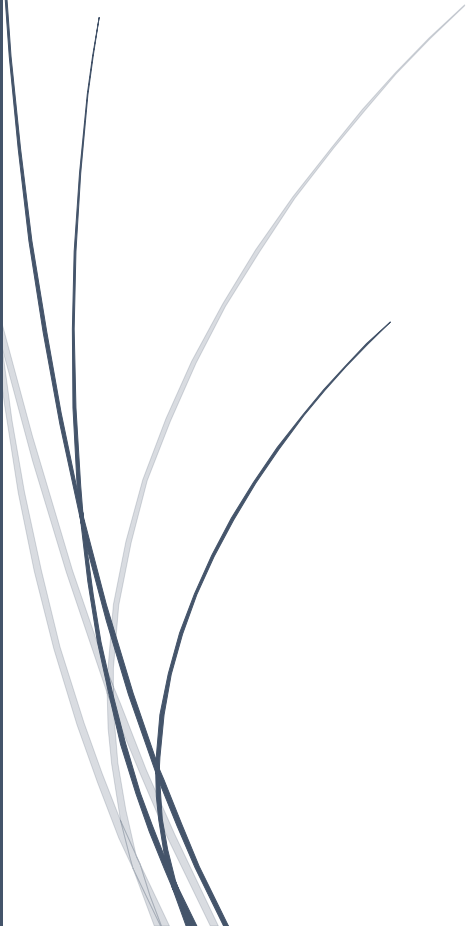


A dark blue vertical bar runs down the left side of the page. A blue arrow points to the right from this bar, containing the date.

10/17/2022

ECSA Graduate Attributes Project

Quality Assurance 344

Several thin, curved lines in shades of blue and grey originate from the bottom left and sweep upwards and to the right.

Ruben van Eck
23573627

Table of Contents

Table of Figures	3
List of Tables	4
Introduction	5
1. Data Wrangling.....	6
2. Descriptive Statistics.....	6
Age of customers who made a purchase	7
Age difference over time in class	7
Ages over different months	8
Number of sales in Price intervals	9
Number of sales in Price intervals per class.....	9
Number of sales in Price intervals per why the purchase was made	10
Delivery times over time.....	10
Delivery times over time per class	11
Delivery times per month	12
Number of deliveries per month.....	12
Number of deliveries per month per class.....	13
Number of deliveries per year	13
Number of deliveries per year per class	14
Process Capability Indices	14
3. Statistical Process Control	15
X-Chart	15
S-Chart	15
Out of Control charts	16
Moderately in Control Charts.....	17
.....	17
In Control Charts	17
.....	17
Control Charts with sample size of 30	18
4. Optimizing Delivery Process.....	20
4.1) Optimizing the delivery process	20
4.1.b) largest number of consecutive samples within 1 sigma	20
4.2) Probability of making a Type 1 error	20
4.3) Delivery time reduction analysis	21
4.4) Probability of a Type 2 error	22
5. MANOVA.....	23

Delivery time per Class	24
Delivery time per Month.....	24
Boxplot of Delivery time and Class.....	25
Boxplot of Price and Class.....	25
6. Reliability of the service and products.....	26
6.1) Problem 6 and 7	26
Problem 6:.....	26
Problem 7.a).....	26
Problem 7.b).....	26
6.2) Problem 27	26
6.3) Vehicle and personnel availability analysis	27
Conclusion.....	31
References	32

Table of Figures

1. Age of customers who made a purchase	7
2. Age difference over time in class	7
3. Ages over different months	8
4. Number of sales in Price intervals	9
5. Number of sales in Price intervals per class	9
6. Number of sales in Price intervals per why the purchase was made	10
7. Delivery times over time	10
8. Delivery times over time per class	11
9. Delivery times per month	12
10. Number of deliveries per month	12
11. Number of deliveries per month per class	13
12. Number of deliveries per year	13
13. Number of deliveries per year per class	14
14.. Luxury Control Chart	16
15.. Gifts Control Chart	16
16. Household Control Chart	16
17.. Technology Control Chart	17
18.. Clothing Control Chart	17
19.. Sweets Control Chart	17
20.. Food Control Chart	17
21. Control Chart of smaller sample size (Sweets)	18
22. Control Chart of smaller sample size (Household)	18
23. Control Chart of smaller sample size (Gifts)	18
24. Control Chart of smaller sample size (Technology)	18
25. Control Chart of smaller sample size (Luxury)	18
26. Control Chart of smaller sample size (Food)	18
27. Control Chart of smaller sample size (Clothing)	19
28. Total Cost vs Change in Delivery times	22
29. Type 2 error	22
30. Delivery time per Class	24
31. Delivery time per Month	24
32. Boxplot of Delivery time and Class	25
33. Boxplot of Price and Class	25
34. Reliability of machines	27
35. pbinom() of vehicles available	28
36. Vehicle and personnel availability code	29

List of Tables

Table 1. Process Capability Indices.....	14
Table 2. X-Chart.....	15
Table 3. S-Chart.....	15
Table 4. Out of Control samples.....	20
Table 5. Consecutive sample sequence.....	20
Table 6. Probability of being inside 1 standard deviation	21
Table 7. Manova (Age, Price)	23
Table 8. Manova (Month, Day)	23

Introduction

The report was constructed to give feedback and recommendations to management based all the different problems and scenarios. The first part of the report the given data was wrangled and was divided into usable data and unusable data, the usable data was then used in descriptive statistics, to visualize and analyse the data. This helped us find a better understanding of the nature of the data and the different features. Statistical process control techniques were used to try and find ways to optimize the possible solutions to the data, control charts were used to display and understand the features. The delivery times were optimized with statistical techniques. A Multi variable analysis of variance (Manova) was used to find the influence that the different features have on each other. The last part was using similar field problems to calculate solutions, to give further positive feedback and recommendations to management. Recommendations were then made based on the information that these field problems provided.

1. Data Wrangling

To work with the data that was given to us, (SalesTable2022) we needed to remove the entries that contained NA, in other words the entries that had missing values. This was done by using an R function to save the data set in another variable that was called **Validdata**. The data was split into 2 groups, where one group still contained the original data (Invaliddata) and the other the valid data that could be used for the remainder of the project.

After doing this the data set contained 179983 entries, with 17 unusable entries being removed. The data was then ordered according to the Year, Month, and the Day.

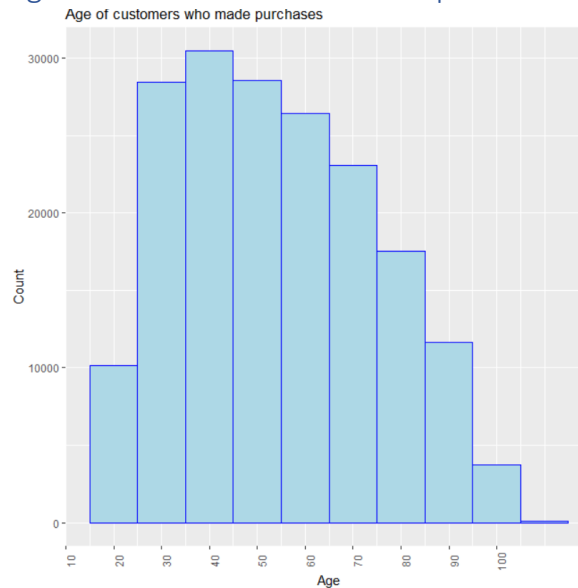
2. Descriptive Statistics

The first thing to do, when trying to analyse the data, was to identify all the different types of features in the data set, the features that were given were the following:

- X (Which acts as the order of the entries)
- ID
- Age
- Class
- Price
- Year
- Month
- Day
- Delivery time
- Why the purchase was made

There are 179983 entries in the data set, and it is physically impossible to manually read through the entries and have a better idea of the data, thus certain features needed to be identified to analyse and compare the data more thoroughly. It was decided to look at the Age of people who made purchases, the Class, the month in which purchase was made, the year in which the purchase was made, the price of the sales and what the delivery time of the sale was as well as why the sale was made. The features were analysed through using extensive histograms and graphs that were plotted in R studio, to gain a better understanding of the data.

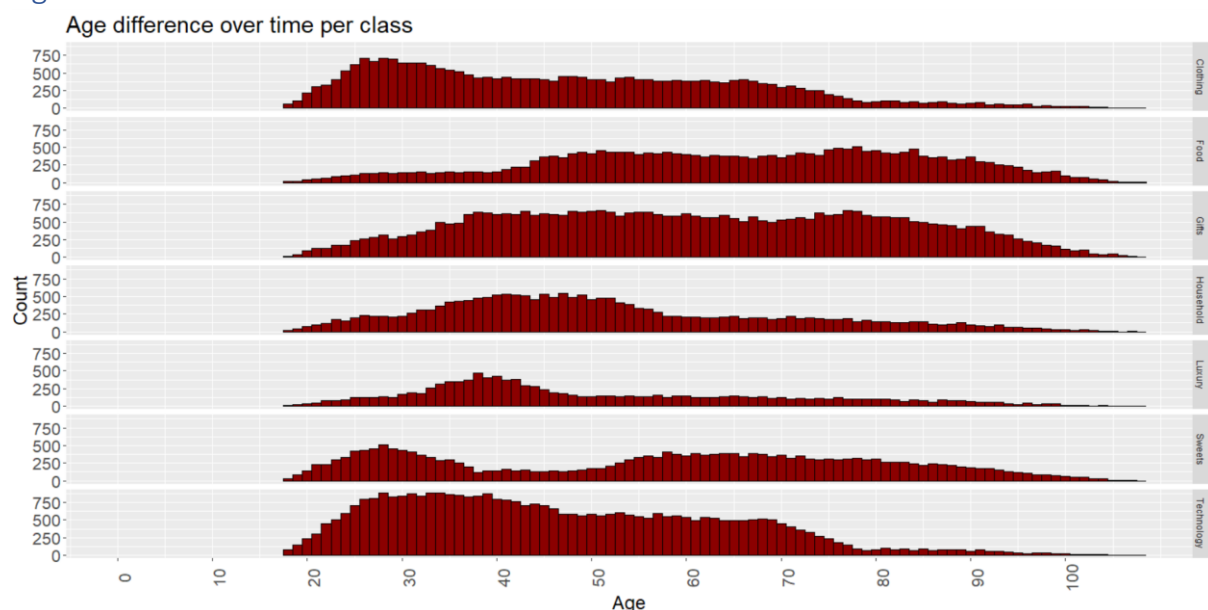
Age of customers who made a purchase



1. Age of customers who made a purchase

The first feature that was analysed was the Age feature, as I tried to establish if there was a correlation with the Age of purchasers and the rest of the features. The average age was 54.65, and according to the histogram is also clearly visible, there is a definite decline in sales after customers reach the age of 45, as well as before they are 25.

