# QUALITY ASSURANCE

ECSA Project

Luka Spies
23692790

October 2022

# Contents

## Table of figures

# Table of tables

# Introduction

Client data for an online business is give, and the data salesTable2022 includes different variables about sales. The variables include the sales ID, Age, Class, Price, Year, Month, Day, Delivery time and why bought. The data provides the opportunity to investigate certain aspects of the sales and perform statistical analysis on the data.

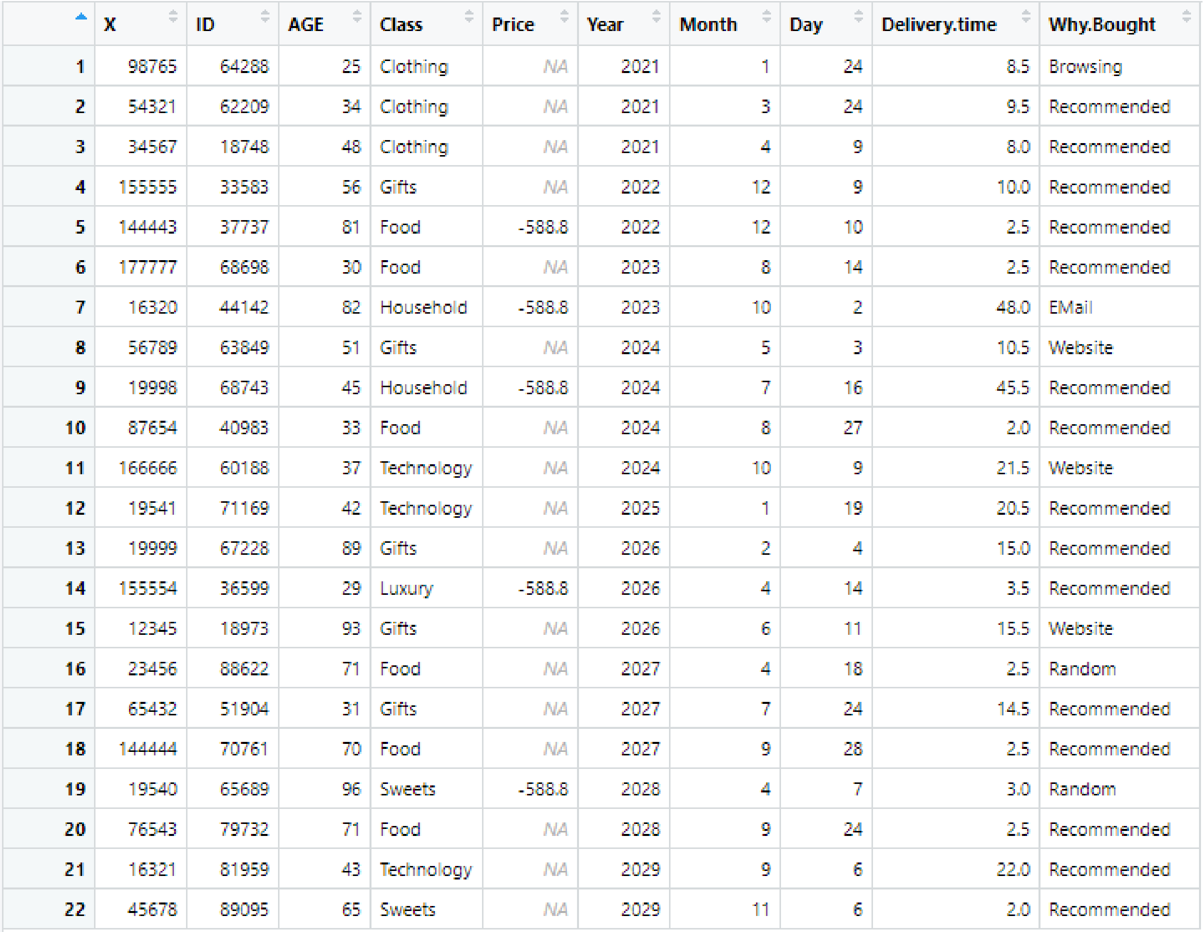
The first part of the report will focus on gaining a deeper understanding of the data as a whole, which will then provide the opportunity to know which relationships to fucus on. From a business and profit perspective it's important to focus on the quality of the service provided.

Next, the aim is to gain a deeper understanding of control limits and the quality of the business's performance. To determine whether a process is in statistical control its first necessary to initialise charts to determine the limits for which a process will be considered in control. These charts are initialised using the first 30 samples. These limits will then be used as a benchmark for future analysis. After performing analysis on all the data, it will be necessary to investigate if the current performance can be improved, and what the cost implications of those decisions are.

## Part 1: Data Wrangling

The original data set had a lot of entries, of which some were invalid. The data was therefore separated into Valid Data and Invalid Data. Invalid data entries were those with NA values or negative price values. These instances were removed from the original data set and formed the Invalid Data set. The remaining data was therefore the Valid Data set.

*Table 1: Invalid Data*

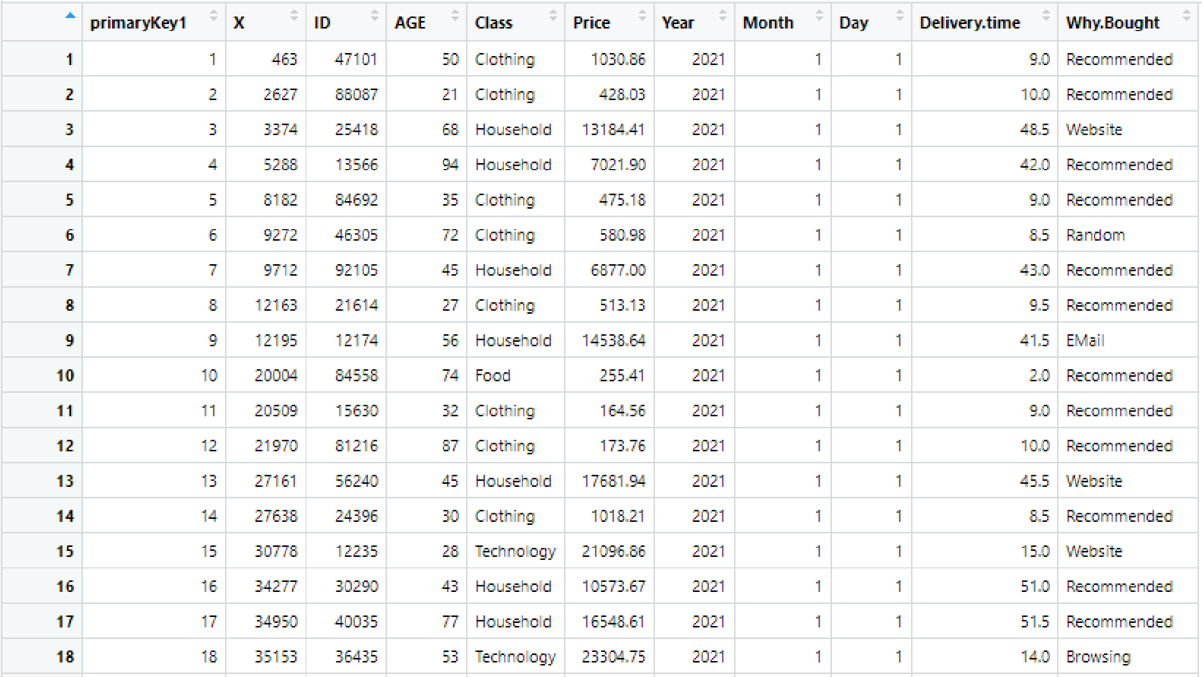


| | X | ID | AGE | Class | Price | Year | Month | Day | Delivery.time | Why.Bought |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 98765 | 64288 | 25 | Clothing | NA | 2021 | 1 | 24 | 8.5 | Browsing |
| 2 | 54321 | 62209 | 34 | Clothing | NA | 2021 | 3 | 24 | 9.5 | Recommended |
| 3 | 34567 | 18748 | 48 | Clothing | NA | 2021 | 4 | 9 | 8.0 | Recommended |
| 4 | 155555 | 33583 | 56 | Gifts | NA | 2022 | 12 | 9 | 10.0 | Recommended |
| 5 | 144443 | 37737 | 81 | Food | -588.8 | 2022 | 12 | 10 | 2.5 | Recommended |
| 6 | 177777 | 68698 | 30 | Food | NA | 2023 | 8 | 14 | 2.5 | Recommended |
| 7 | 16320 | 44142 | 82 | Household | -588.8 | 2023 | 10 | 2 | 48.0 | EMail |
| 8 | 56789 | 63849 | 51 | Gifts | NA | 2024 | 5 | 3 | 10.5 | Website |
| 9 | 19998 | 68743 | 45 | Household | -588.8 | 2024 | 7 | 16 | 45.5 | Recommended |
| 10 | 87654 | 40983 | 33 | Food | NA | 2024 | 8 | 27 | 2.0 | Recommended |
| 11 | 166666 | 60188 | 37 | Technology | NA | 2024 | 10 | 9 | 21.5 | Website |
| 12 | 19541 | 71169 | 42 | Technology | NA | 2025 | 1 | 19 | 20.5 | Recommended |
| 13 | 19999 | 67228 | 89 | Gifts | NA | 2026 | 2 | 4 | 15.0 | Recommended |
| 14 | 155554 | 36599 | 29 | Luxury | -588.8 | 2026 | 4 | 14 | 3.5 | Recommended |
| 15 | 12345 | 18973 | 93 | Gifts | NA | 2026 | 6 | 11 | 15.5 | Website |
| 16 | 23456 | 88622 | 71 | Food | NA | 2027 | 4 | 18 | 2.5 | Random |
| 17 | 65432 | 51904 | 31 | Gifts | NA | 2027 | 7 | 24 | 14.5 | Recommended |
| 18 | 144444 | 70761 | 70 | Food | NA | 2027 | 9 | 28 | 2.5 | Recommended |
| 19 | 19540 | 65689 | 96 | Sweets | -588.8 | 2028 | 4 | 7 | 3.0 | Random |
| 20 | 76543 | 79732 | 71 | Food | NA | 2028 | 9 | 24 | 2.5 | Recommended |
| 21 | 16321 | 81959 | 43 | Technology | NA | 2029 | 9 | 6 | 22.0 | Recommended |
| 22 | 45678 | 89095 | 65 | Sweets | NA | 2029 | 11 | 6 | 2.0 | Recommended |

From the invalid data it can be observed that only the Price feature had invalid entries. A negative price is invalid because the cost of something must always be positive. The NA price entries is invalid because an item must always have a price. A possible reason for these invalid data instances could be from sales or because the business decided to give out of season products aways as gifts.

