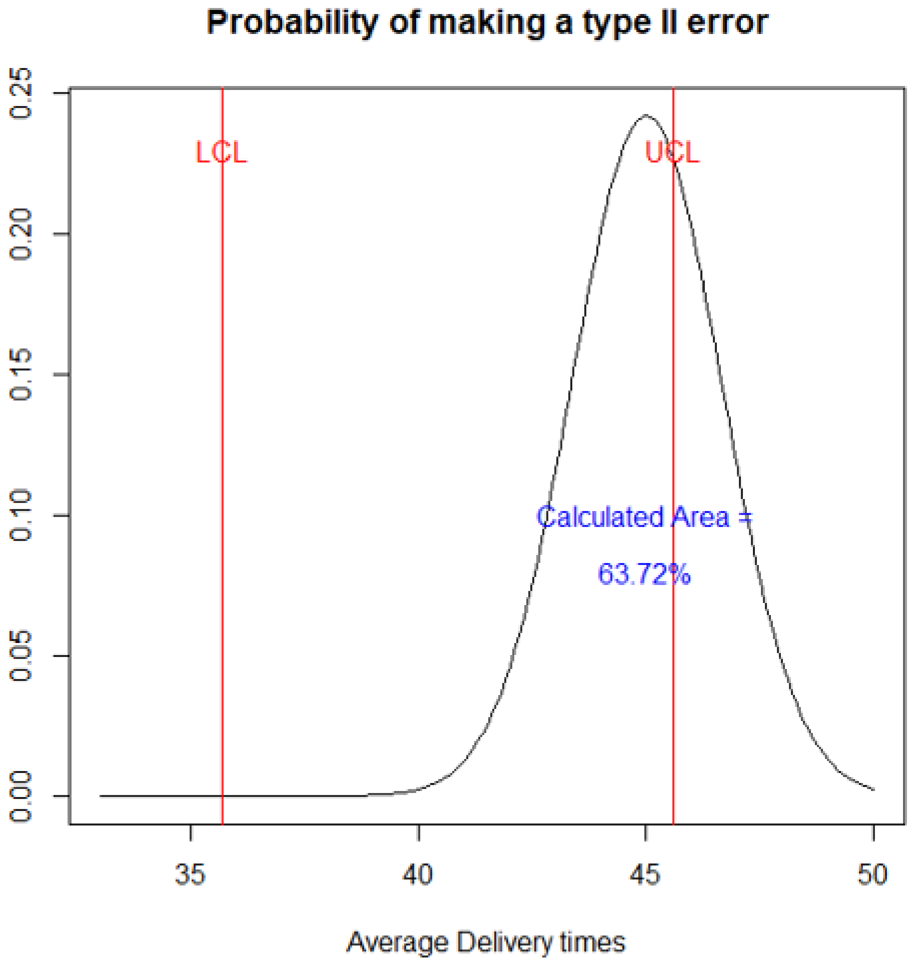


Figure 28: Optimal number of delivery days

The average hours that delivery takes are 20.0195 hours=20hours. You can reduce the number of hours by about 2 hours (to make the optimal amount 24 hours). The additional cost currently is because of lost sales is R446124. If you reduce the hours with 2 hours, you can save R298201.

To determine the appropriate number of delivery days on which to centre the delivery process for maximum profit, the cost of lowering the average delivery days by one day is determined. The lowest possible cost will result in the highest possible profit.

4.4 Type 2 error



In this scenario, a type II error for the delivery time for the class Technology happens when the product is delivered late, yet the company believes the technology product is supplied on time. The graphs outside control limits (UCL and LCL) are highlighted by red lines.

The probability of committing a type II error is 0.6372. (Indicated as the area of the graph between the two control limits).

This chance is relatively high, and the corporation must be cognizant of ensuring that the product is delivered on time rather than simply assuming that the product is supplied on time.

Part 5: MANOVA tests

To determine if there is any connection between the dependent variables and the independent factors, a MANOVA table is used to provide a p-value for each dependent variable (Dissertation, 2022).