Age difference over time in class

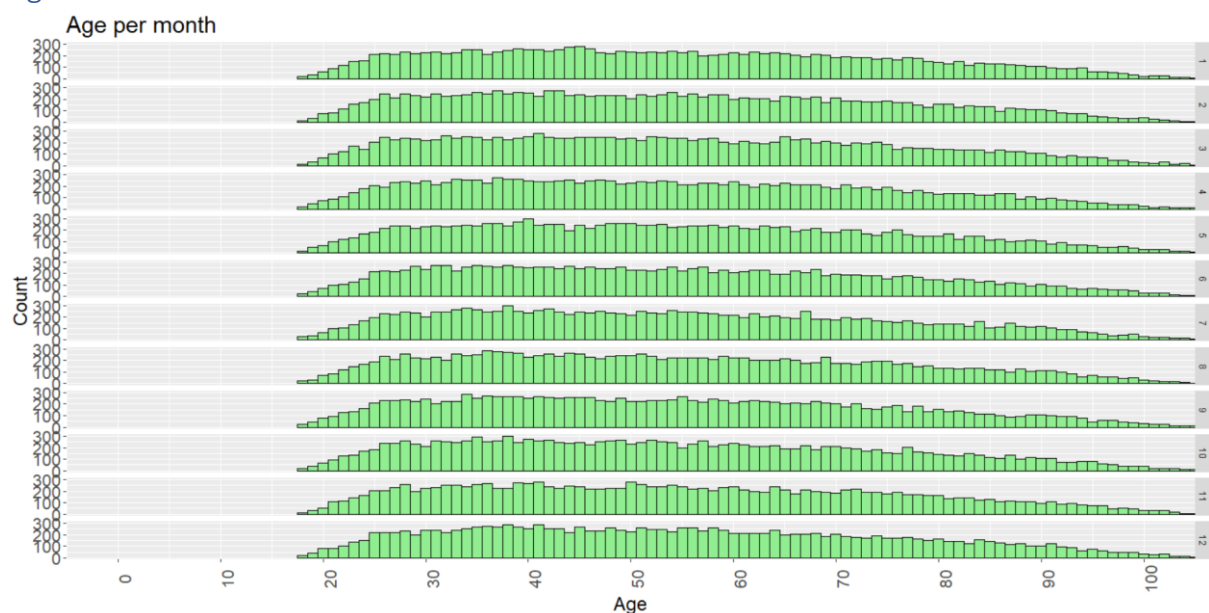


2. Age difference over time in class

By plotting a histogram that shows the different ages of types of purchases, it is evident to see that there is a difference in behaviour. The class Gifts has a normal distribution, as gifts is an item that is bought by purchasers from all ages, it has small deviations and remains relatively uniform. The food class is also a normal distribution as the food sales increase at 40 years old and then declines again at 80 years old, it is obvious as people that start families are those ages and their need for more food increases, buyers of younger ages do not need as much food or don't even buy their own food yet,

and then again older buyers do not need to buy food for their families anymore, and they themselves eat less food at that age. Clothing and Technology is skewed to the right as these are items that would be bought primarily by younger buyers, younger buyers would still have the desire to impress others, which is visually shown in the spike of sales with younger buyers, and technology is a relatively modern thing, as older buyers would maybe not have the need for technology or not know how to use it. Buyers between the ages of 20 and 50 would have a lot more use for technology items. Household and luxury items have very similar normal distributions, where they do not contain most of the sales, but the age in which these sales are made are very similar. This is evident, because these are also the ages that people start families and buy houses, they are more interested to buy new household items, and luxury items to improve their lifestyles, it is also clear that buyers who are older do not really have the need to spend money on household and luxury items, as they may have already bought these items when they were younger. Sweets has a bimodal distribution, where the histogram has 2 local maximums. This specific class was very interesting to me as the data shows that the sale of sweets until the age of 30 is regular and then there is a decrease in sales, but the sales increase again when buyers are at the age of 50. Almost as if people develop a sweet tooth when they are in their youth and lose interest in sweets after a couple of years, but they regain their sweet tooth later in their lives and then consistently enjoy sweets.

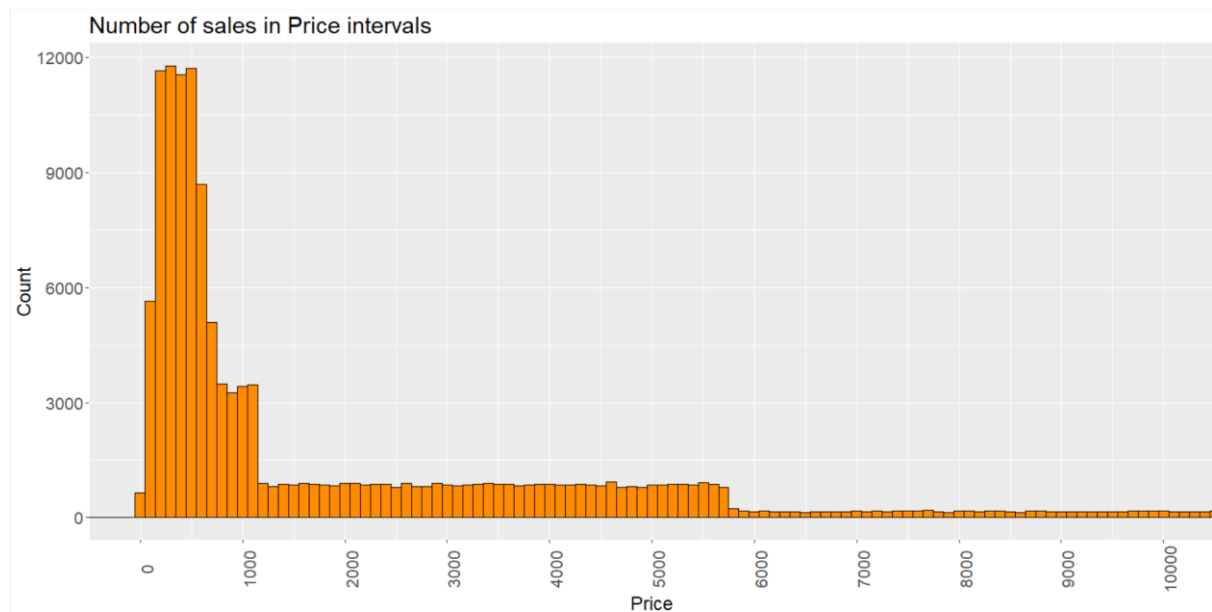
Ages over different months



3. Ages over different months

Not a lot could be said regarding to these histograms, only that the sales over each month remain very constant, and the age at which sales are made is normally distributed, according to the age at which more sales are usually made. The youth and the elderly make significantly less sales than the buyers who are middle-aged.

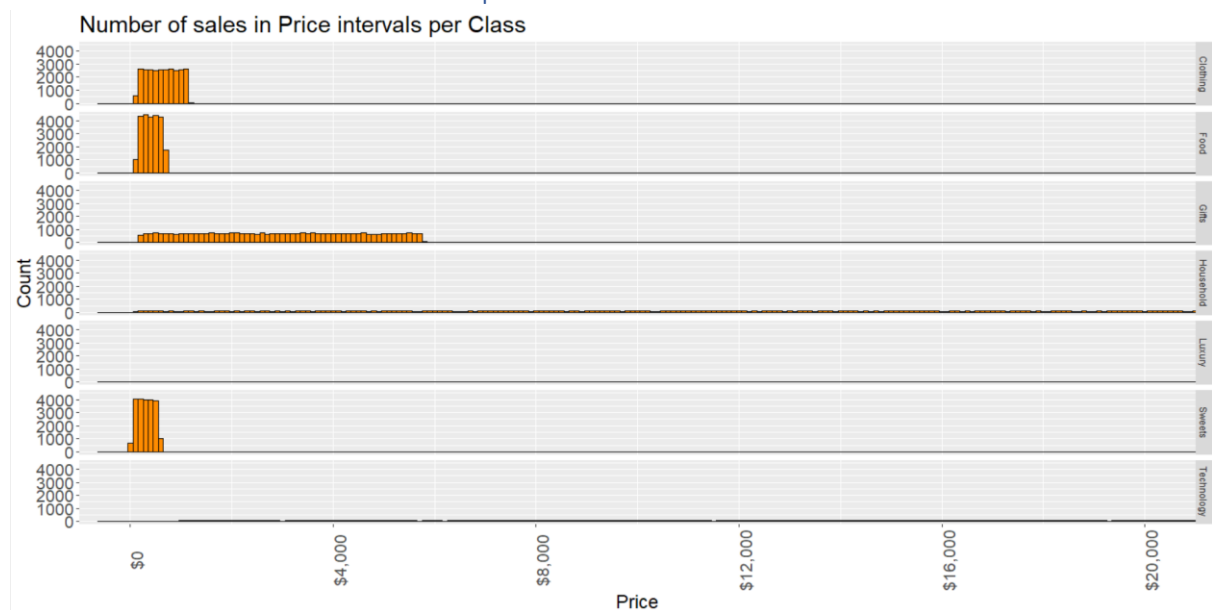
Number of sales in Price intervals



4. Number of sales in Price intervals

The price feature was analysed to determine the spread of the prices of the products. The number of purchases made in each price class was skewed to the right, but still had very distinct sections, it was clear that the majority of sales were in a lower price class, and the occurrence was investigated further, which later showed that the types of sold products (Class) had a impact on the price.

Number of sales in Price intervals per class

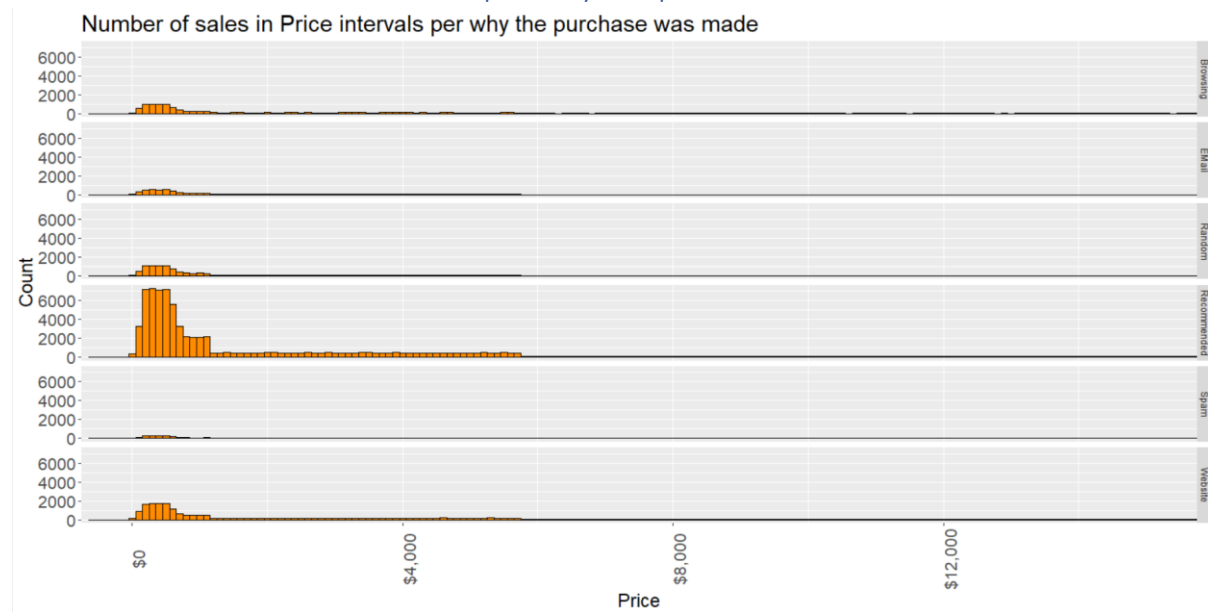


5. Number of sales in Price intervals per class

As discussed in the previous occurrence most of the sales fall in lower price classes, the price distributions of each class differ a lot, where the clothing, food and sweet items have low prices compared to the other products. The gift class is more evenly spread but still has a relatively low price, whereas the household class has prices up to \$20,000 and is much more evenly spread. The luxury and technology class are also evenly spread, and they have a very large price range where it

ranges to over \$50000 and \$100000 respectively. The food and sweets class are normally distributed, whereas the clothing, gifts, household, luxury, and technology class are uniformly distributed.