The original data set had 180 000 entries, and then after removing the 22 invalid entries the Valid Data set had 179 978 entries. The Valid Data set was then ordered by date and given another index to the ordered data. The Valid Data set will be the data used for all future analysis.

*Table 2: Valid Data*

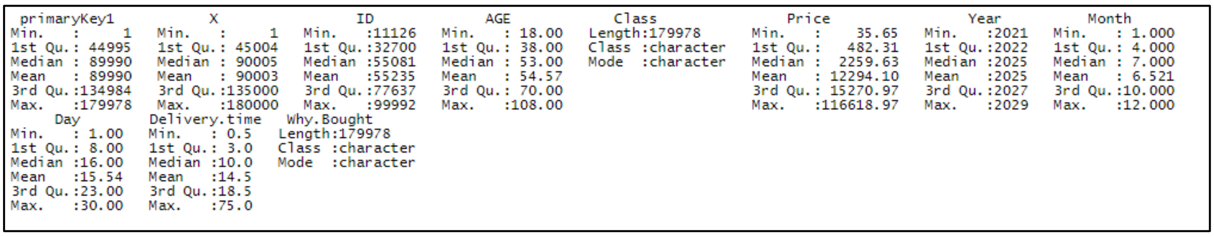


| | primaryKey1 | X | ID | AGE | Class | Price | Year | Month | Day | Delivery.time | Why.Bought |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 1 | 463 | 47101 | 50 | Clothing | 1030.86 | 2021 | 1 | 1 | 9.0 | Recommended |
| **2** | 2 | 2627 | 88087 | 21 | Clothing | 428.03 | 2021 | 1 | 1 | 10.0 | Recommended |
| **3** | 3 | 3374 | 25418 | 68 | Household | 13184.41 | 2021 | 1 | 1 | 48.5 | Website |
| **4** | 4 | 5288 | 13566 | 94 | Household | 7021.90 | 2021 | 1 | 1 | 42.0 | Recommended |
| **5** | 5 | 8182 | 84692 | 35 | Clothing | 475.18 | 2021 | 1 | 1 | 9.0 | Recommended |
| **6** | 6 | 9272 | 46305 | 72 | Clothing | 580.98 | 2021 | 1 | 1 | 8.5 | Random |
| **7** | 7 | 9712 | 92105 | 45 | Household | 6877.00 | 2021 | 1 | 1 | 43.0 | Recommended |
| **8** | 8 | 12163 | 21614 | 27 | Clothing | 513.13 | 2021 | 1 | 1 | 9.5 | Recommended |
| **9** | 9 | 12195 | 12174 | 56 | Household | 14538.64 | 2021 | 1 | 1 | 41.5 | EMail |
| **10** | 10 | 20004 | 84558 | 74 | Food | 255.41 | 2021 | 1 | 1 | 2.0 | Recommended |
| **11** | 11 | 20509 | 15630 | 32 | Clothing | 164.56 | 2021 | 1 | 1 | 9.0 | Recommended |
| **12** | 12 | 21970 | 81216 | 87 | Clothing | 173.76 | 2021 | 1 | 1 | 10.0 | Recommended |
| **13** | 13 | 27161 | 56240 | 45 | Household | 17681.94 | 2021 | 1 | 1 | 45.5 | Website |
| **14** | 14 | 27638 | 24396 | 30 | Clothing | 1018.21 | 2021 | 1 | 1 | 8.5 | Recommended |
| **15** | 15 | 30778 | 12235 | 28 | Technology | 21096.86 | 2021 | 1 | 1 | 15.0 | Website |
| **16** | 16 | 34277 | 30290 | 43 | Household | 10573.67 | 2021 | 1 | 1 | 51.0 | Recommended |
| **17** | 17 | 34950 | 40035 | 77 | Household | 16548.61 | 2021 | 1 | 1 | 51.5 | Recommended |
| **18** | 18 | 35153 | 36435 | 53 | Technology | 23304.75 | 2021 | 1 | 1 | 14.0 | Browsing |

# Part 2: Descriptive statistics

## Analyse data set

Using the summary() function as a starting point to get a summative overview of the data gives the following result:



```
  primaryKey1          X                 ID             AGE            Class               Price              Year          Month       
 Min.   :     1   Min.   :     1   Min.   :11126   Min.   : 18.00   Length:179978      Min.   :    35.65   Min.   :2021   Min.   : 1.000  
 1st Qu.: 44995   1st Qu.: 45004   1st Qu.:32700   1st Qu.: 38.00   Class :character   1st Qu.:   482.31   1st Qu.:2022   1st Qu.: 4.000  
 Median : 89990   Median : 90005   Median :55081   Median : 53.00   Mode  :character   Median :  2259.63   Median :2025   Median : 7.000  
 Mean   : 89990   Mean   : 90003   Mean   :55235   Mean   : 54.57                      Mean   : 12294.10   Mean   :2025   Mean   : 6.521  
 3rd Qu.:134984   3rd Qu.:135000   3rd Qu.:77637   3rd Qu.: 70.00                      3rd Qu.: 15270.97   3rd Qu.:2027   3rd Qu.:10.000  
 Max.   :179978   Max.   :180000   Max.   :99992   Max.   :108.00                      Max.   :116618.97   Max.   :2029   Max.   :12.000  
      Day         Delivery.time   Why.Bought       
 Min.   : 1.00   Min.   : 0.5   Length:179978     
 1st Qu.: 8.00   1st Qu.: 3.0   Class :character  
 Median :16.00   Median :10.0   Mode  :character  
 Mean   :15.54   Mean   :14.5                     
 3rd Qu.:23.00   3rd Qu.:18.5                     
 Max.   :30.00   Max.   :75.0                     
```

Some features, such as the primary key, x and ID isn't valuable to gaining an understanding of the data since these are merely admin entries to keep track of the number of sales taking place.

The valuable information can be summed up as follows:

*Table 3: Summary of features*

| | AGE | PRICE | YEAR | MONTH | DAY | DELIVERY TIME |
|---|---|---|---|---|---|---|
| Minimum | 18.00 | 35.65 | 2021 | 1.00 | 1.00 | 0.5 |
| 1st Quadrant | 38.00 | 482.31 | 2022 | 4.00 | 8.00 | 3.0 |
| Median | 53.00 | 2259.63 | 2025 | 7.00 | 16.00 | 10.0 |
| Mean | 54.57 | 12294.10 | 2025 | 6.521 | 15.54 | 14.5 |
| 3rd Quadrant | 70.00 | 15270.97 | 2027 | 10.00 | 23.00 | 18.5 |
| Maximum | 108.00 | 116618.77 | 2029 | 12.00 | 30.00 | 75.0 |

Bar plots were used to get a better understanding of the categorical features. Each bar represents one categorical value, and the length of the bar represents the number of entries in that category.

**Why Bought:**

From this bar plot it can be seen that the majority of items were bought because of recommendations. Second and third most were Website and Browsing, while the least items were bought because of spam. This is helpful for the business to know because it's clear that good service and good products leads to people recommending the business to friends and family. Additionally, the business doesn't have to spend a lot of money on sending out spam because it doesn't give good results in terms of why people buy products.

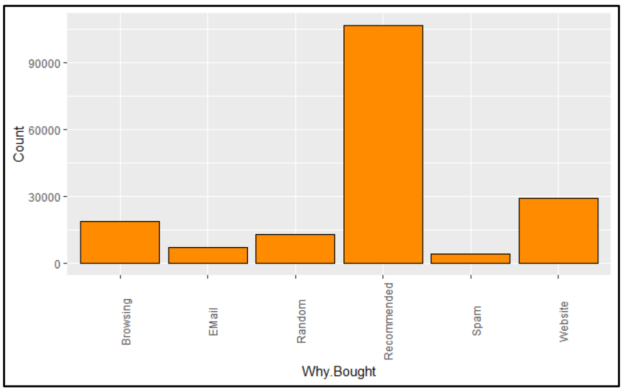


*Figure 1: Bar plot of Why Bought*

**Class**

The bar plot below shows the most popular and less popular class of products. The most items sold belong to the Gifts class, and technology is the second most popular class. Luxury items are the least popular class of items.

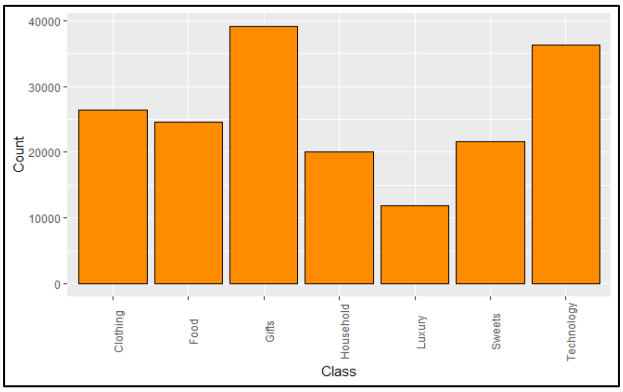


*Figure 2: Bar plot of Class*

Histograms were plotted to gather a better understanding of the numeric features. A histogram is a chart that plots the distribution of a numeric variable's values as a series of bars. “Each bar typically covers a range of numeric values called a bin or class; a bar's height indicates the frequency of data points with a value within the corresponding bin.” (Hessing, 2019)

## Age

The age is slightly skewed to the right. A distribution skewed to the right is said to be positively skewed. This kind of distribution has a large number of occurrences in the lower value cells and few in the upper value cells. A skewed distribution can result when data is gathered from a system with has a boundary such as zero. In other words, all the collected data has values greater than zero. This is true for age because all ages are higher than zero, and most instances are in the lower value cells because the majority of this businesses’ customers are younger than 60.