P-value: I chose a value for p=0.05, as this is the most common p value to choose from and the most universally used.

First hypothesis

The product's class serves as the dependent variable in this scenario, while its price and delivery time serve as its independent variables.

independent variables	Price, Delivery times and Age
dependent variable:	Class of each product
H0	The purchase patterns for each class did not significantly vary as a result of price, delivery periods, or age.
H1	The purchase pattern is affected by at least one feature.

Table 10: First hypothesis

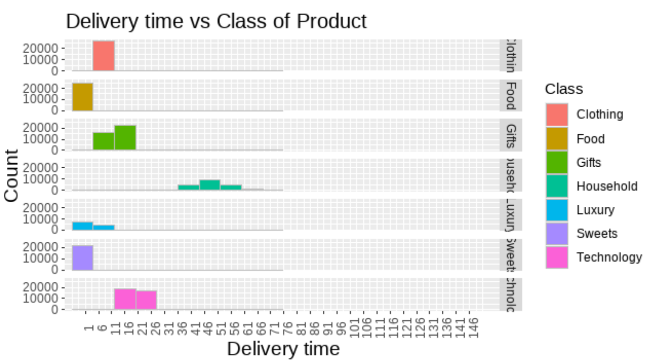
P-value:

```
> manova1 <- manova(cbind(Delivery.time, Price, AGE) ~ Class, data = valid)
> summary(manova1)
              Df Pillai approx F num Df den Df    Pr(>F)    
Class          6 1.7577    42438     18 539928 < 2.2e-16 ***
Residuals 179976                                            
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As seen in image above after the code was executed, p is less than 2.2e-16 which is smaller than 0.05. Thus:

- Reject Null Hypotheses.
- The average of at least one dependent variable varies.
- Price, delivery dates, and age are extremely significant differences between classes, according to this result.

Visualising the results:



Boxplots:

Figure 29: Box plot of delivery time vs class

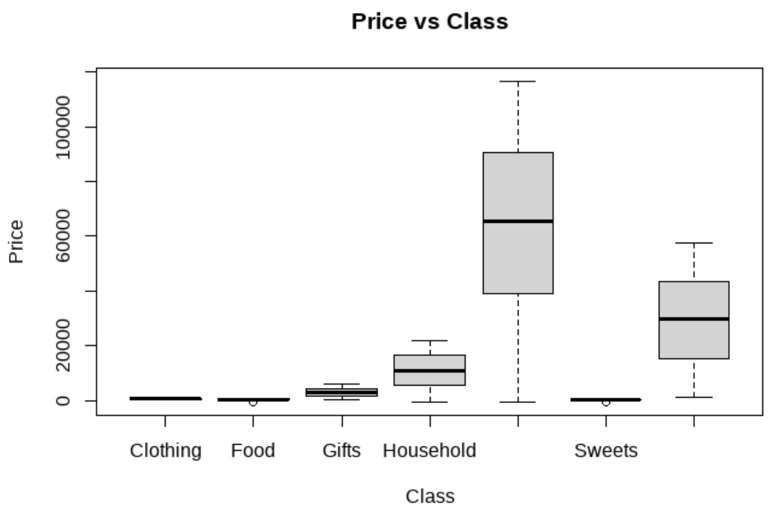


Figure 30: Box plot of price vs class

Conclusion for hypothesis 1:

The household delivery time is far quicker than the other delivery timings, although the x bar chart renders this conclusion unreliable (the chart indicated the process is not in control). Reliability of service provision will decline.

Because luxury goods are more valued as commodities, the buying behaviour suggests that luxury goods are more costly. To boost sales, luxury goods must have high levels of reliability and service. Because younger individuals are more likely to utilize technology, emphasis should be made on marketing technology to this age. Each class has an even distribution of ages.

Second hypothesis

The product's reason for purchase serves as the dependent variable in this scenario, while price and delivery time serve as its two independent variables.

independent variables	Price, Delivery times and Age
dependent variable:	Why bought
H0	The purchase patterns for why each product was bought did not significantly vary as a result of price, delivery periods, or age.
H1	The purchase pattern is affected by at least one feature.

Table 11: Second hypothesis

P-value:

```
> summary(manova2)
                 Df   Pillai approx F num Df den Df
Why.Bought        5 0.044145   537.59     15 539931
Residuals    179977
                Pr(>F)
Why.Bought < 2.2e-16 ***
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As seen in image above after the code was executed, p is less than 2.2e-16 which is smaller than 0.05. Thus:

- Reject Null Hypotheses.
- The average of at least one dependent variable varies.
- Price, delivery dates, and age are extremely significant differences between why the product is bought, according to this result.