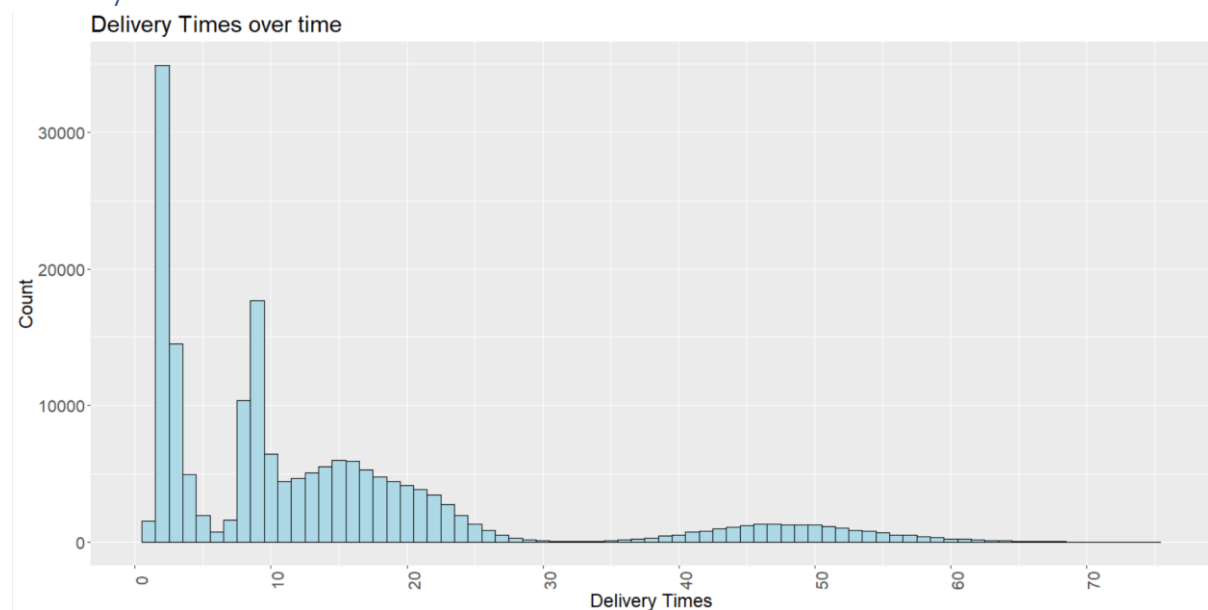
Number of sales in Price intervals per why the purchase was made



6. Number of sales in Price intervals per why the purchase was made

By analysing this histogram, it is evident that most cheaper sales were based of a recommendation. A very small portion of sales were made by spam and email, and only cheaper price products were advertised by these methods. Most of the expensive sales were made possible by browsing on the internet, and visiting the website, the distribution for all methods is the same, with only the number of sales that are different. The methods are all normally distributed.

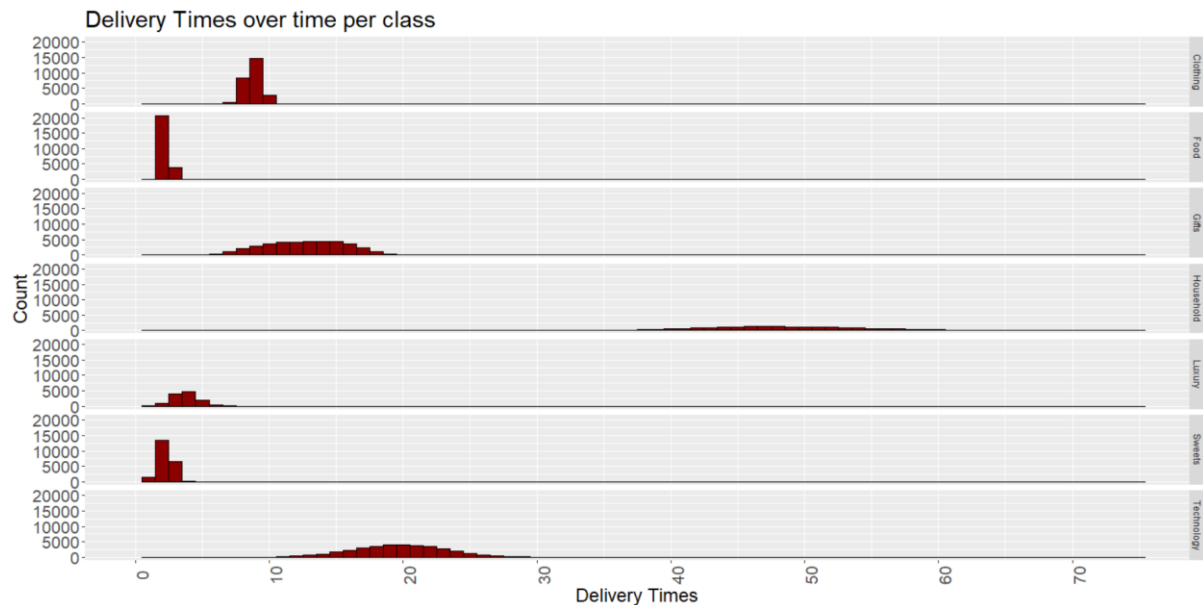
Delivery times over time



7. Delivery times over time

The delivery times are also an interesting feature to use when analysing the data, the delivery time is distributed in a bimodal manner, with a couple of local maximums. The most common delivery time is 2 hours, and then the second most common delivery time was 9 hours, it makes sense that most of the delivery times are short in length, as the company would try its best to minimize the delivery times to obtain customer satisfaction. There is a decline in the amount of delivery times from 20-40 hours, and it increases again between 40-60 hours. This may be evident to larger household items or clothes that needed to be shipped and ordered by the company only after the customer placed their order.

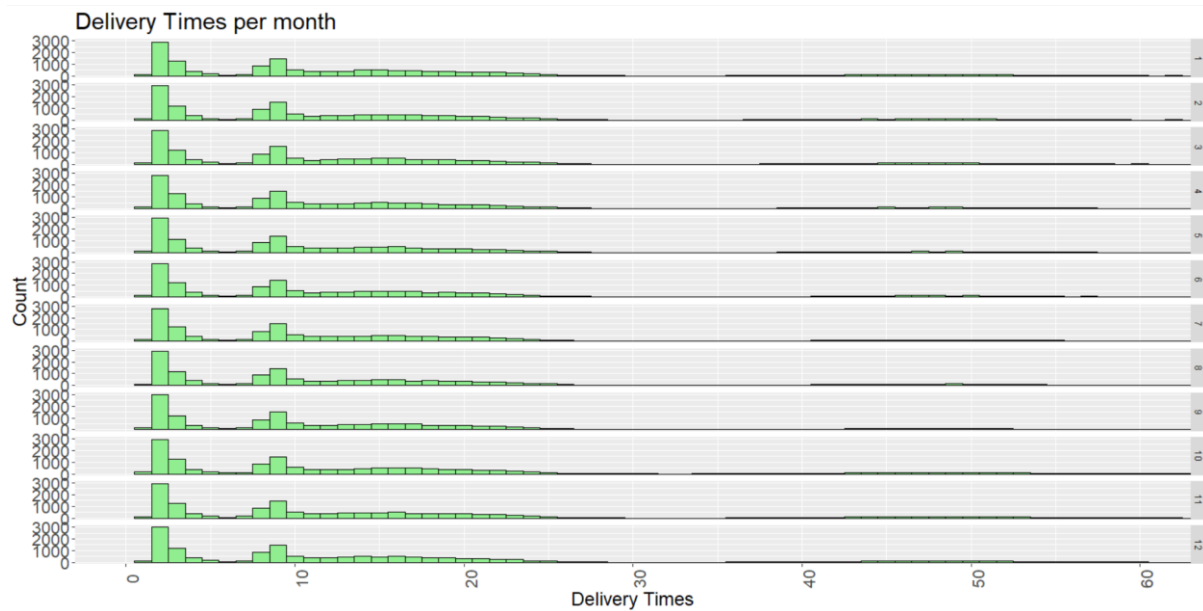
Delivery times over time per class



8. Delivery times over time per class

All the classes are normally distributed, except food, which is skewed to the right, referring back to the previous occurrence in the previous histogram it is evident that most of the products have short lead times. Sweets, luxury, food, and clothing items have very similar delivery times, and the delivery times are not spread out, they all also have relatively short delivery times. Gifts, household, and technology items have more spread-out delivery times, and has increase in their delivery time range. The reason being these items could take longer to order, or even physically bigger, thus the transportation of the item takes longer. Household items have on average the longest delivery times, probably because of most possible items, and that on average household items are rare and need to be ordered and shipped.

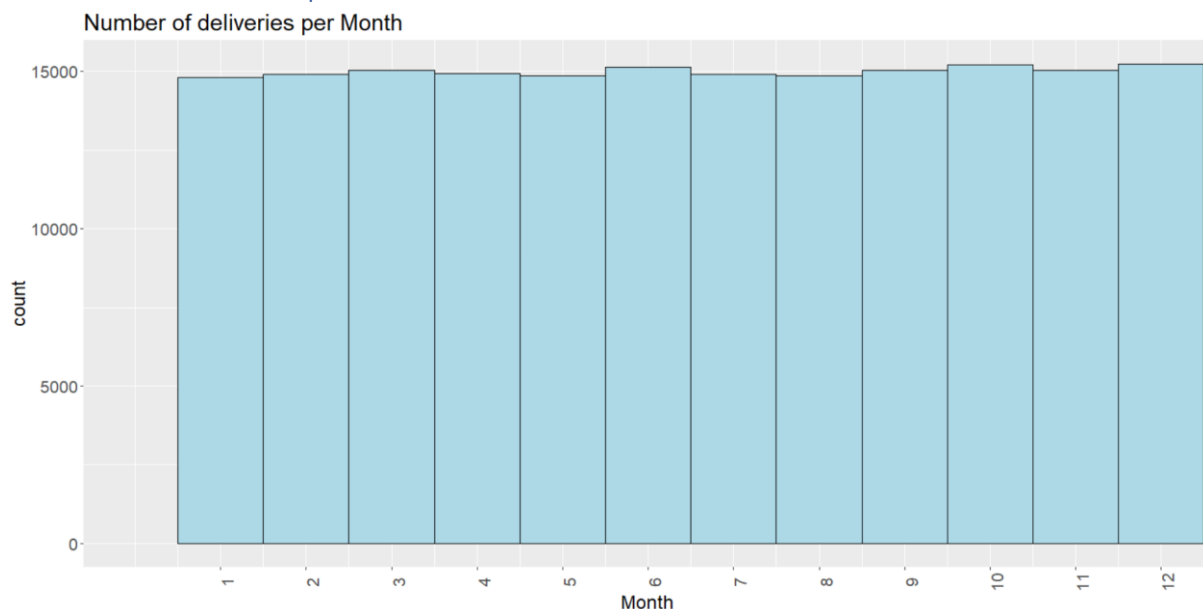
Delivery times per month



9. Delivery times per month

The delivery times are distributed bimodally, and the delivery times are constant throughout the year, and the possibility is ruled out that for instance delivery times were slower in the December holidays.