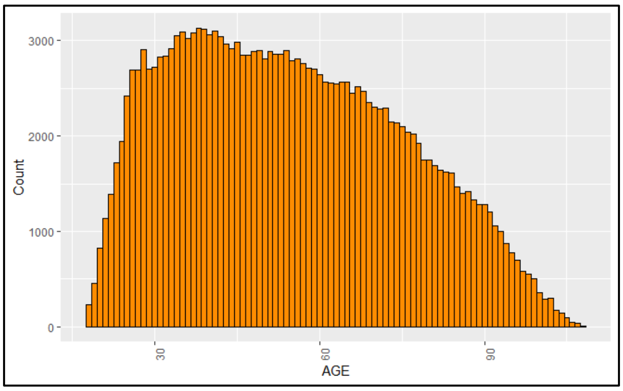


*Figure 3: Histogram of Age*

## Price

Price is skewed right, with a very long tail, indicating that most of the items sold have a lower price but there are some extreme outliers. This is because the business sells a wide variety of products, from food to luxury items, and these products vary a lot in price. To get a better understanding of the price distribution it’s beneficial to look at the price per class. The boxplot showing the Price for each specific class provides more insight into Price. From the Boxplot it’s clear that luxury items are the most expensive, since the mean price of luxury items are the highest and the upper tail of luxury is also the highest. The mean price of technology is second highest. Clothing, Food and Sweets have the lowest process.

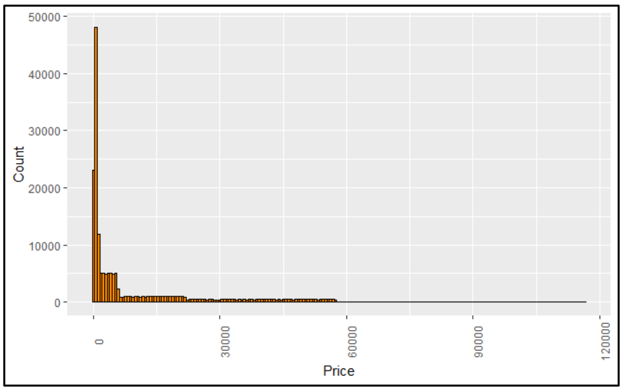


*Figure 4: Histogram of Price*

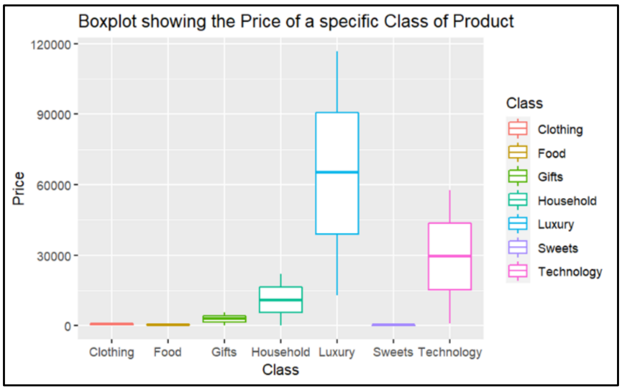


*Figure 5: Boxplot Price per Class*

## Delivery time

The histogram of all the delivery times is random, with a few peaks. The random distribution has no apparent pattern and describes a distribution that has several modes. This is because several sources of variation have been combined, so it's necessary to analyse them separately. A random distribution often means there are too many classes, and in this scenario it's not beneficial to plot all the classes together.

Once all the classes are separated into their own histograms it becomes clear where the different peaks come from.

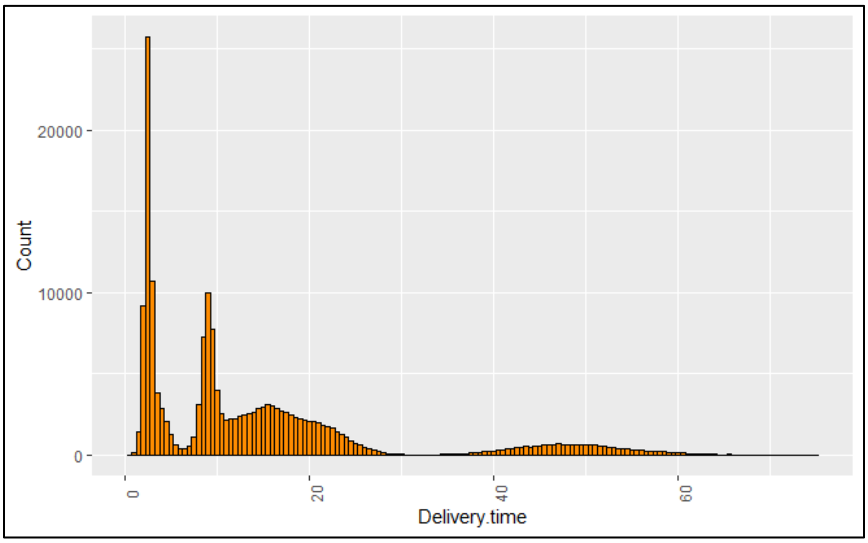


*Figure 6: Histogram of Delivery time*

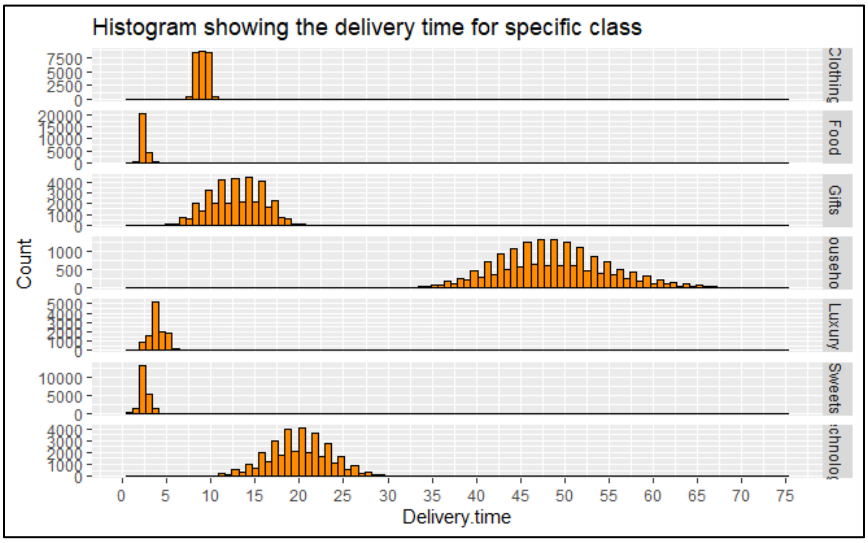
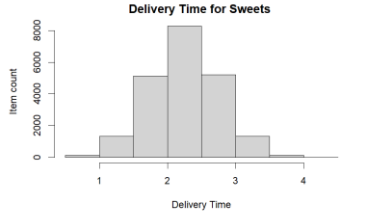
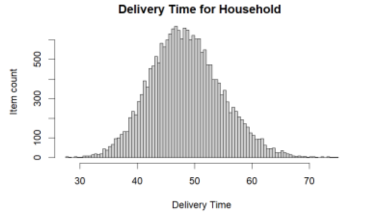
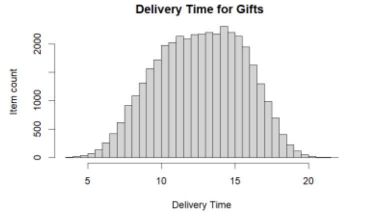
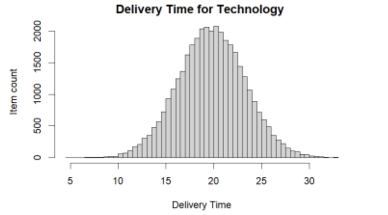
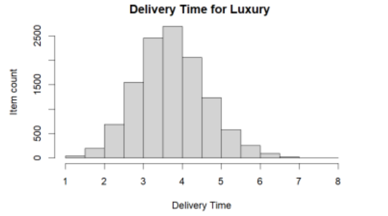
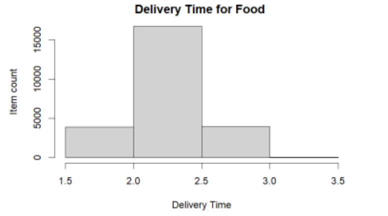
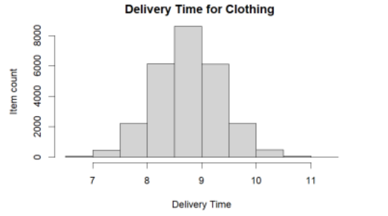


*Figure 7: Histogram of Delivery time per class*

*Table 4: Delivery times per Class*

| Class | Graph | Comment |
|---|---|---|
| Sweets |  | Normally distributed with a mean of 2.501206 |
| Household |  | Normally distributed with a mean of 48.71956 |
| Gifts |  | Normally distributed with a mean of 12.89055 |
| Technology |  | Normally distributed with a mean of 20.01095 |
| Luxury |  | Normally distributed with a mean of 3.97152 |
| Food |  | Normally distributed with a mean of 2.502014 |
| Clothing |  | Normally distributed with a mean of 8.999527 |

The following histogram summarises the mean delivery time per class, and here it can be observed that items belonging to the Household class takes the longest to deliver. The second longest mean delivery time and the shortest mean delivery time is for Food items. This is likely because food items are perishable so it's important to deliver these items as fast as possible.

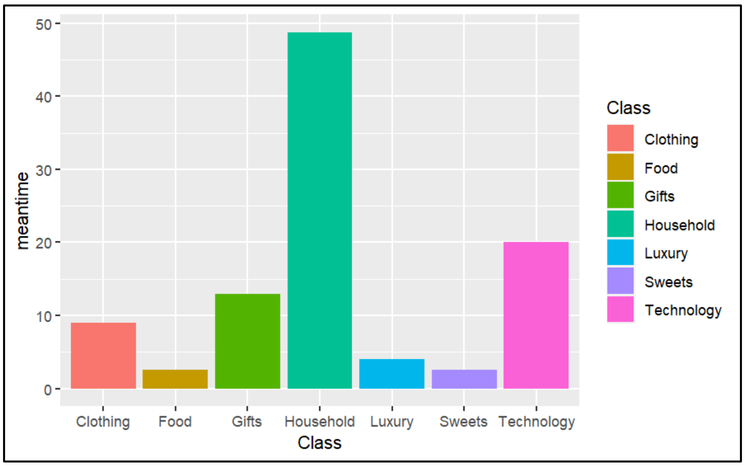


*Figure 8: Histogram of mean delivery times*

The boxplot below summarises the distribution of age per class. Here it can be observed that food items have the highest mean age, while clothing and luxury items have the smallest mean age. This figure also indicates that all classes have similar minimum and maximum ages, so although the mean age per class is different, the range of ages per class is the same.