Visualising the results: Boxplots:

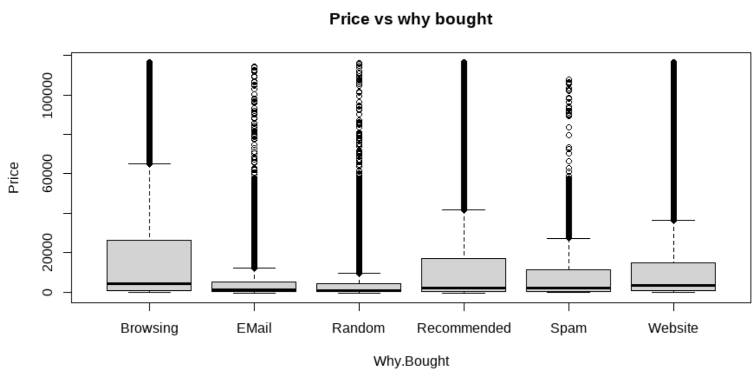


Figure 31:Box plot of price vs why bought

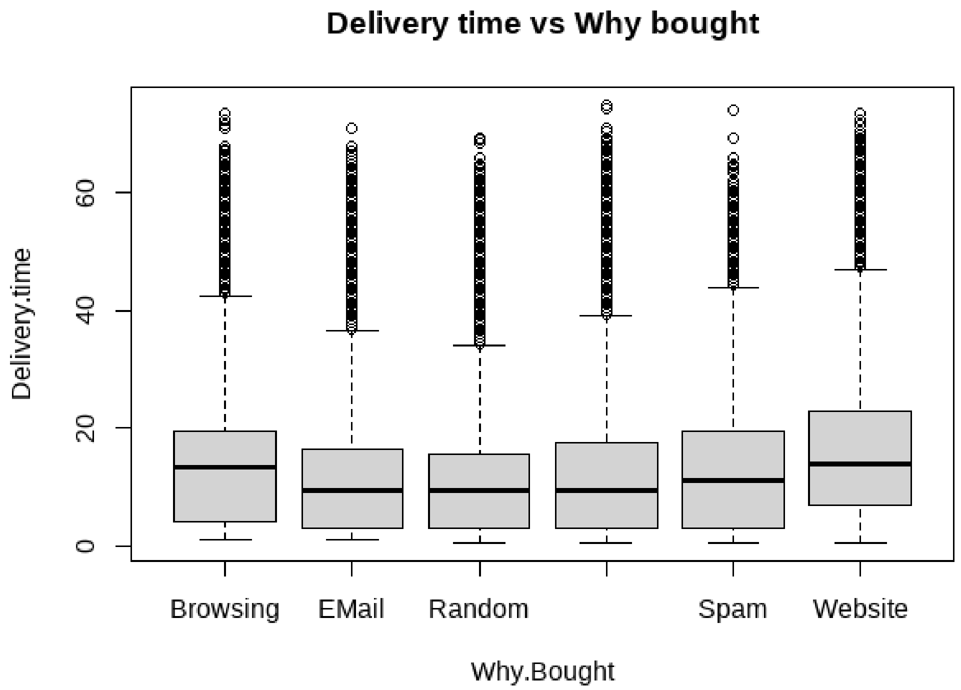


Figure 32: Box plot of delivery time vs why bought

Conclusion for hypothesis 2:

When a product is purchased for "website"-related reasons, the delivery time is greater. The age is a factor in sales that is widely spread.