Number of deliveries per month

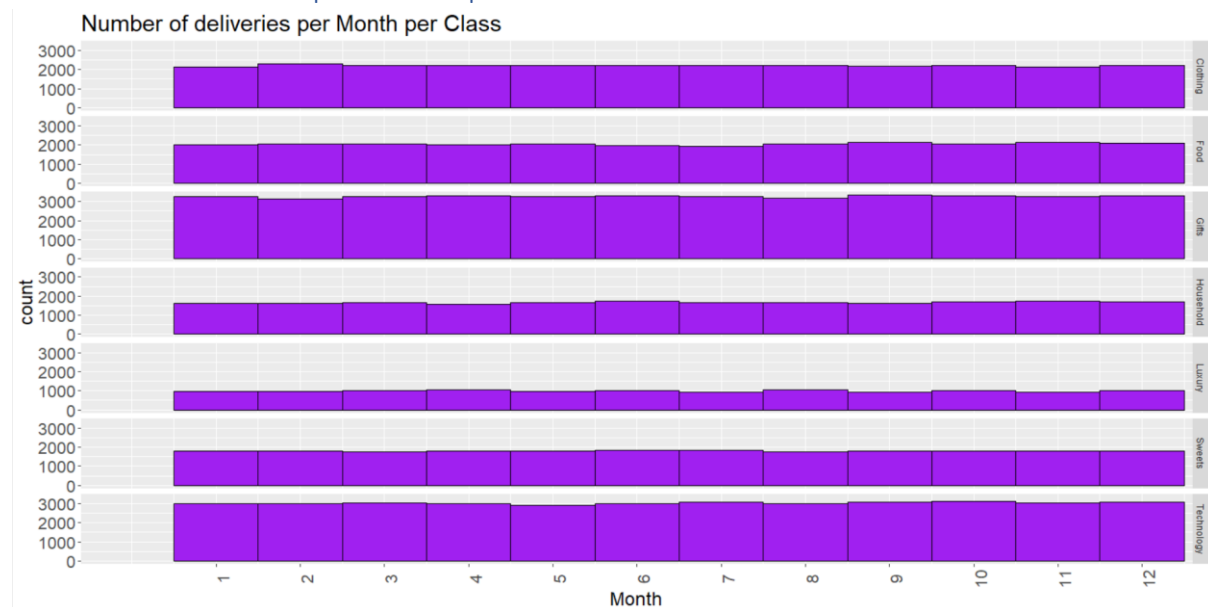


10. Number of deliveries per month

The number of deliveries per month is uniformly distributed and very evenly spread, there is maybe a slight increase in deliveries in March, June, October, and December. Which could be because of the 4 holidays in a year that take place in those months. Buyers have more times on their hands to order

in those times, and thus the delivery takes place in those times, but overall, the number of deliveries stay the same each month.

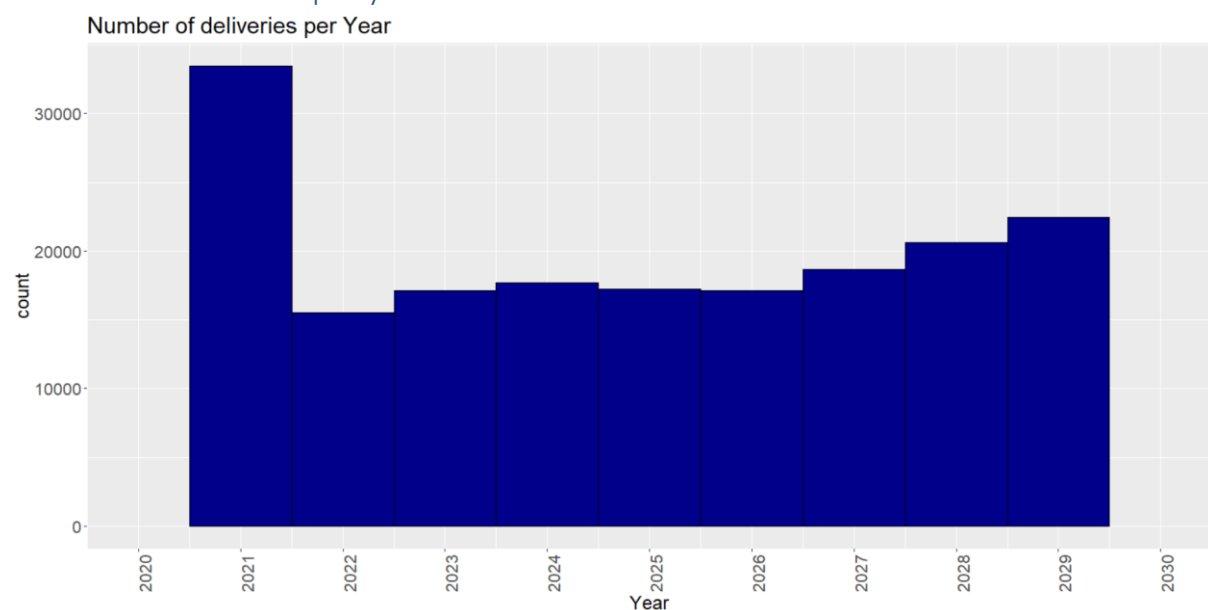
Number of deliveries per month per class



11. Number of deliveries per month per class

All the delivery times of the classes are uniformly spread over the 12 months, and the delivery times of all months and classes stay relatively constant. Technology and gifts visually have the greatest number of deliveries throughout the year, as they are the most bought items. The household items have less deliveries, probably something to do with their items not being ordered as often because of their price and their necessity. The luxury class has the least deliveries because buyers don't all have the need to buy luxurious items, and if the buyer has the need for luxury items, it is not an item that is bought monthly, but rather an item that is bought on occasion.

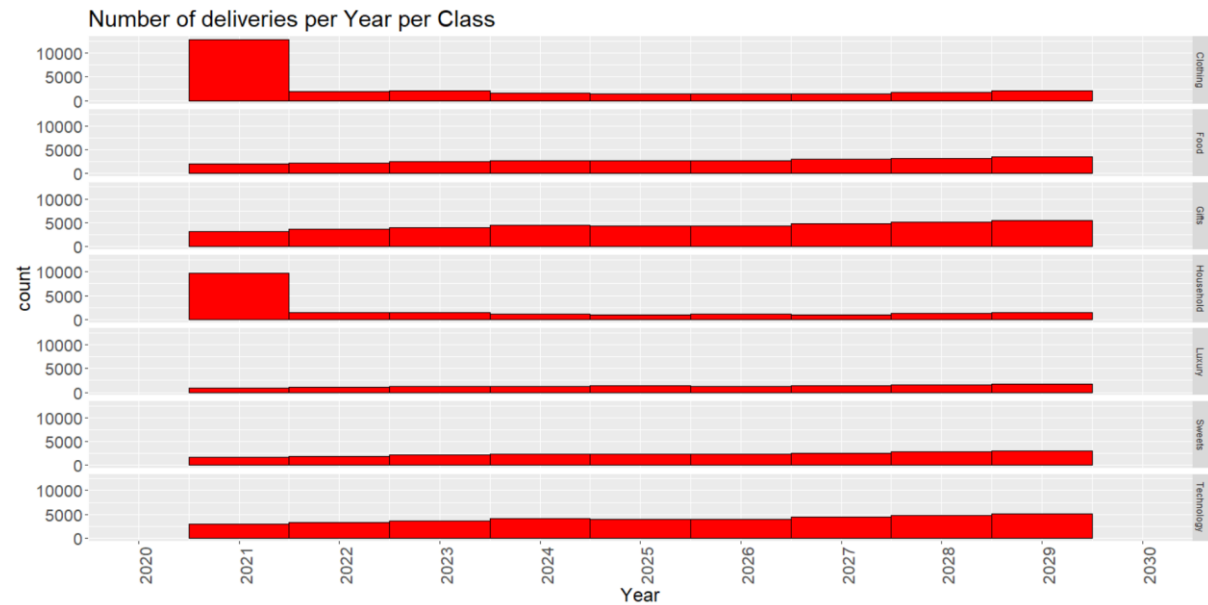
Number of deliveries per year



12. Number of deliveries per year

There is spike in deliveries in 2021 and then the deliveries drastically decrease between 2021-2022, after 2022 the deliveries start to gradually increase again, we would need to figure out what happened between 2021-2022, and for that we would plot the number of deliveries per year per class.

Number of deliveries per year per class



13. Number of deliveries per year per class

The occurrence makes more sense with these histograms, as it is evident that in 2021 there was a big spike in the deliveries of Household and Clothing items. It is difficult to give a reason for this scenario. Excluding the spike of sales in 2021 the delivery times are distributed uniformly throughout all the classes. After 2021 clothing and household items constantly had of the lowest number of deliveries. It is again clear that technology and gift items were constantly responsible for the highest number of deliveries.

Process Capability Indices

After the carefully done visual analysis was completed the process capability indices for the technology class was calculated. An upper service level of 24 days and a lower service level of 0 days was used. The lower service level of 0 is logical because it is possible for a item to be ready delivered the moment it is completed. The indices were then calculated by the same statistical calculations that were covered in Quality Assurance 344:

Table 1. Process Capability Indices

Index	Equations	Results
Cp	$\frac{(USL - LSL)}{6\sigma}$	1.142207
Cpu	$\frac{(USL - \mu)}{3\sigma}$	1.90472
Cpk	$\min(Cpl, Cpu)$	0.3796933
Cpl	$\frac{(\mu - LSL)}{3\sigma}$	0.3796933

3. Statistical Process Control

The next step in analysing the data was to generate process control charts for each class. This was done by first splitting the ordered Valid data (Ord.Validdata) into the various classes, and then I extracted the delivery times from each class by subsetting the data per class. After the data for each class was extracted, it was condensed by making sample groups of 15. The samples were then saved in a new variable where the sample size was created. The size was 30*15 which resulted in 450 samples.

Arrays were then created to calculate the UCL, U2Sigma, U1Sigma, CL, L1Sigma, L2Sigma and LCL for the X-chart and the S-chart. The calculations were the done by using Statistical calculations that were done by using a variety of equations with the help of constants like d2, B3, B4 and the standard deviation.