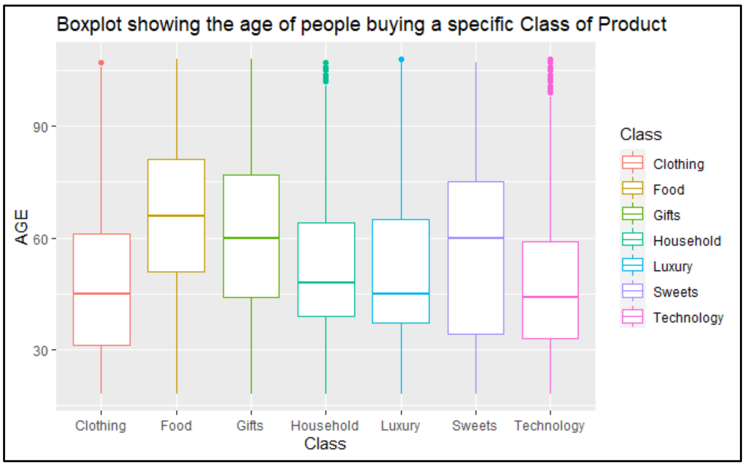


*Figure 9: Boxplot of age per class*

The histogram below summarises the reason bought by class. This figure once again shows that most items are bought because of recommendations. Browsing has the biggest impact on items belonging to the clothing and gifts classes. Luxury items are never bought because of emails, random or spam.

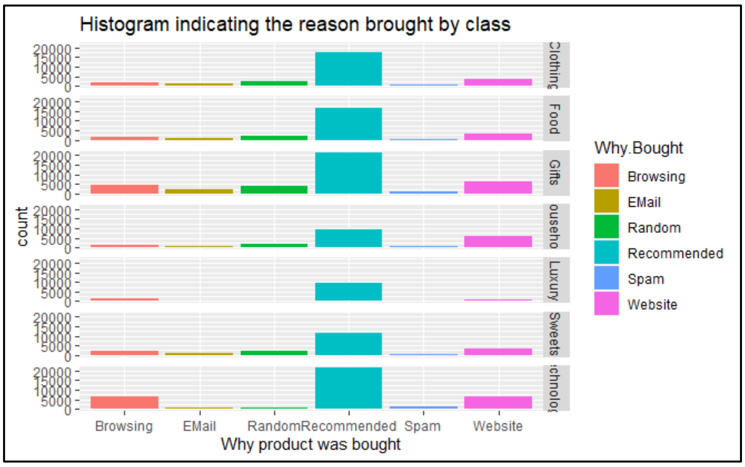
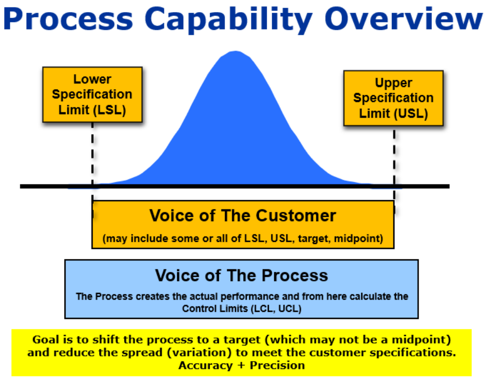


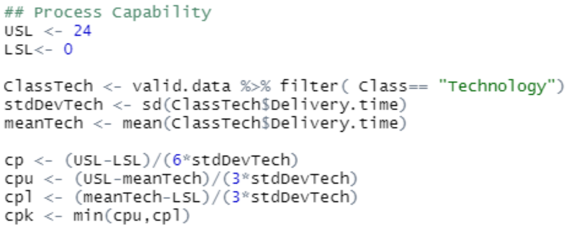
*Figure 10: Histogram on why bought per class*

## Process capability indices

Process Capability may be defined as the ability of a process to meet specifications. Capability potential (Cp) and Capability performance (Cpk) illustrate a process's ability to meet specifications. The Cp ratio shows how well the process spread (expressed as six standard deviations of the process) fits into the specification range. This measurement is determined by dividing the specification limit (voice of the customer) by the process spread (voice of the process) ie VOC/VOP. The Cpk ratio also measures what Cp does and in addition it measures how close the process mean is to the target value of the specification.


Process Capability Overview
Lower Specification Limit (LSL)
Upper Specification Limit (USL)
Voice of The Customer
(may include some or all of LSL, USL, target, midpoint)
Voice of The Process
The Process creates the actual performance and from here calculate the Control Limits (LCL, UCL)
Goal is to shift the process to a target (which may not be a midpoint) and reduce the spread (variation) to meet the customer specifications.
Accuracy + Precision

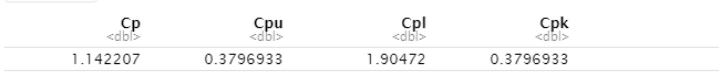
To calculate the indices, we select a lower limit of LSL = 0. This is logical as the minimum days to deliver is 0. The upper limit has been set at USL = 48.



```
## Process Capability
USL <- 24
LSL<- 0

ClassTech <- valid.data %>% filter( Class== "Technology")
stdDevTech <- sd(ClassTech$Delivery.time)
meanTech <- mean(ClassTech$Delivery.time)

cp <- (USL-LSL)/(6*stdDevTech)
cpu <- (USL-meanTech)/(3*stdDevTech)
cpl <- (meanTech-LSL)/(3*stdDevTech)
cpk <- min(cpu,cpl)
```



```
         Cp          Cpu          Cpl          Cpk
      <dbl>        <dbl>        <dbl>        <dbl>
   1.142207    0.3796933      1.90472    0.3796933
```

The Process Capability Index (Cp) is equal to 1.142. A Cp of less than one indicates that the process spread is greater than the specification, which means that some of the data lies outside of the specification. The calculated Cpk value is 0.380, and a Cpk values of less than one is considered poor, and the process is not capable. Since the Cpk is less than the Cp it can be concluded that the distribution of the data is not perfectly centred.

## Part 3: Statistical process control

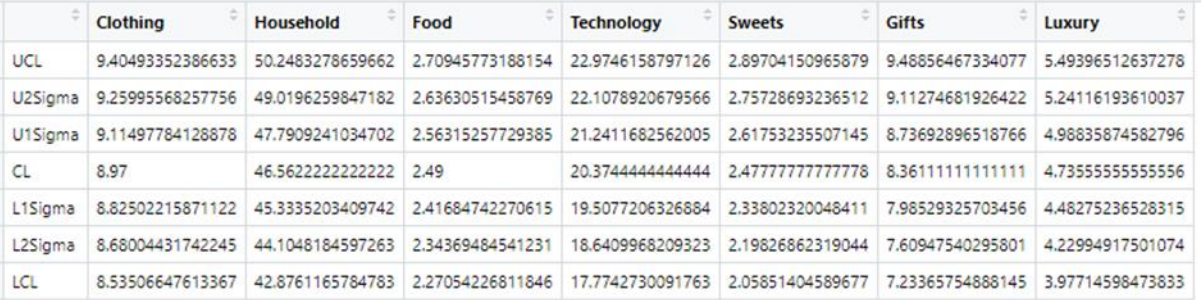
Statistical process control (SPC) is defined as the use of statistical techniques to control a process or production method. SPC tools and procedures can help you monitor process behaviour, discover issues in internal systems, and find solutions for production issues.

## 3.1 Initializing Control Charts using First 30 samples

To calculate the initial control limits, the first (chronological) 25 samples of 15 instances each were used to determine centre lines, outer control limits, the 2-sigma-control limits and the 1-sigma-control limits for the delivery times for each class. Tables below show the control limits calculated for each class based on the first 25 samples:

X chart

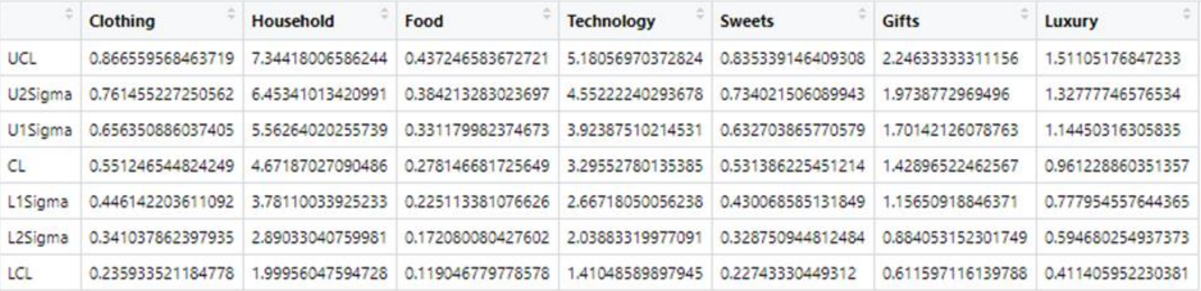
*Table 5: X chart*



| | Clothing | Household | Food | Technology | Sweets | Gifts | Luxury |
|---|---|---|---|---|---|---|---|
| UCL | 9.40493352386633 | 50.2483278659662 | 2.70945773188154 | 22.9746158797126 | 2.89704150965879 | 9.48856467334077 | 5.49396512637278 |
| U2Sigma | 9.25995568257756 | 49.0196259847182 | 2.63630515458769 | 22.1078920679566 | 2.75728693236512 | 9.11274681926422 | 5.24116193610037 |
| U1Sigma | 9.11497784128878 | 47.7909241034702 | 2.56315257729385 | 21.2411682562005 | 2.61753235507145 | 8.73692896518766 | 4.98835874582796 |
| CL | 8.97 | 46.5622222222222 | 2.49 | 20.3744444444444 | 2.47777777777778 | 8.36111111111111 | 4.73555555555556 |
| L1Sigma | 8.82502215871122 | 45.3335203409742 | 2.41684742270615 | 19.5077206326884 | 2.33802320048411 | 7.98529325703456 | 4.48275236528315 |
| L2Sigma | 8.68004431742245 | 44.1048184597263 | 2.34369484541231 | 18.6409968209323 | 2.19826862319044 | 7.60947540295801 | 4.22994917501074 |
| LCL | 8.53506647613367 | 42.8761165784783 | 2.27054226811846 | 17.7742730091763 | 2.05851404589677 | 7.23365754888145 | 3.97714598473833 |