Third hypothesis

The product's reason for purchase as the dependent variable in this scenario, while day month and year serve as its two independent variables.

Dependent variables	Why bought
Independent variable:	Day month and year
H0	Day, Month and year made no significant change to the buying pattern of why the product is bought.
H1	The purchase pattern is affected by at least one feature.

Table 12: Third hypothesis

P-value:

```
> summary(manova3)
                Df    Pillai approx F num Df den Df    Pr(>F)
Why.Bought       5 0.002323   27.894     15 539931 < 2.2e-16 ***
Residuals   179977
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

As seen in image above after the code was executed, p is less than 2.2e-16 which is smaller than 0.05. Thus:

- Reject Null Hypotheses.
- The average of at least one dependent variable varies.
- Price, delivery dates, and age are extremely significant differences between why the product is bought, according to this result.

Visualising the results: Boxplots:

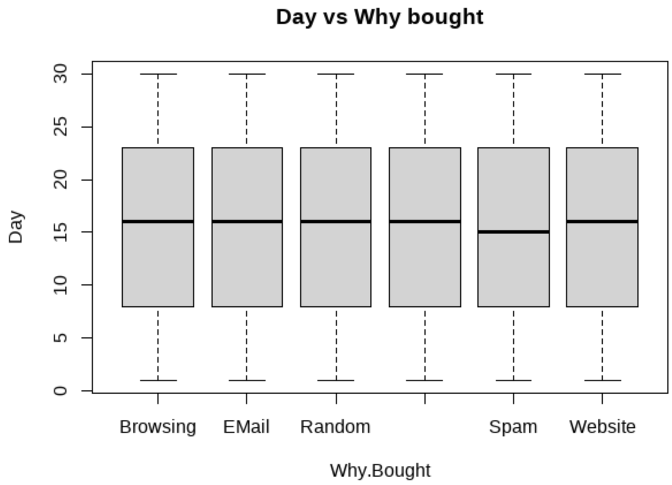


Figure 33: Box plot of price vs why bought

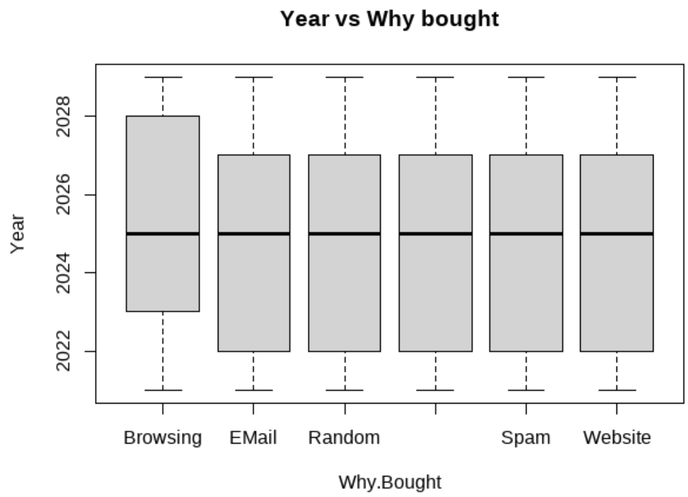


Figure 34: Box plot of Month vs why bought

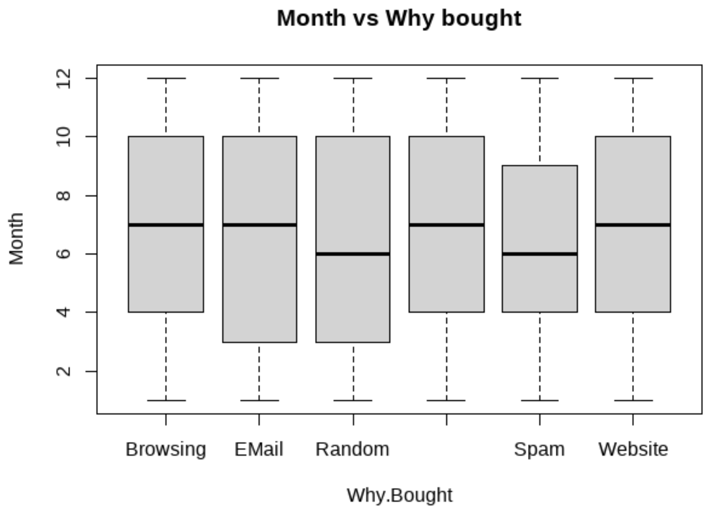


Figure 35:Box plot of price vs why bought

Conclusion for hypothesis 3:

The day and month of sales are unaffected by the product's class. However, the year the goods is purchased is influenced by the class. Between 2025 and 2029, less customers purchased products on "recommendation" and "website" grounds. Even though this decline is not large, it has to be looked into.