The control limits were then tabulated as follows:

X-Chart

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	9.390601	9.250401	9.110200	8.970000	8.829800	8.689599	8.549399
Food	2.702226	2.631484	2.560742	2.490000	2.419258	2.348516	2.277774
Gifts	9.451412	9.087978	8.724545	8.361111	7.997678	7.634244	7.270811
Household	50.126859	48.938647	47.750435	46.562222	45.374010	44.185798	42.997585
Luxury	5.468973	5.224501	4.980028	4.735556	4.491083	4.246610	4.002138
Sweets	2.883225	2.748076	2.612927	2.477778	2.342629	2.207479	2.072330
Technology	22.888932	22.050770	21.212607	20.374444	19.536282	18.698119	17.859957

Table 2. X-Chart

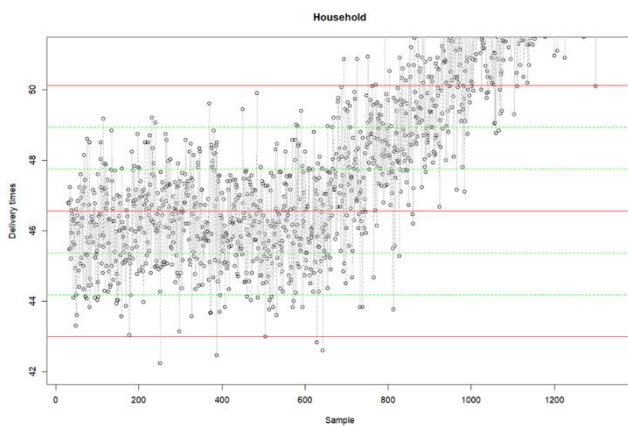
S-Chart

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	0.5661217	0.4259213	0.2857210	0.1455206	0.005320229	-0.13488014	-0.2750805
Food	0.2772340	0.2064920	0.1357501	0.0650081	-0.005733868	-0.07647584	-0.1472178
Gifts	1.4721776	1.1087441	0.7453106	0.3818771	0.018443639	-0.34498985	-0.7084233
Household	4.7468725	3.5586602	2.3704478	1.1822355	-0.005976865	-1.19418920	-2.3824015
Luxury	1.0194840	0.7750115	0.5305389	0.2860664	0.041593837	-0.20287870	-0.4473512
Sweets	0.5520133	0.4168641	0.2817149	0.1465656	0.011416392	-0.12373284	-0.2588821
Technology	3.4535377	2.6153751	1.7772125	0.9390500	0.100887383	-0.73727519	-1.5754378

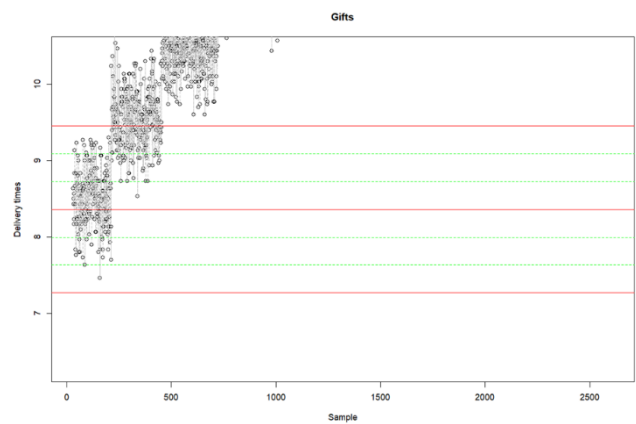
Table 3. S-Chart

The remaining samples were plotted on control charts to determine if any of the processes went out of control. The process control charts were then grouped into out of control, moderately out of control and in control.

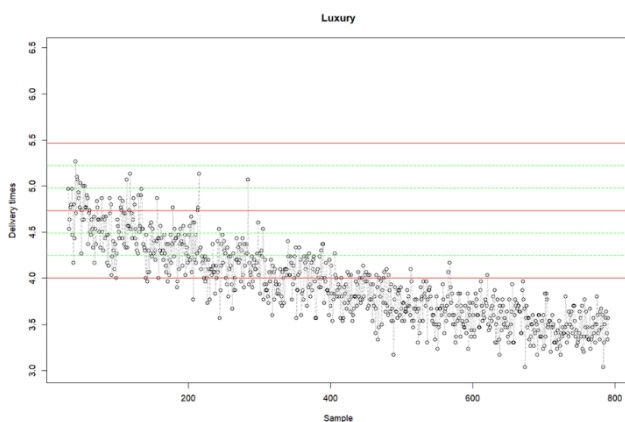
Out of Control charts



16. Household Control Chart



15.. Gifts Control Chart



14.. Luxury Control Chart

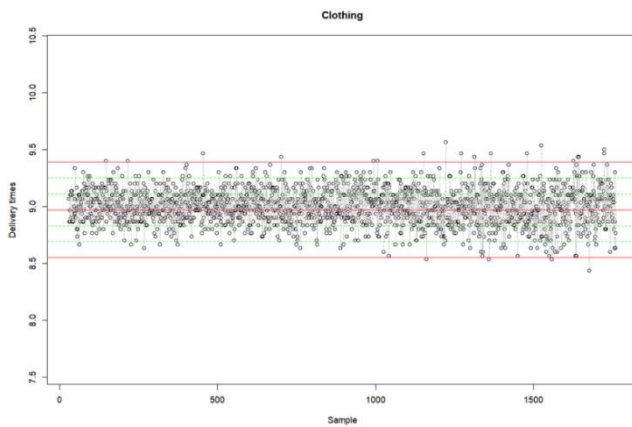
The luxury and the household items as well as the Gift items had charts that went out of control. The luxury delivery items decreased drastically after about 100 samples. The almost immediate decrease in delivery times means the real process control limits were unknown or incorrectly calculated by whoever oversaw the process. The control limits need to be re-evaluated and then stabilised to have more constant delivery times. The person who was in charge with the first samples need to reevaluate their strategy, luckily the delivery times stabilised from about 600 samples and

onward.

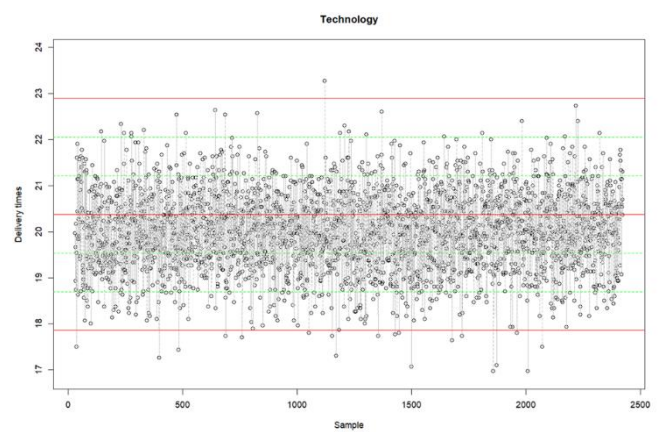
The household delivery times were stable until 500 samples and then it started increasing and going out of control. It is a possibility that the real control limits were used up to then, and a new change or variable came into play that was not correctly implemented into the calculations. The process should be reevaluated and the same process as the 1st half should be implemented once again. The original control limits created a much more stable delivery time.

The gift items almost instantly went out of control and behaved very suspicious. I am not sure if my calculations were correct when plotting the gift samples, due its extreme nature. Regardless of it seems as if the control limits changed on a constant basis. It could also be due to the variety of products that gift items provide. Whatever the case is it should be reevaluated, to try and produce more stable control limits.

Moderately in Control Charts



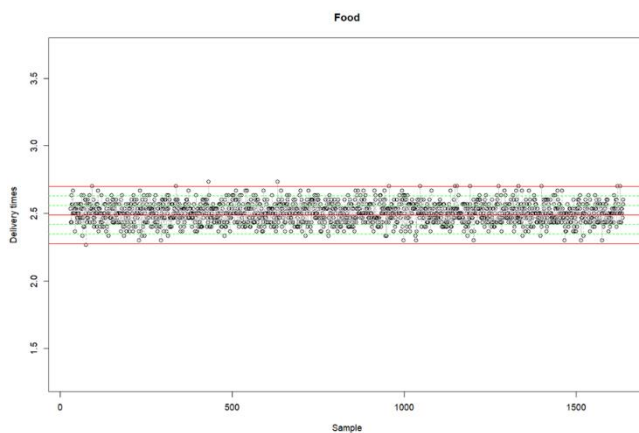
18.. Clothing Control Chart



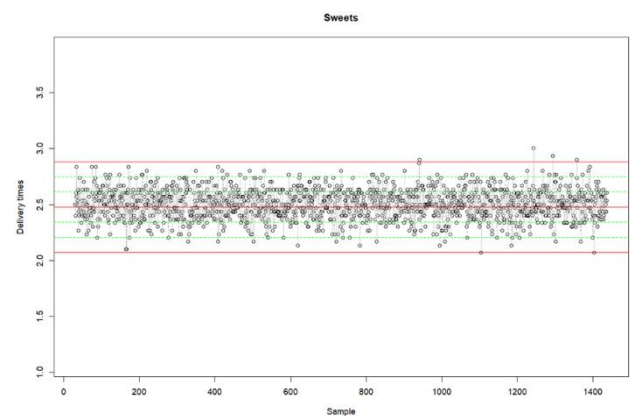
17.. Technology Control Chart

The Clothing and Technology items are both moderately in control, with a very small number of out-of-control samples, there are samples that are above and under the control limit, as it could be due to expired products as well as products that needed to be delivered as fresh items.

In Control Charts



20.. Food Control Chart

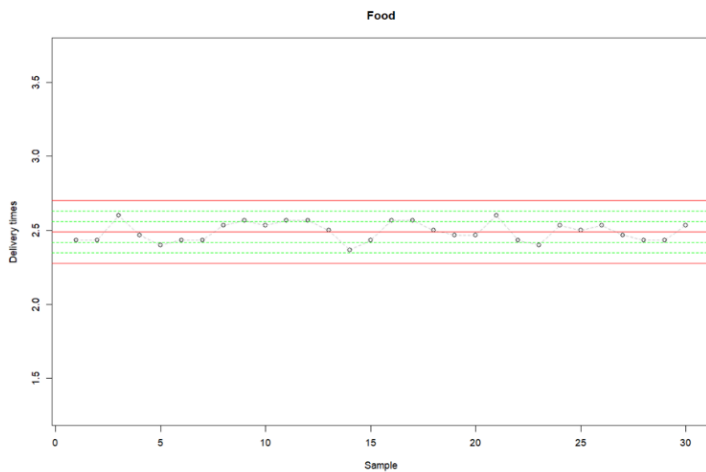


19.. Sweets Control Chart

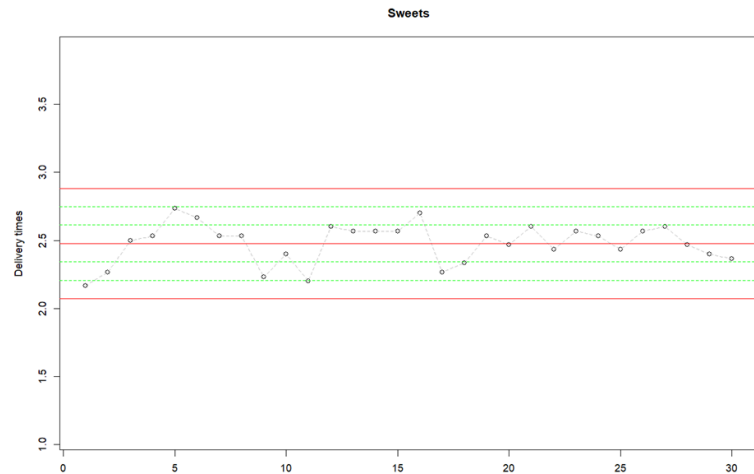
The Food and Sweets delivery times remain in control until the end, with only a few out-of-control samples. They follow the control limits and remain stable; the manager could now focus on making reducing the variance even further to optimize the delivery times to the maximum.