SD chart

*Table 6: SD chart*



| | Clothing | Household | Food | Technology | Sweets | Gifts | Luxury |
|---|---|---|---|---|---|---|---|
| UCL | 0.866559568463719 | 7.34418006586244 | 0.437246583672721 | 5.18056970372824 | 0.835339146409308 | 2.24633333311156 | 1.51105176847233 |
| U2Sigma | 0.761455227250562 | 6.45341013420991 | 0.384213283023697 | 4.55222240293678 | 0.734021506089943 | 1.9738772969496 | 1.32777746576534 |
| U1Sigma | 0.656350886037405 | 5.56264020255739 | 0.331179982374673 | 3.92387510214531 | 0.632703865770579 | 1.70142126078763 | 1.14450316305835 |
| CL | 0.551246544824249 | 4.67187027090486 | 0.278146681725649 | 3.29552780135385 | 0.531386225451214 | 1.42896522462567 | 0.961228860351357 |
| L1Sigma | 0.446142203611092 | 3.78110033925233 | 0.225113381076626 | 2.66718050056238 | 0.430068585131849 | 1.15650918846371 | 0.777954557644365 |
| L2Sigma | 0.341037862397935 | 2.89033040759981 | 0.172080080427602 | 2.03883319977091 | 0.328750944812484 | 0.884053152301749 | 0.594680254937373 |
| LCL | 0.235933521184778 | 1.99956047594728 | 0.119046779778578 | 1.41048589897945 | 0.22743330449312 | 0.611597116139788 | 0.411405952230381 |

From the data in the tables above, control charts could be created for each class. Control charts shows where each sample case falls with respect to the control limits. In the figures below, the red line is CL, the blue lines are the +- 1 sigma limits, the yellow lines are the +- 2 sigma limits, and the green lines are UCL and LCL.

## Clothing

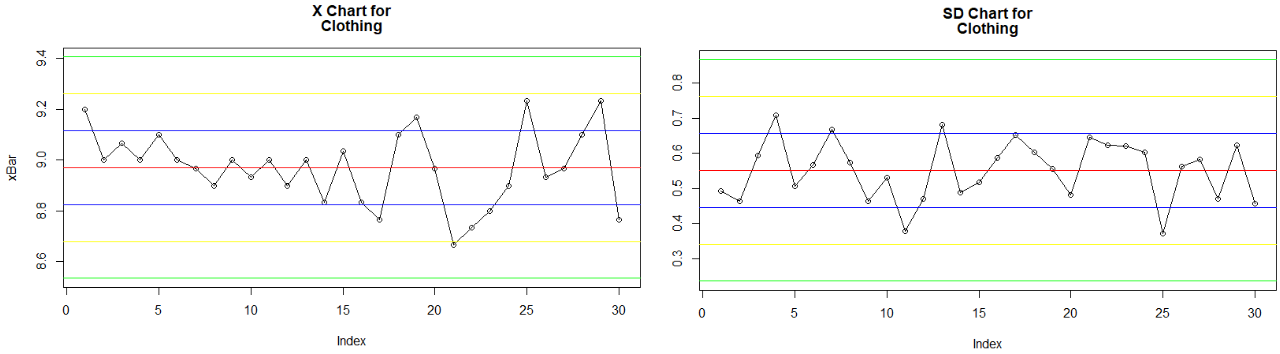


*Figure 11: Clothing X and SD charts*

For clothing all points are within the 3-sigma limit, so sample data for clothing is in control. Another thing to notice is that majority of the points fall between the +- 1 sigma lines, which indicated that this process is tightly controlled.

## Household

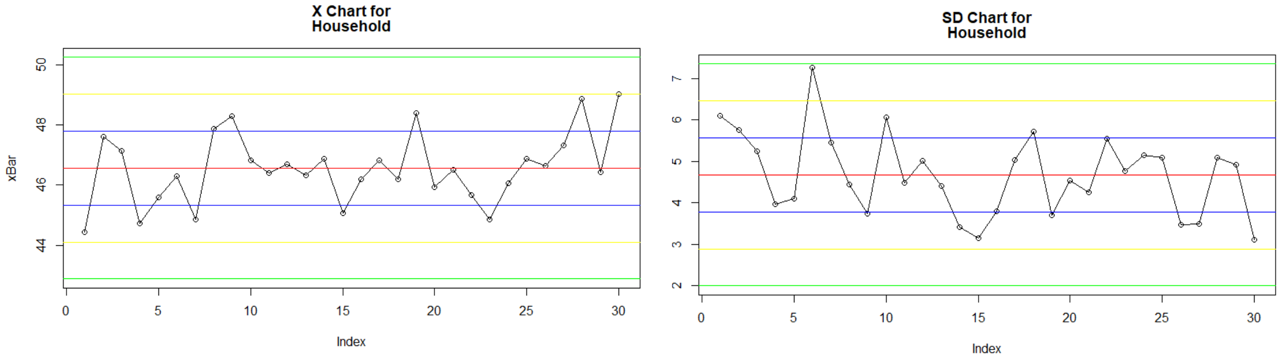


*Figure 12: Household X and SD charts*

For household it can be seen that all points are within the 3-sigma limit, so household sample data is in statistical control. Another thing to notice is that majority of the points fall between the +- 1 sigma lines, which indicated that this process is tightly controlled.

## Food

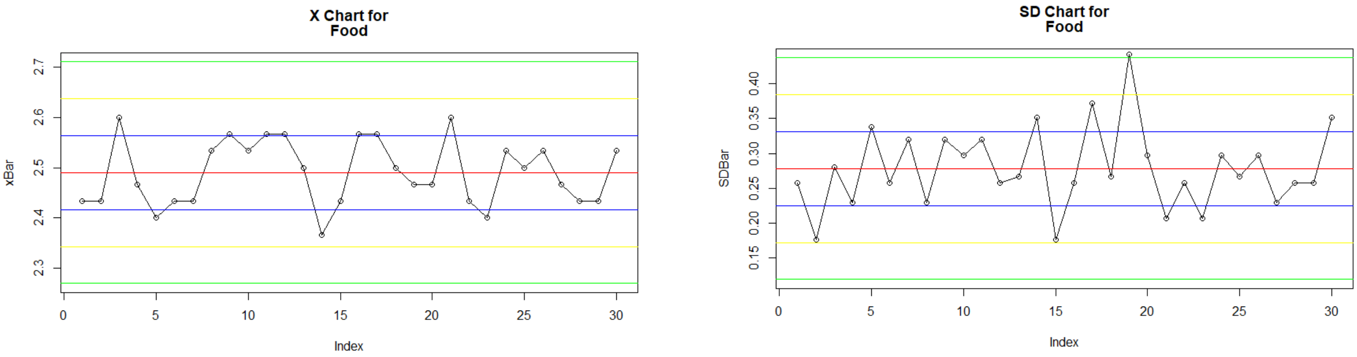


*Figure 13: Food X and SD charts*

The X bar chart shows that no points are out of control which means the process mean is in control. On the X chart majority of the points fall between the +- 1 sigma lines, which indicated that this process is tightly controlled. The S chart has one point outside control, which is subgroup 19.

## Technology

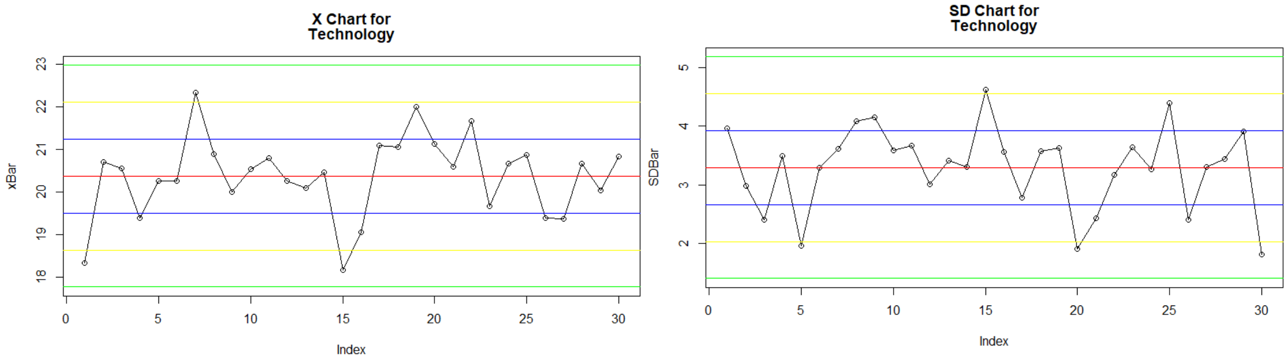


*Figure 14: Technology X and SD charts*

## Sweets

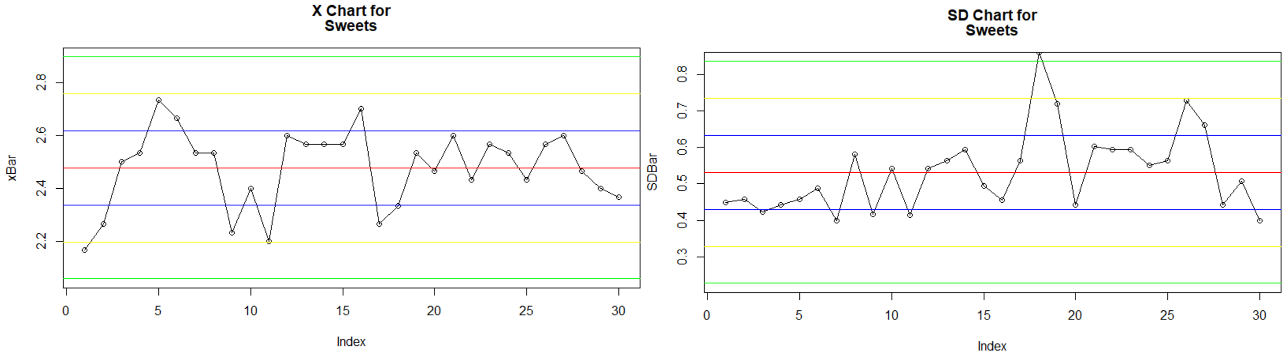


*Figure 15: Sweets X and SD charts*

The S chart shows that there is a point out of the limits and thus subgroup 18 is out of control points. The X bar chart shows that no points are out of control which means the process mean is in control.