Part 6: Reliability of the service and products

Problem 6.1

According to the Taguchi Loss Function, deviation from the aim would result in unsatisfied customers. Customer discontent rises as the deviation from the target increases (BIZPI, 2022).

Calculate the constant by the following variables:

t <- 0.04 #target

T <- 0.06 #deviation

L <- 45 #This is the scrap value

k <- L/(t^2) #calculate the constant

= 28125

Loss function(x) = $\boldsymbol{k}\ \mathbf{(}\boldsymbol{x} - \boldsymbol{T}\mathbf{)^2}$

$L(x)$ = $\mathbf{28125\ (}\boldsymbol{x}\mathbf{ - 0.06)^2}$

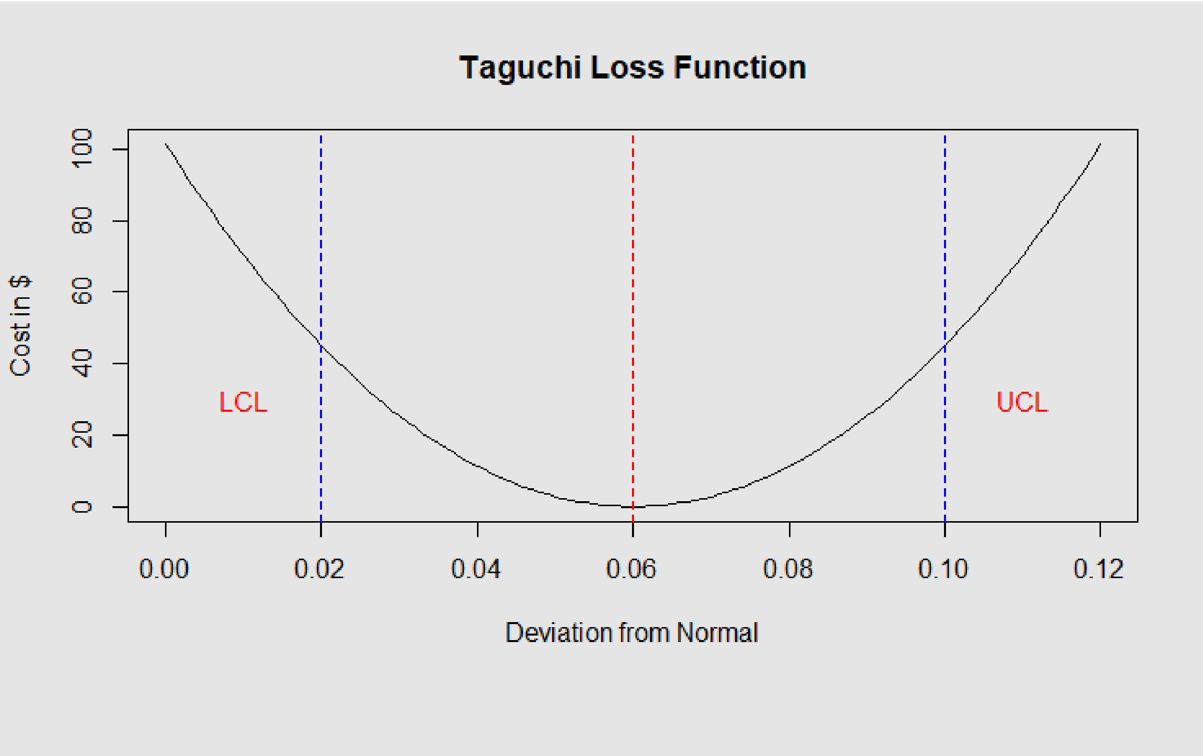


Figure 36: Taguchi Loss function graph

Conclusion:

The quality of the product is worse and the cost to the company is higher the more a particular product's characteristic deviates from the target value (0.06). As a result, the service will be less effective, and the goods will be unreliable.

Problem 7a

Calculate constant:

$L(x) = k\,(x - T)^2$ #loss function

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

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

$=$**21875** #constant

Loss function(x) = $\boldsymbol{k\,(x - T)^2}$

$L(x)$ =**21875** $\boldsymbol{(x - 0.06)^2}$

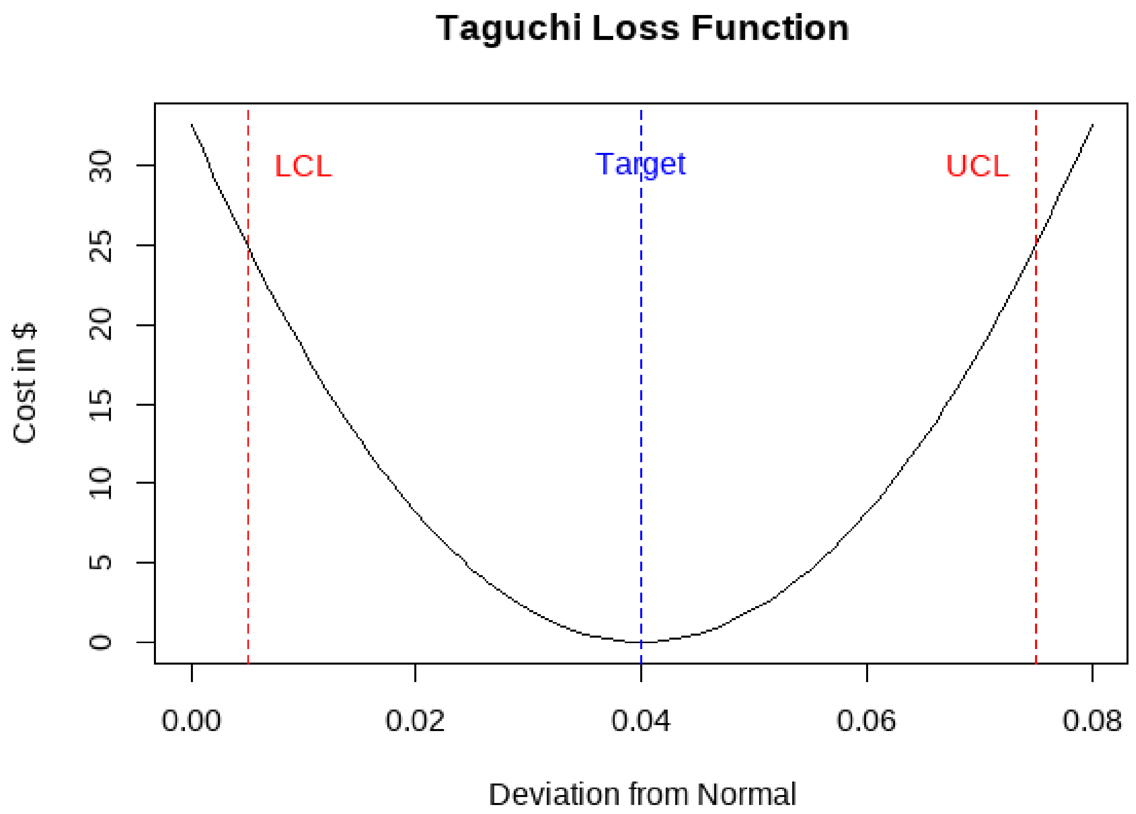


Figure 37: Taguchi loss function of problem 7