Control Charts with sample size of 30

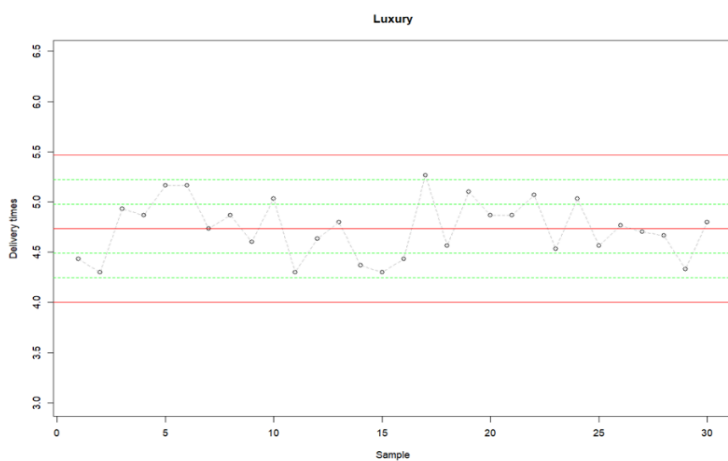
To see what the nature of the samples were the start of the classes a control chart for each item was constructed for only 30 samples, it gave us interesting information that the larger control charts didn't necessarily pick up:



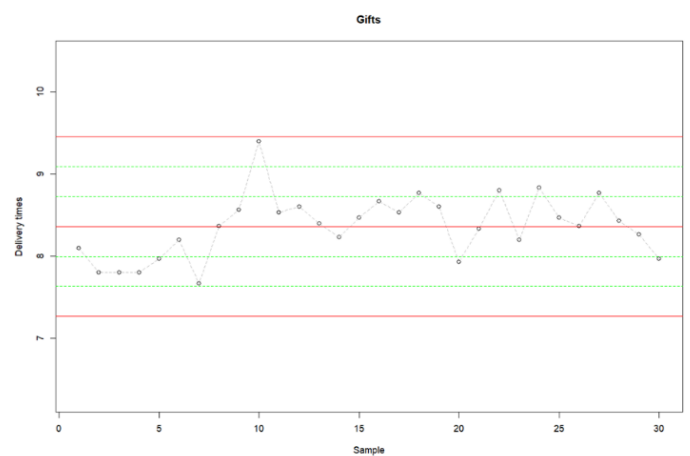
26. Control Chart of smaller sample size (Food)



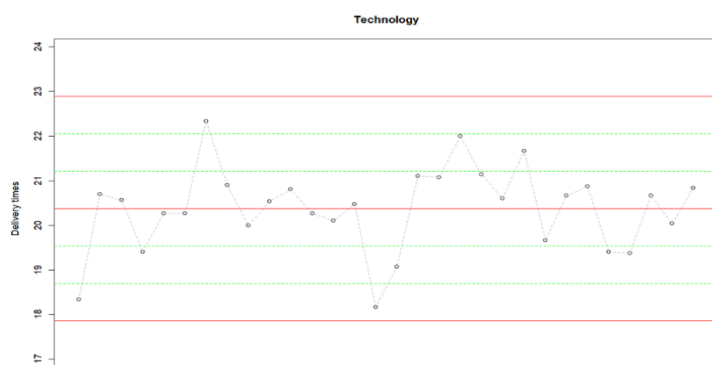
21. Control Chart of smaller sample size (Sweets)



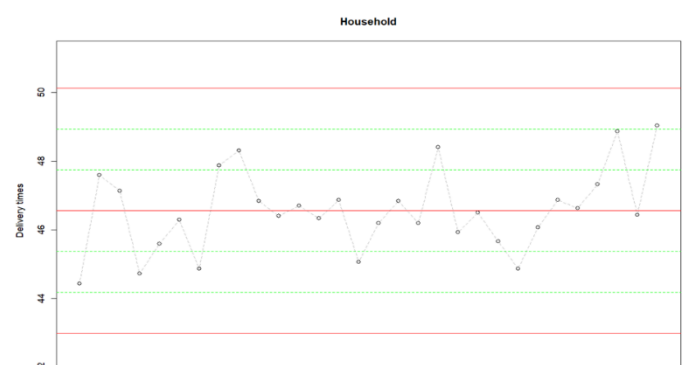
25. Control Chart of smaller sample size (Luxury)



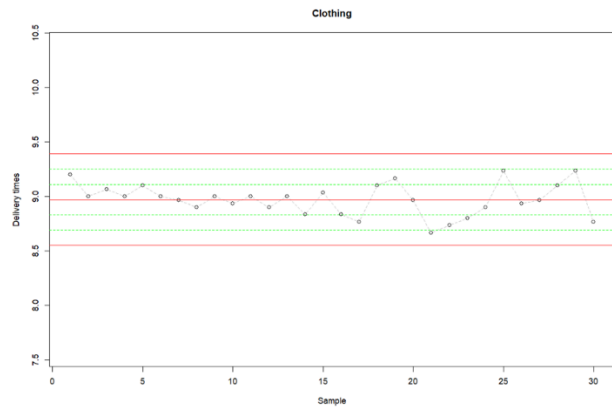
23. Control Chart of smaller sample size (Gifts)



24. Control Chart of smaller sample size (Technology)



22. Control Chart of smaller sample size (Household)



27. Control Chart of smaller sample size (Clothing)

Not one of the items are drastically out of control at this stage but is clear to see that Food and Sweets is the most compact and stable from the start, and that the other are a bit more volatile.

4. Optimizing Delivery Process

4.1) Optimizing the delivery process

The number of samples that were out of control were located by finding the number of samples that were above the UCL and below the LCL, the samples were then sorted by their index.

Class	Number of samples that were out of control	First and last 3 samples that were out of control
Technology	20	7, 368, 453..... 1931, 1979, 2041
Clothing	22	118, 187, 425..... 1647, 1693, 1694
Sweets	6	912, 1074, 1213, 1264, 1328, 1373
Gifts	2296	183, 186, 188..... 2577, 2578, 2579
Household	406	222, 357, 599..... 1305, 1306, 1307
Luxury	453	68, 110, 112..... 759, 760, 761
Food	3	45, 402, 603

Table 4. Out of Control samples

As seen in the control charts is obvious that the gifts, luxury, and household items, still show that they have an alarming number of samples that are out of control.

4.1.b) largest number of consecutive samples within 1 sigma

The length of a sequence is simply the largest difference in starting point. And the following table displays the longest sequence of each class, and what the length of the sequence was:

Class	Sequence	Sequence length
Technology	291:306	15
Sweets	1304:1325	21
Clothing	1645:1668	23
Food	390:405	15
Luxury	637:716	79
Household	1191:1228	37
Gifts	282:296	14

Table 5. Consecutive sample sequence

4.2) Probability of making a Type 1 error

A Type I error occurs when a manager believes the process to be out of control when in fact the process is in control. Calculating the probability to make an error in 4.1A could be found by calculating the probability that a sample is outside of 3*standard deviations from the mean, the following calculation was done in R:

$$P(\text{Type 1 error}) = pnorm(-3) * 2 = 0.002699796$$

If a process stays in control that the likelihood of an out-of-control sample occurring is quite low, 0.2699%.

For the question 4.1.B the probability of making a Type 1 error is the probability of a sample being outside 1 standard deviation from the mean, the following calculation was done in R:

$$P(\text{Type 1 error}) = 1 - \text{pnorm}(0) = 0.5$$

The probability of a sample being inside 1 standard deviation was worked out by using $1 - 0.5 = 0.5$, and using this the probability of the longest sequence that could occur was calculated:

Class	Sequence length	Probability
Technology	15	$0.5^{15} = 3.05 * 10^{-5}$
Sweets	21	$0.5^{21} = 4.768 * 10^{-7}$
Food	15	$0.5^{15} = 3.05 * 10^{-5}$
Gifts	14	$0.5^{14} = 6.103 * 10^{-5}$
Luxury	79	$0.5^{79} = 1.654 * 10^{-24}$
Household	37	$0.5^{37} = 7.276 * 10^{-12}$
Clothing	23	$0.5^{23} = 1.192 * 10^{-7}$

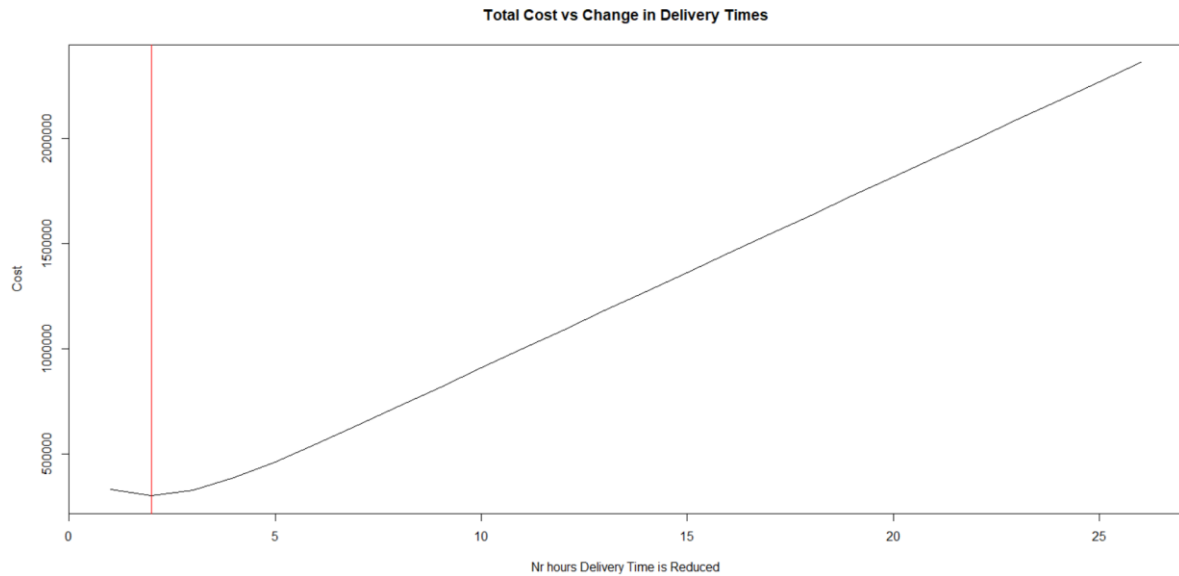
Table 6. Probability of being inside 1 standard deviation

All the probabilities or the given sequence lengths are very small, especially those of 15 and longer, the 79 and 37 sequence length is especially low, having a percentage probability of less than 1%. The management clearly did something right, and it should be investigated, to make sure that the same plans are implemented.

4.3) Delivery time reduction analysis

Given that the company loses R329/item-late/hour in lost sales if they deliver technology items slower than 26 hours, and it costs them R2.5/item/hour to reduce the average time by one hour. It was calculated that the amount of money the delivery times were late, and sales were lost was \$446124.

The average delivery times are 20.01095 hours, and to give a better visual understanding, the total change in cost was plotted against the change in delivery times.



28. Total Cost vs Change in Delivery times

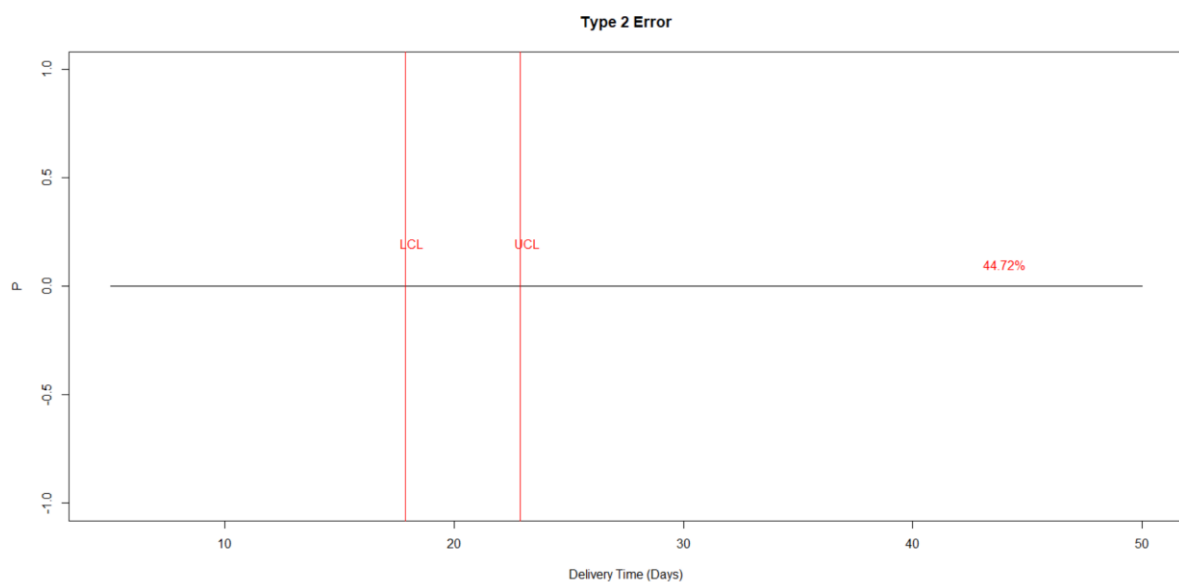
The red line is the optimal number of hours that the delivery time hours need to be reduced by, which will provide us with the minimum possible cost at \$298201, which is almost \$150000 less, than the original amount of money and cost that was started with.