## Gifts

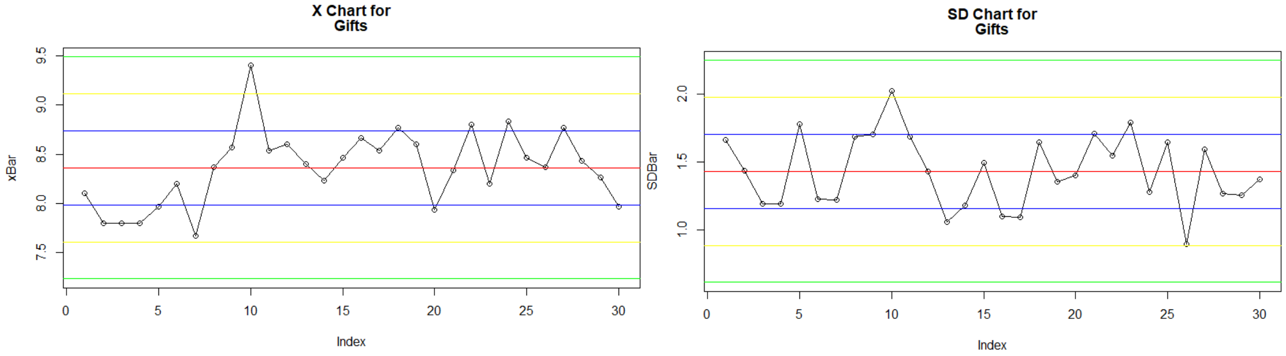


*Figure 16: Gifts X and SD charts*

The S chart shows no points are out of control. This means that the process variation is in control. The X bar chart shows that no points are out of control which means the process mean is in control.

### Luxury

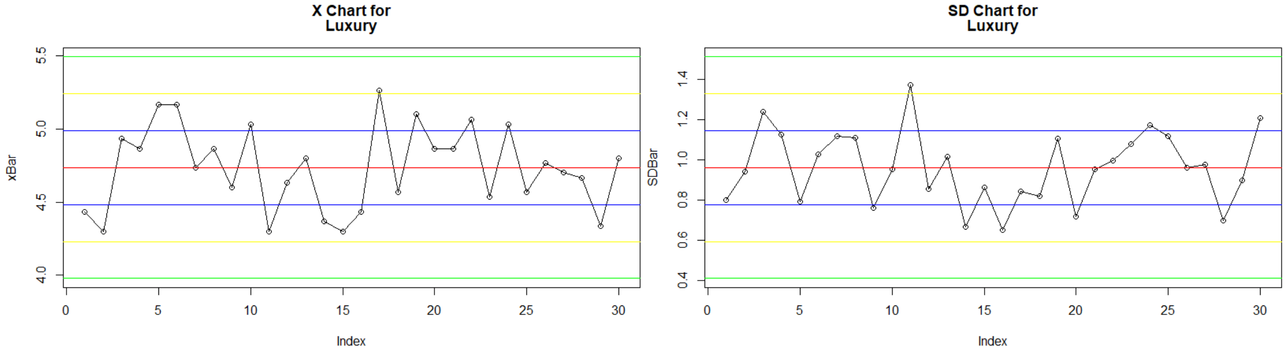


*Figure 17: Luxury X and SD charts*

The S chart shows no points are out of control. This means that the process variation is in control. The X bar chart shows that no points are out of control which means the process mean is in control.

## Part 3.2 Continuation of samples

After setting up control charts, the rest of the data is plotted onto the control charts to see whether the delivery processes for each class continues to either stay in control, come under control, go out of control, or stay out of control.

In the figures below any points outside the control limits will be in purple.

### Clothing

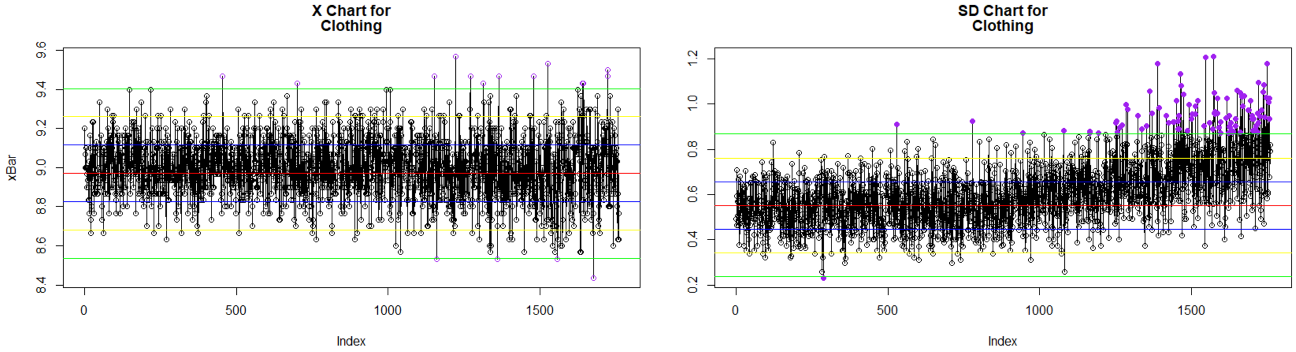


*Figure 18: Clothing extended X and SD charts*

Clothing has stayed in control. There only seems to be a very few select instances where the sample means fall outside the control limits. It can also be seen that a vast majority of samples fall within 2-sigma and a fair amount within 1-sigma of the centre line. Samples are randomly distributed either side of the centre line. The SD chart starts trending more upwards after 1000 samples, with more instances outside the control limits.

## Household

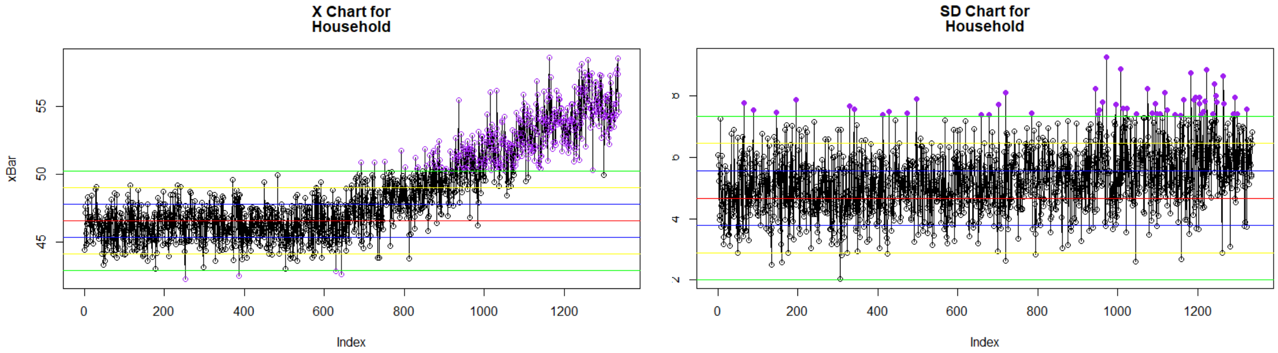


*Figure 19: Household extended X and SD charts*

Delivery time for Household products displays an upward trend after the 800th sample, where the sample mean moves out of the upper control limit. Up until this point the process seemed to be in control. The distribution for household delivery time has remained constant at the start but exhibits an increasing mean delivery time.

## Food

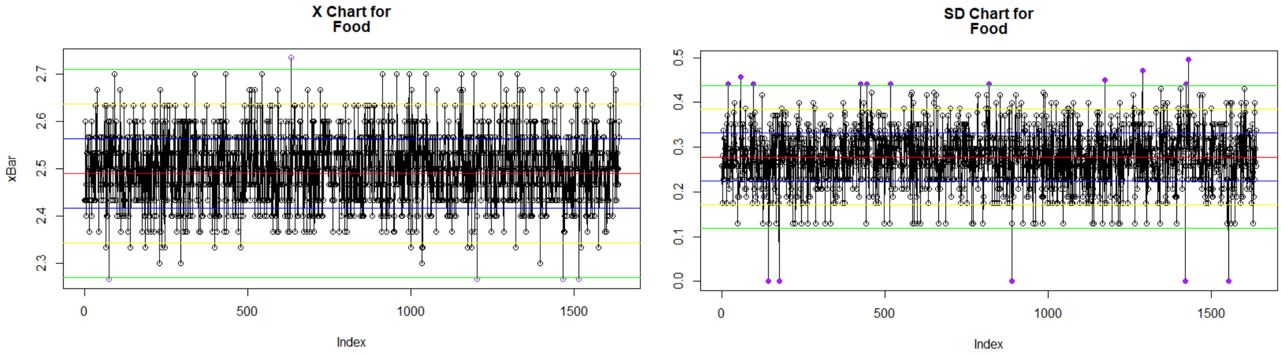


*Figure 20: Food extended X and SD charts*

Food has stayed in control. There only seems to be a very few select instances where the sample means fall outside the control limits. It can also be seen that a vast majority of samples fall within 2-sigma and a fair amount within 1-sigma of the centre line. Samples are randomly distributed either side of the centre line.