The product varies from the target as seen above on figure 37. The target is 0.06cm. If the quantity can be decreased, the company's losses will increase which will lead to inefficiency and unreliability.

Problem 7b

If process variation from goal is decreased to 0.027, Taguchi loss will occur:

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

L (0.027)=21875(0.027)^2

=$15.95

This concludes that the company would make a loss (per item) of the value in dollar stated above. As a result, the company's services are now of worse quality.

6.2. Problem 27

a. The probability of one machine at every stage
Reliability = $Reliability\ (Machine\ A) \times Reliability(Machine\ B) \times Reliability(Machine\ C)$
= 0.85 × 0.92 × 0.90 **= 0.7038** =70.38%

b. The probability when both machines are used
Reliability= = $Reliability\ (A1\ \&\ \mathrm{A2}) \times Reliability(\mathrm{B1}\ \&\ \mathrm{B2}) \times Reliability(\mathrm{C1}\ \&\ \mathrm{C2})$=
$(1\text{-}(1\text{-}0.85)^{2}$ X $(1\text{-}(1\text{-}0.92)^{2}$ X $(1\text{-}(1\text{-}0.90)^{2}$) **= 0.9615 =96.15%**

Bar plot of situation a and b:

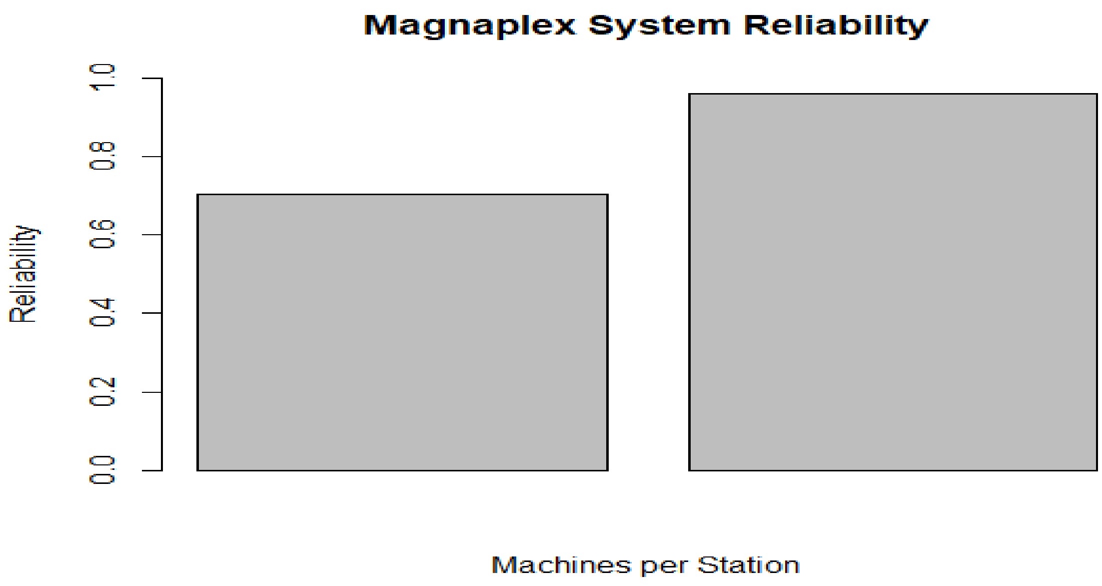


Figure 38:Barplot of situation a and b

Conclusion from Magnaplex problem:

As a result, parallelizing two identical machines will increase dependability by 26%. This is so that if one machine malfunctions, a parallel version of the same machine can continue to function. Running the two units concurrently would increase reliability for the business.

6.3. Binomial probability