The optimal delivery time was calculated as 24 hours, where the following calculation was made:

$$\text{Optimal delivery} = 26 - (\text{Optimal amount to reduce by}) = 26 - 2 = 24 \text{ hours}$$

4.4) Probability of a Type 2 error

A Type 2 error is when a manager makes the mistake to assume the sample is in control, but in reality, it is out of control. The mean delivery time moves to 23, and thus the Type 2 error is shown as follows:



29. Type 2 error

The red lines are the UCL and LCL for technology delivery times, and the percentage is the probability to make a type 2 error. It is calculated in the following way:

$$p(\text{Type II Error}) = pnorm\left(UCL, 45, \left(\frac{UCL - LCL}{6}\right)\right) - pnorm\left(LCL, 45, \left(\frac{UCL - LCL}{6}\right)\right) = 0.4472$$

5. MANOVA

A Multi-Variable Analysis of variance (MANOVA) was performed on the sales data by using the `manova()` function in R. Then the `summary()` and `summary.aov()` function was used to find what each features influence on delivery time was. All the features were used, and an example of AGE and Price's as well as Month and Day's influence is given in the following figure:

```

      Df    Pillai approx F num Df den Df    Pr(>F)
Delivery.time 1 0.019638   1802.6      2 179980 < 2.2e-16 ***
Residuals    179981
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary.aov(man)
Response AGE :
      Df    Sum Sq Mean Sq F value    Pr(>F)
Delivery.time 1 1020180 1020180    2488 < 2.2e-16 ***
Residuals    179981 73800804      410
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response Price :
      Df    Sum Sq    Mean Sq F value    Pr(>F)
Delivery.time 1 6.8160e+11 6.8160e+11 1575.7 < 2.2e-16 ***
Residuals    179981 7.7853e+13 4.3256e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Table 7. Manova (Age, Price)

```

      Df    Pillai approx F num Df den Df    Pr(>F)
Delivery.time 1 1.7495e-05    1.5744      2 179980 0.2071
Residuals    179981
> summary.aov(man)
Response Month :
      Df    Sum Sq Mean Sq F value    Pr(>F)
Delivery.time 1    21  21.249  1.7813  0.182
Residuals    179981 2146992 11.929

Response Day :
      Df    Sum Sq Mean Sq F value    Pr(>F)
Delivery.time 1    102 101.71  1.3598  0.2436
Residuals    179981 13462535  74.80

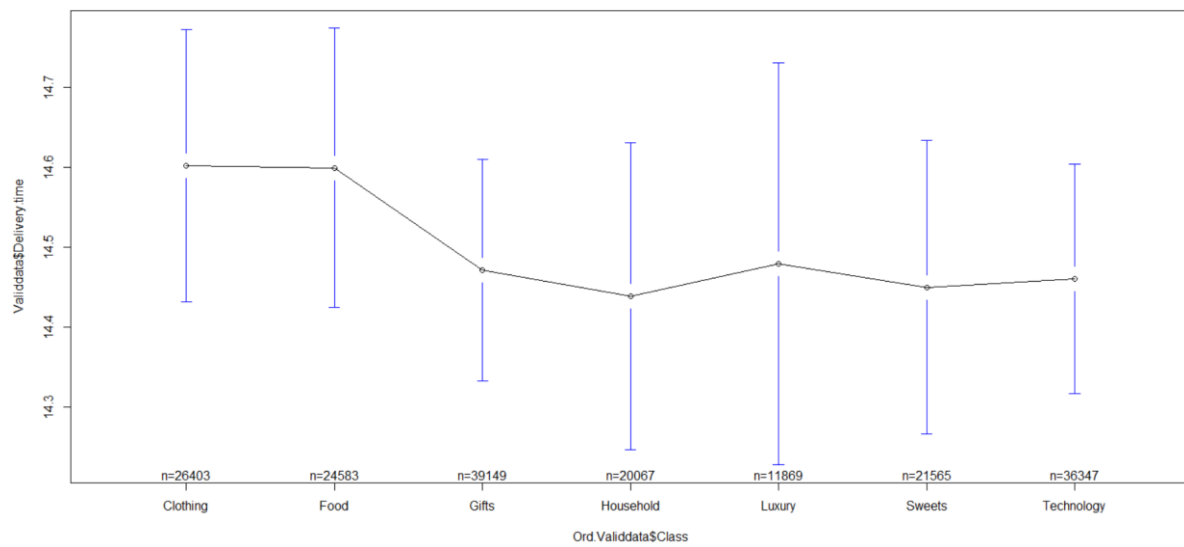
```

Table 8. Manova (Month, Day)

By using the Manova function for all the features, it was found that all the features, except Day and Month had a strong influence with delivery time. The *** shows the strong influence the features have, except Day and month have p-values of 0.2436 and 0.182 respectively. Features such as Class and Why Bought had strong impacts on delivery time.

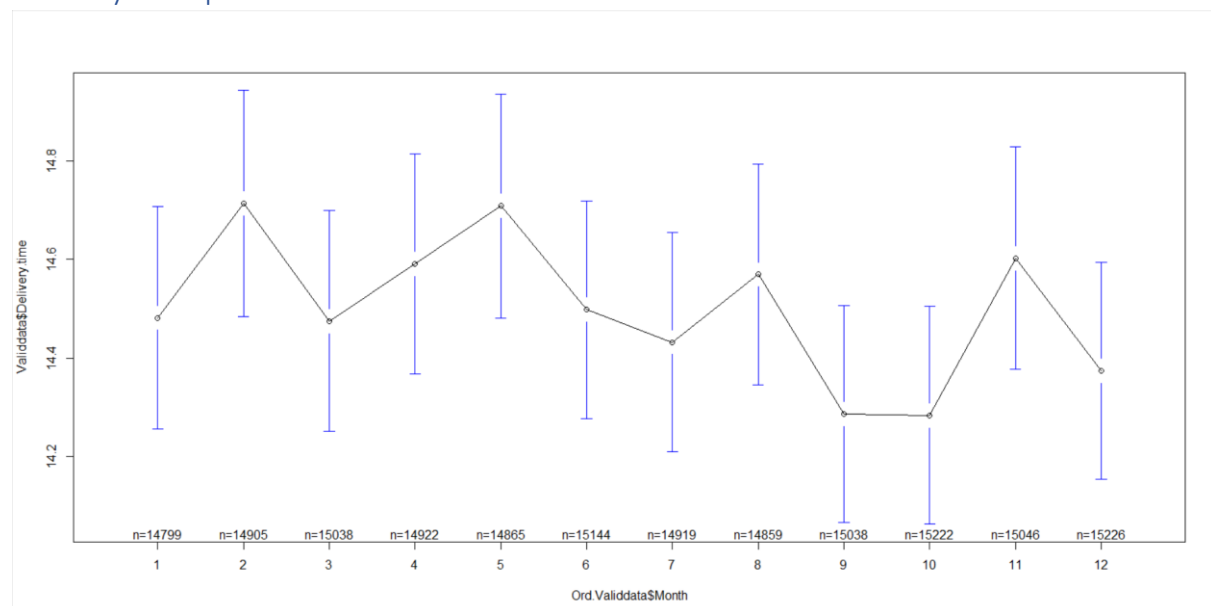
`Plotmeans()` function was used to visually plot the features against each other:

Delivery time per Class



30. Delivery time per Class

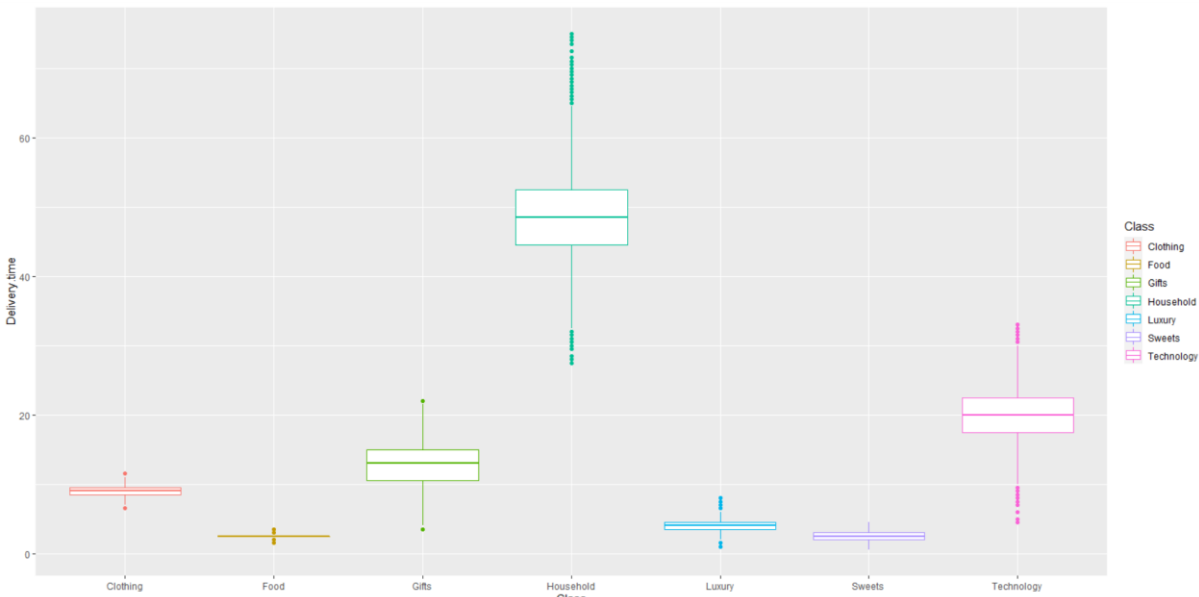
Delivery time per Month



31. Delivery time per Month

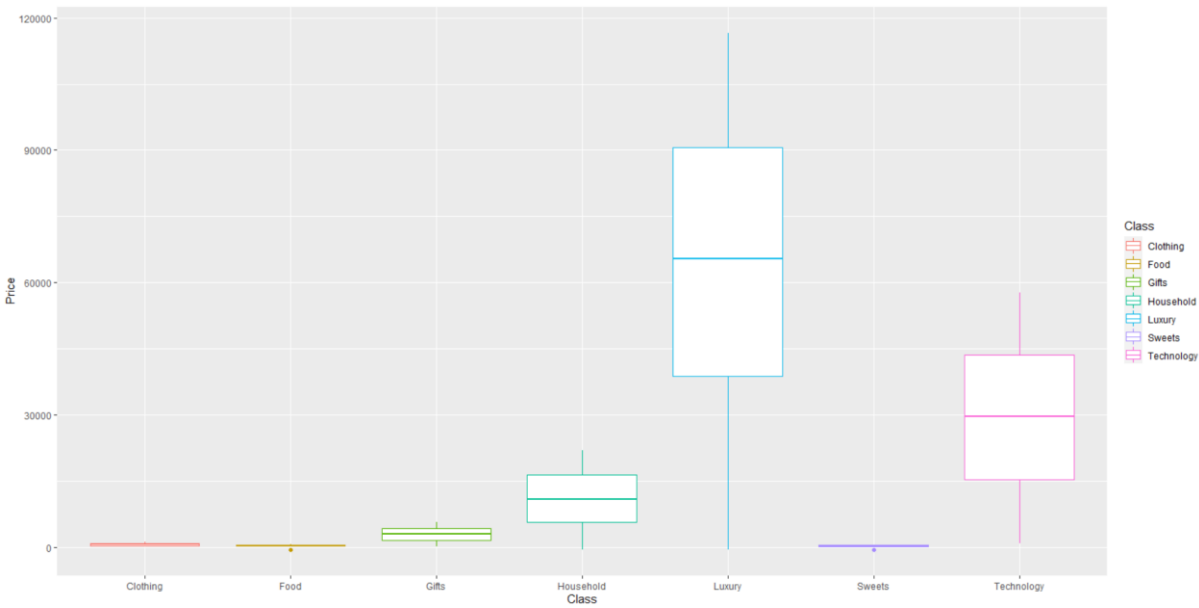
Boxplot charts were then further used, to visualise the features:

Boxplot of Delivery time and Class



32. Boxplot of Delivery time and Class

Boxplot of Price and Class



33. Boxplot of Price and Class

6. Reliability of the service and products

6.1) Problem 6 and 7

The goal of problem 6 and 7 was to illustrate to management the impact that scrap and variance/deviation has on the costs of production.

Problem 6:

For problem 6 the goal was to determine a Taguchi loss function for a scenario given the allowed variance and the scrap cost of a part. In this example the blueprint specification was 0.06 +- 0.04 centimetres, and the scrap cost was \$45 per part.

$$*m = 0.06$$

Taguchi loss function: $L = k(y - m)^2$

Constant k:

$$45 = k(0.04)^2$$

$$k = 28125$$

Thus:

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

Problem 7.a)

For the new scrap cost a new loss function should now be determined. This can be done in the same way as in Problem 6. For the new loss function, the scrap cost is \$35.

Constant K:

$$35 = k(0.04)^2$$

$$k = 21875$$

Thus:

$$L = 21875(y - 0.06)^2$$

Problem 7.b)

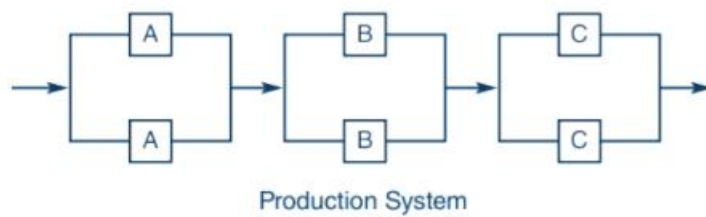
The process deviation from target is then reduces to 0.027cm, the new Taguchi loss is:

$$L = 21875 * 0.027^2 = 15.946875$$

This means that for each part the loss is reduced by \$15.95 if the deviation is reduced to 0.027cm.

6.2) Problem 27

Problem 27 was used to illustrate to management the important need for backup machines



The reliabilities of the machines are as follows:

Machine	Reliability
A	0.85
B	0.92
C	0.90

34. Reliability of machines

To figure out if there was a need for backup machines, the reliability needed to be calculated, with backup machines and without back up machines. The first calculation is simply the reliability of having only 1 machine at each station:

$$reliability = r(A) * r(B) * r(C)$$

$$reliability = 0.85 * 0.92 * 0.9 = 0.7038 = 70.38\%$$

To calculate the reliability for each station when 2 machines are used (backup machine also used), we needed to use this equation for each station:

$$reliability(A) = (1 - (1 - r(A))^2)$$

Therefore, the final reliability was:

$$reliability = (1 - (1 - r(0.85))^2) * (1 - (1 - r(0.92))^2) * (1 - (1 - r(0.9))^2) = 0.9615 = 96.15\%$$

The reliability of the system increases with 25.77%, when the backup machines are used. It increases productivity by almost a quarter.

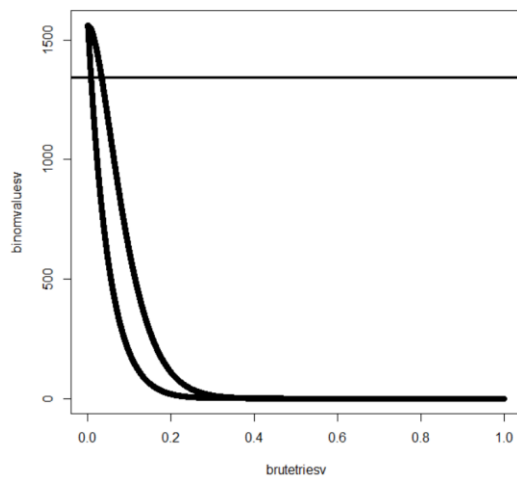
6.3) Vehicle and personnel availability analysis

The number of vehicles and drivers have an impact on the delivery process, and the scenario needs to be solved to provide answers to management. Due to the fact that on a given day the service can either be reliable, or not reliable, the problem is based on binomial probabilities. The approach to solving the problem was to determine on how many days there were more than 16 vehicles available on a day, and to find the binomial probability resulting in that number of days. The most effective approach I found was to use brute force, and work with increments of 0.000001.

By subtracting all the days that we know vehicles were available:

$$Days = 1560 - 190 - 22 - 3 - 1 = 1344$$

The pbinom() function was then used with the increments of 0.000001 to find which probabilities would reach the 1344 days:



35. *pbinom()* of vehicles available

It is visible that the line at 1344 intersects the binomial distributions at two separate locations. These are the possible probabilities that could be used in our calculations. The following equation, is the equation that is used in binomial equations:

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

R was the used to do the actual calculations, and the following figure was the code that was used:

```

#Part 6
#vehicles
pv<-(1344/1560)
n_v<-20

Car.20<-dbinom(20, size = n_v, prob= pv)
Car.19<-dbinom(19, size = n_v, prob= pv)
Car.18<-dbinom(18, size = n_v, prob= pv)
Car.17<-dbinom(17, size = n_v, prob= pv)

p_enough_vehicle<-Car.20+Car.19+Car.18+Car.17
p_enough_vehicle

#drivers
pd<-(1458/1560)
n_d<-21

d.21<-dbinom(21, size = n_d, prob= pd)
d.20<-dbinom(20, size = n_d, prob= pd)
d.19<-dbinom(19, size = n_d, prob= pd)
d.18<-dbinom(18, size = n_d, prob= pd)

p_enough_drivers<-d.21+d.20+d.19+d.18
p_enough_drivers
p_enough_vehicle*p_enough_drivers*365

pv<-(1344/1560)
n_v<-21

Car2.21<-dbinom(21, size = n_v, prob= pv)
Car2.20<-dbinom(20, size = n_v, prob= pv)
Car2.19<-dbinom(19, size = n_v, prob= pv)
Car2.18<-dbinom(18, size = n_v, prob= pv)
Car2.17<-dbinom(17, size = n_v, prob= pv)

p_enough_vehicle_21<-Car2.21+Car2.20+Car2.19+Car2.18+Car2.17
p_enough_vehicle_21

p_enough_vehicle_21*p_enough_drivers*365

```

36. Vehicle and personnel availability code

The probability that there are enough vehicles is **0.7031868** and it stored in the **p_enough_vehicle** variable. The dbinom() function is used to calculate this probability, where all the separate probabilities are added to each other to reach the probability for enough vehicles.

We then calculate the probability for the reliability of the drivers, by using the same dbinom() function to calculate the probabilities. We add all the calculated probabilities together and we have a probability of **0.9553064** that is saved in the **p_enough_drivers** variable.

To calculate the number of days in a year that we can rely on reliable delivery times we multiply the probabilities with each other and then multiply the answer with 365:

$$\text{Reliable delivery days(per year)} = (\text{P}_{\text{enough vehicles}} * \text{P}_{\text{enough drivers}}) * 365 = 245.192$$

With 20 vehicles and 21 drivers the calculations show that we will have 245 reliable days.

21 vehicles:

We now add an extra vehicle to see if it will give us an advantage on our reliability, we used the same calculations as in the previous scenario:

We used the dbinom() function again to calculate the probability of if there is enough vehicles, this time we had an extra vehicle to use, thus the expectance was that we would have a better probability, the forecast was correct and the new probability was **0.8445316** and was saved in the **p_enough_vehicle_21** variable.

We repeat our process and to calculate the number of days in a year that we can rely on reliable delivery times we multiply the probabilities with each other and then multiply the answer with 365:

$$\text{New Reliable delivery days} = (p_{\text{enough}_{\text{vehicle}_{21}}} * p_{\text{enough}_{\text{drivers}}}) * 365 = 294.477$$

Thus, the final answer is, if we add another vehicle we will have 294 reliable days in a year, and it will improve our reliable days a year with almost 50 days.

Conclusion

In conclusion the following was done, the sales data that was supplied, was divided into Validdata and Invaliddata, which could be translated to usable data and unusable data. The usable data was then ordered and sorted in a more approachable manner. The ordered data was then analysed through a variety of different approaches. Firstly, the data was analysed through descriptive statistics and various histograms to get a better visual understanding of the data. Thereafter process control indices techniques as well as control charts and statistical process control techniques were used to further analyse the delivery times and the data. The delivery times was carefully analysed, and recommendations were made to the management. Next was a MANOVA analysis, which looked deeper into the data from another angle and finally separate similar field problems were done to supply the management with new recommendations.

Recommendations that could be made is that Household immediately needs attention to its process control and the delivery times need to be investigated, I would recommend that a closer look be taken at the control limits of the luxury items, and that more realistic control limits be used. The probability that a sequence was within 1 standard deviation, was very small. This should be revaluated and implemented throughout the process.

References

Bedre, R., n.d. *MANOVA using R (with examples and code)*. [Online]

Available at: <https://www.reneshbedre.com/blog/manova.html>

Hernandez, F., 2015. *Data Analysis with R - Exercises*. [Online]

Available at: <http://fch808.github.io/Data-Analysis-with-R-Exercises.html>

Institute of Quality and Reliability, n.d. *Tables of Constants for Control charts*. [Online]

Available at:

<https://web.mit.edu/2.810/www/files/readings/ControlChartConstantsAndFormulae.pdf>

Quality Assurance, 2022. *Statistics*. [Online]

Available at:

https://learn.sun.ac.za/pluginfile.php/3514418/mod_resource/content/1/QA344%20Statistics.pdf

R Coder, n.d. *Binomial distribution in R*. [Online]

Available at: [https://r-coder.com/binomial-distribution-](https://r-coder.com/binomial-distribution-r/#:~:text=The%20binomial%20distribution%20function%20can,and%20a%20probability%20of%20success.)

[r/#:~:text=The%20binomial%20distribution%20function%20can,and%20a%20probability%20of%20success.](https://r-coder.com/binomial-distribution-r/#:~:text=The%20binomial%20distribution%20function%20can,and%20a%20probability%20of%20success.)

Statology, n.d. *How to Fix in R: plot.new has not been called yet*. [Online]

Available at: <https://www.statology.org/r-plot-new-has-not-been-called-yet/>