## Technology

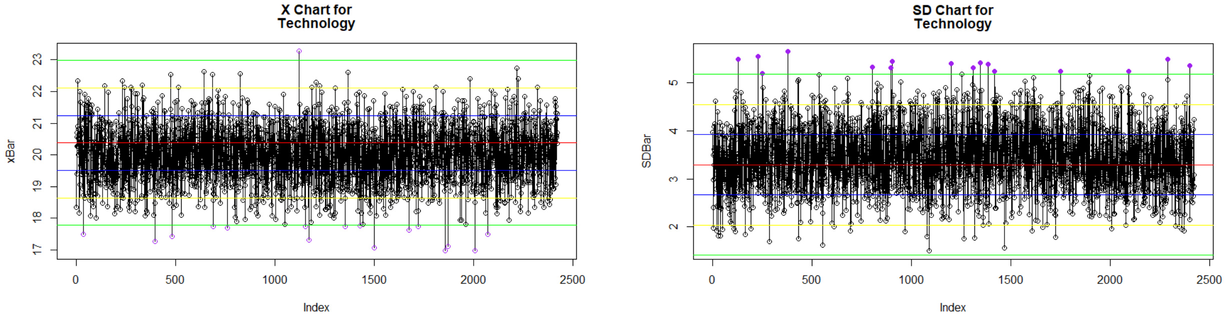


*Figure 21: Technology extended X and SD charts*

Technology has stayed in control. There only seems to be a very few select instances where the sample means fall outside the control limits. It can also be seen that a vast majority of samples fall within 2-sigma and a fair amount within 1-sigma of the centre line. Samples are randomly distributed either side of the centre line

## Sweets

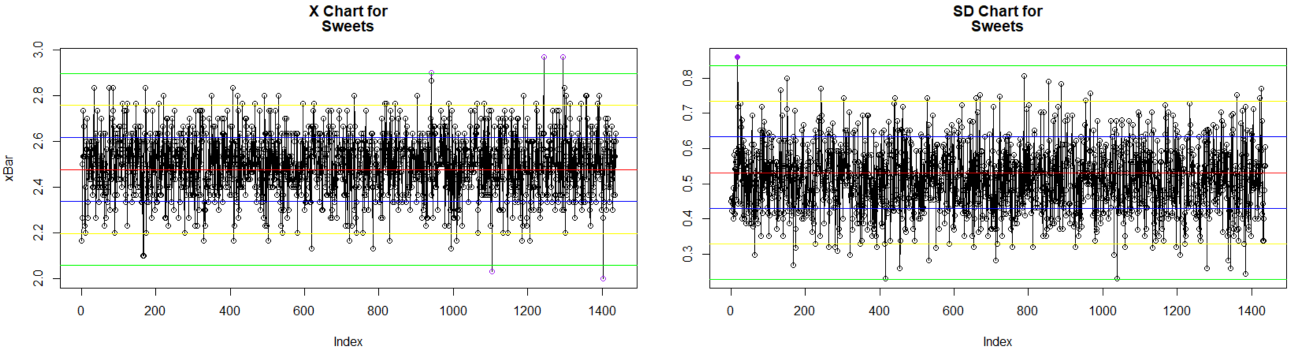


*Figure 22: Sweets extended X and SD charts*

Sweets has stayed in control. There only seems to be a very few select instances where the sample means fall outside the control limits. It can also be seen that a vast majority of samples fall within 2-sigma and a fair amount within 1-sigma of the centre line. Samples are randomly distributed either side of the centre line

## Gifts

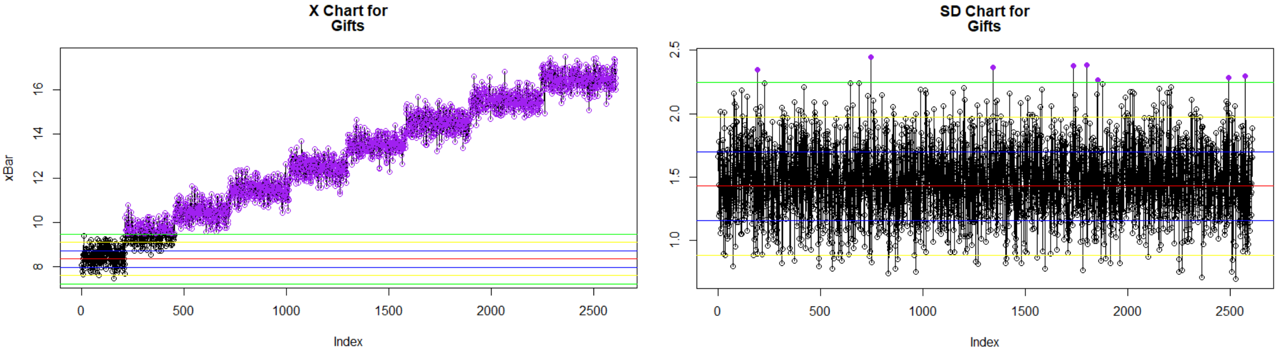


*Figure 23: Gifts extended X and SD charts*

Delivery time for Gifts has seen an upward trend from the start as the sample mean has increased above the centre lines. The distribution has remained fairly constant throughout.

### Luxury

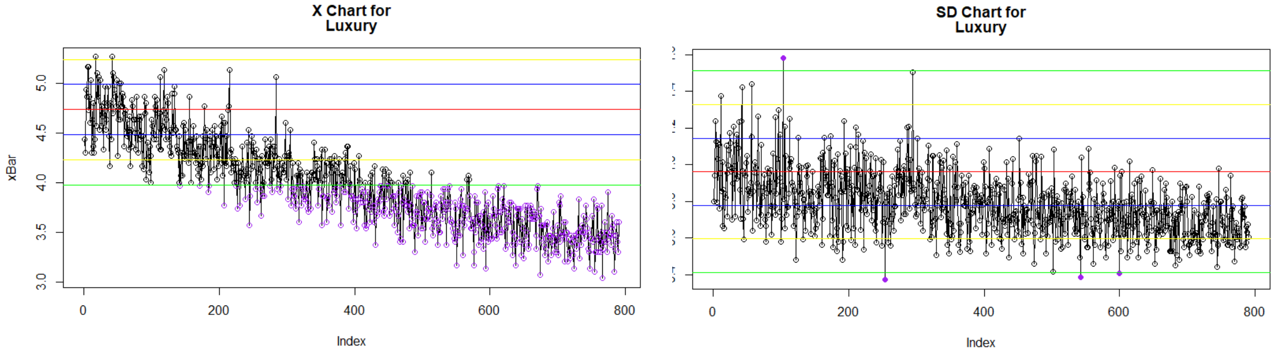


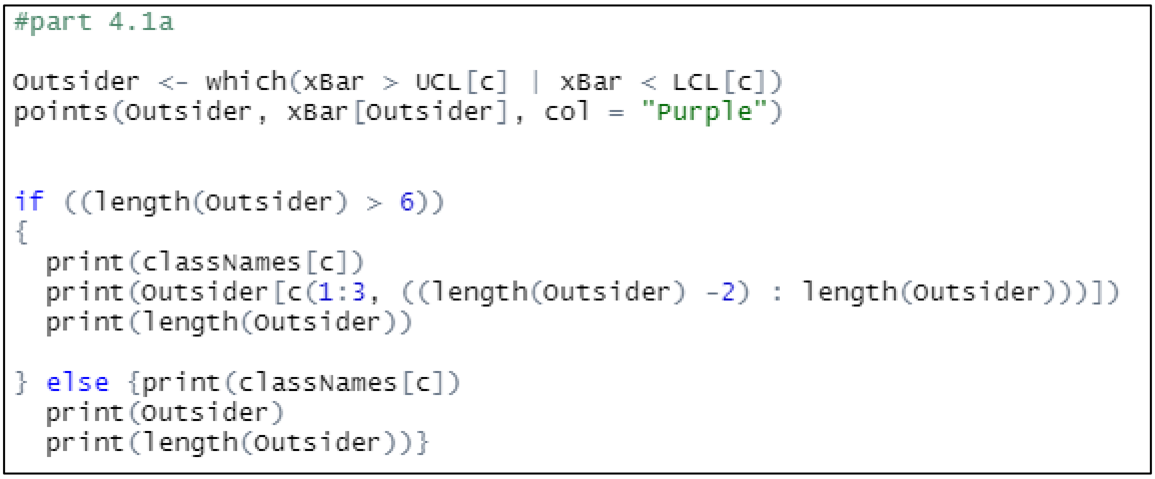
*Figure 24: Luxury extended X and SD charts*

Delivery time for Luxury products has seen a downward trend from the start as the sample mean has decreased below the centre lines. The distribution has remained constant throughout.

## Part 4: Optimising delivery process

### 4.1. A)

The purple points in the extended X and SD charts above are outside the control limits. The results can be summarised as follows:



```
#part 4.1a

Outsider <- which(xBar > UCL[c] | xBar < LCL[c])
points(Outsider, xBar[Outsider], col = "Purple")


if ((length(Outsider) > 6))
{
  print(classNames[c])
  print(Outsider[c(1:3, ((length(Outsider) -2) : length(Outsider)))])
  print(length(Outsider))

} else {print(classNames[c])
  print(Outsider)
  print(length(Outsider))}
```

*Table 7: Rule A results*

| | How many outsiders | Outsiders |
|---|---|---|
| Clothing | 17 | 455 702 1152 1677 1723 1724 |
| Household | 400 | 252 387 629 1335 1336 1337 |
| Food | 5 | 75 633 1203 1467 1515 |
| Technology | 17 | 37 398 483 1872 2009 2071 |
| Sweets | 5 | 942 1104 1243 1294 1403 |
| Gifts | 2290 | 213 216 218 2607 2608 2609 |
| Luxury | 434 | 142 171 184 789 790 791 |

Looking at the table above and the X bar plots household and luxury has many subgroups that are “out-of-control”, and gifts has the most. A possible reason for this could be a change in the process that wasn’t present during the first 30 samples, or a problem in the process exists. Food and sweets show very few “out-of-control” subgroups and thus these processes seem to be within the standard expectations. The most important class to further investigate will be gifts, since the trend of outsides is continuously upward. These investigations can include recording data about gift deliveries and an in-control class, like food, and comparing the results.