We want to calculate the probability of having reliable vehicles

Using the following formula:

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

Figure 39: Binomial formula

1. $Px<1=20C1 * p^{1} * 1-p^{20-1} * 1560=1560-190+22+3+1=1344$

   p= **0.03280011**

   $P2=20C2 * p^{2} * 1-p^{20-2} * 1560=1560-190+22+3+1=190$

   p= **0.0348579**

2. $P3=20C3 * p^{3} * 1-p^{20-3} * 1560=1560-190+22+3+1=21$

   p**= 0.02701039**

3. $P4=20C4 * p^{4} * 1-p^{20-4} * 1560=1560-190+22+3+1=3$

   P=**0.02812168**

   $P5=20C5 * p^{5} * 1-p^{20-5} * 1560=1560-190+22+3+1=1$

   p**= 0.03740828**

Thus, the weighted average is:

w1 <- (x1*1344+x2*190+x3*22+x4*3+x5*1)/1560

(Where xi={1,2,3,4,5} and is the probabilities calculated above)

= 0.03296304

Expected reliable delivery days in a year:

Px<2=20C2∗0. **0.0348579** ^{2} ∗ 1−**0.0348579** $^{20-}$ * 365 days= 355.2037

Probability of reliable drivers: same binomial equations

1. $Px<3=20C3 * p^{3} * 1-p^{20-3} * 1560=1560-190+22+3+1=1344$

p= 0.0779277

2. $P4=20C4 * p^4 * 1 - p^{20-4} * 1560=1560-190+22+3+1=190$

   P=0.08493793

3. $P5=20C5 * p5 * 1 - p^{20-5} * 1560=1560-190+22+3+1=21$

   P=0.0569216

4. $P6= 20C6 * p6 * 1 - p^{20-6} * 1560=1560-190+22+3+1=3$

   p= 0.05803208

Thus, the weighted average is:

w2 <- (1458*x1a+x2b*95+x3c*6+x4d*1)/1560

= 358.8645

Expected reliable delivery days in a year:

348.9175 days

Part 2: Increase vehicles by 1 to 21.

Code: I calculated the binomial distribution of p1 to p5, where the x value varied from 1 to 5.

```
p1 <- dbinom(0,21,prob=w1,log=FALSE)

p2 <- dbinom(1,21,prob=w1,log=FALSE)

p3 <- dbinom(2,21,prob=w1,log=FALSE)

p4 <- dbinom(3,21,prob=w1,log=FALSE)

p5 <- dbinom(4,21,prob=w1,log=FALSE)
```

The total probabilities, thus the sum of P1 to P5, is **0.9994914.** That multiplied by 365 to get the days**= 364.8144.**

Conclusion

In order to better comprehend and analyse the data, visual representations were developed once the given data about the internet company was filtered.

The process capability indices were also generated and utilized to provide a comparison between the distribution and the specifications' range. The control limits were established using the first 30 samples in the statistical process control calculations. For each of the many classes, X-bar and S-charts were created, and the various classes that are out of control are shown.

The type l and type ll errors were computed together with the technology delivery timeframes, and it is obvious that the likelihood for a type l error is suggestively lower than a type ll error.

In order to determine if the selected independent variables and selected dependant variables are correlated with one another, a MANOVA table is created. Boxplots shows this result visually.

The company will receive useful information from this data analysis to help it grow and increase its earnings.

Bibliography

BIZPI, 2022. *Taguchi Loss Function.* [Online]
Available at: www.leansixsigmadefinition.com
[Accessed 12 October 2022].

Dissertation, 2022. *Multiple analysis of variance (MANOVA).* [Online]
Available at: www.statisticssolutions.com
[Accessed 11 October 2022].

EPICOR, 2022. *Why Is Data Important for Your Business?.* [Online]
Available at: www.grow.com
[Accessed 12 October 2022].

Hessing, T., 2022. *Statistical Process Control (SPC).* [Online]
Available at: www.sixsigmastudyguide.com
[Accessed 5 October 2022].

iSixSigma, 2022. *Categorical vs. Continuous Data: What's the Difference?.* [Online]
Available at: www.isixsigma.com
[Accessed 17 October 2022].

NIST, 2022. *What is Process Capability?.* [Online]
Available at: www.itl.nist.gov
[Accessed 11 October 2022].

Stobierski, T., 2021. *Data Wrangling: What It Is & Why It's Important.* [Online]
Available at: www.online.hbs.edu
[Accessed 1 October 2022].