### 4.1.B)

For rule B, the upper and lower limits were calculated as follows:

sigmaPlus4 = CL + 0.4*((UCL- CL)/3)

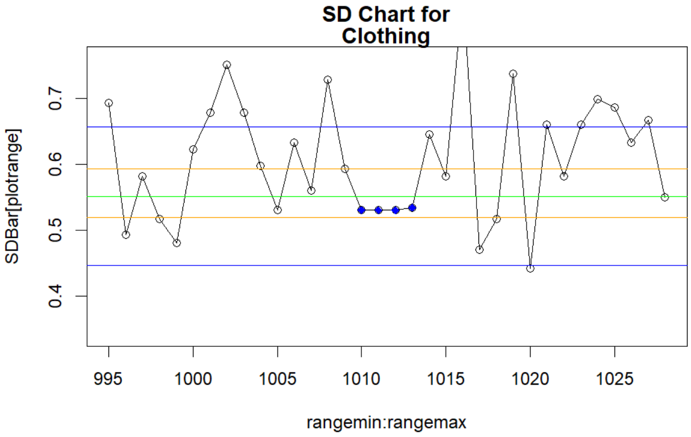
sigmaminus3 = CL - 0.3*((UCL- CL)/3)

The results can be summarised in the following table:

*Table 8: Rule B results*

| | Longest consecutive sample | Samples |
|---|---|---|
| Clothing | 4 | 1010 1011 1012 1013 |
| Household | 3 | 43 44 45 |
| Food | 7 | 946 947 948 949 950 951 952 |
| Technology | 6 | 367 368 369 370 371 372 |
| Sweets | 4 | 91 92 93 94 |
| Gifts | 5 | 250 251 252 253 254 |
| Luxury | 4 | 60 61 62 63 |

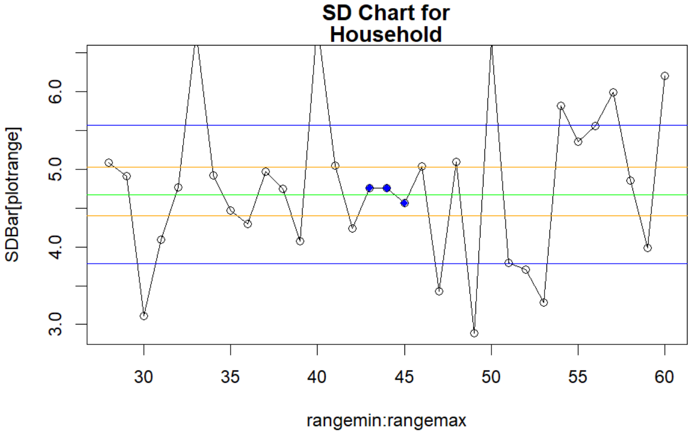
When a control chart has a large number of points between +0.4 sigma and -0.3 sigma it indicates that the process is tightly controlled. A tightly controlled process shows there is minimal variation in the delivery times. In the analyses of the X-bar charts it was mentioned that the sweets and food and processes were tightly controlled, this is verified by foods having the largest consecutive number of 7.



Limits:

0.5197152

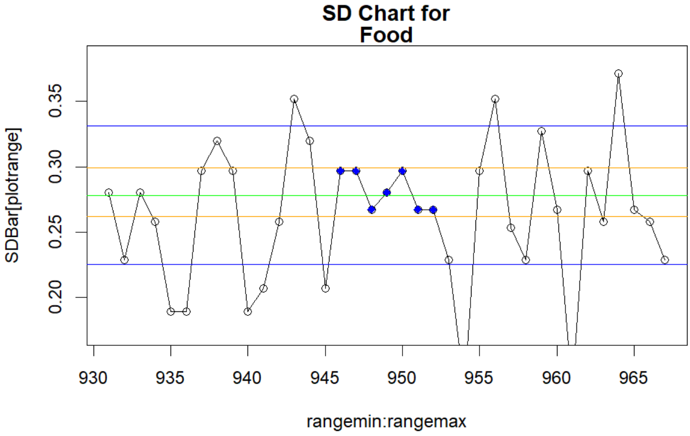
0.5932883



Limits:

4.404639

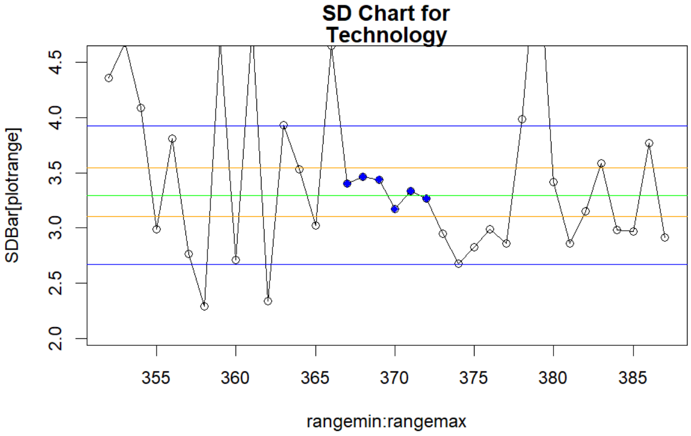
5.028178



Limits:

0.2622367

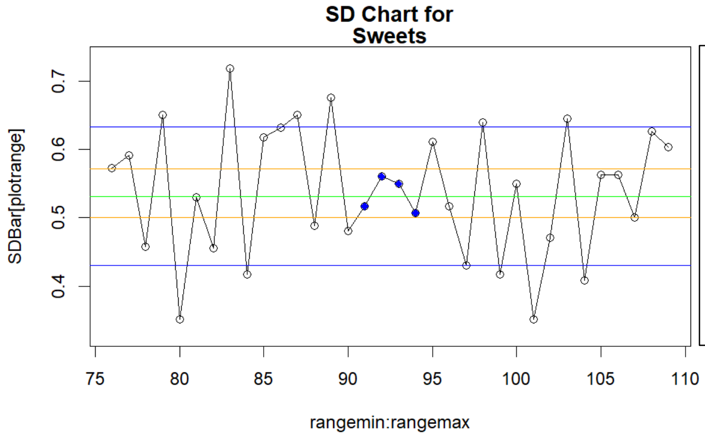
0.29936



Limits:

3.107024

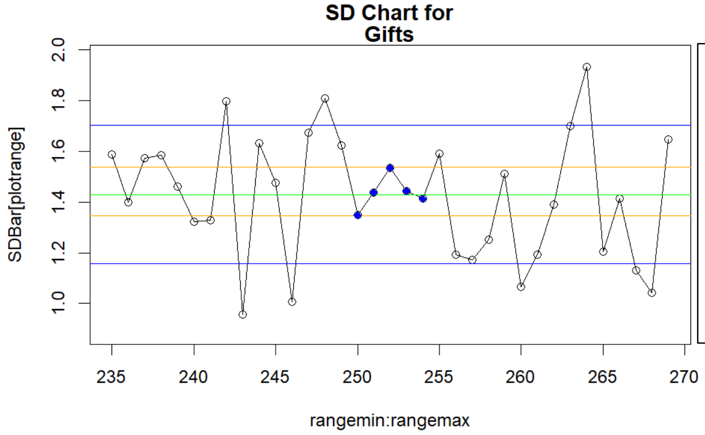
3.546867



Limits:

0.5009909

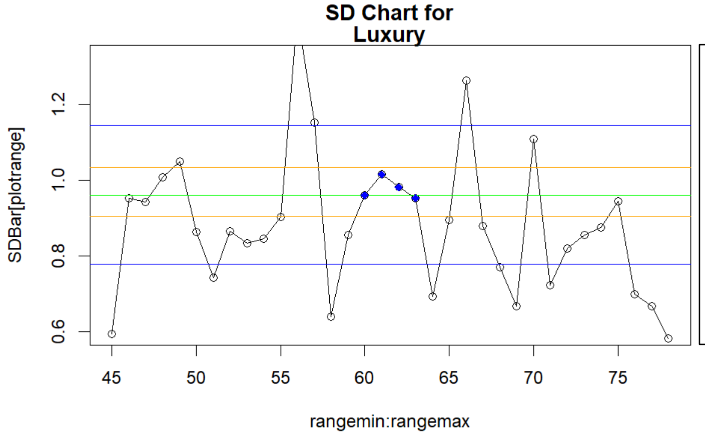
0.5719133



Limits:

1.347228

1.537948



Limits:

0.9062466

1.034539

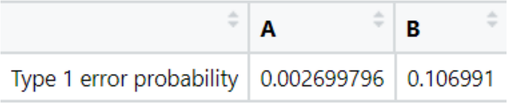
## 4.2 Type I error

The null hypothesis (Ho) assumes that the process is in control. Samples will be normally distributed with the natural limits at LCL and UCL. If samples were to be converted to the standard normal distribution with a mean = 0 and standard deviation =1, then the LCL = -3 and UCL = 3. A type I error, also known as a manufacturer's error, occurs when a process is deemed out of control however it is in fact in control. This error incurs costs to the manufacturer thus the name manufacturers error. Control charts are assumed to follow a normal distribution, therefore for a process to be deemed out of control it means there is a point beyond the +- 3 sigma limit.

For rule A: Probability of Type I Error for an X bar sample outside of the outer control limits: $P\ (z < -3) + P\ (z > 3)$ = 0.0013 + 0.0013 = 0.002699796

For rule B: The pattern length 7 is used because its' the most consecutive samples outside the limits of rule B. Here, the probability of a type I error relates to pattern lengths of 7 consecutive means between the +0.4 sigma and -0.3 sigma line.

*Table 9: type I errors*



| | A | B |
|---|---|---|
| Type 1 error probability | 0.002699796 | 0.106991 |

## 4.3 Centre delivery process

It is important for a business to deliver products fast enough to not pay penalty fees, but not so fast that optimising the process costs more than what the reward would be. Currently, an item is considered late if it takes longer than 26 hours to be delivered, and it costs R329/item-late-hour in lost sales. As can be seen on figure X, where the red line represents a delivery time of 26 hours, there are some deliveries above this point and would be considered late. Figure XX is the histogram of current delivery times, and here it's seen that the mean delivery time is currently 20.01 hours.

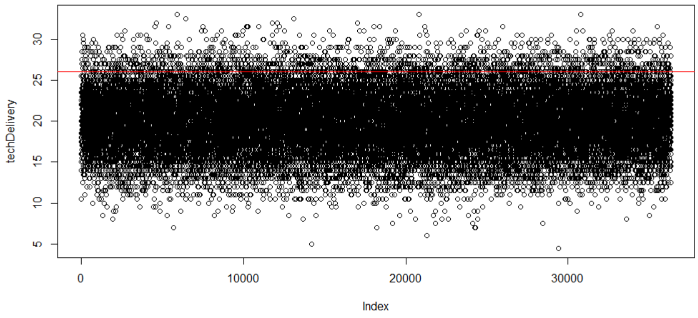


*Figure 25: Delivery times for technology*

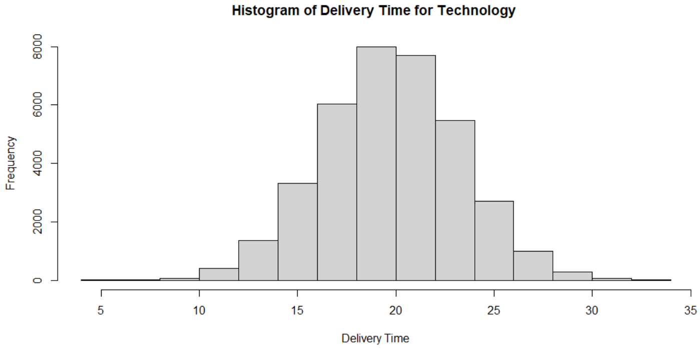


*Figure 26: Histogram of technology delivery times*

The business can choose to speed up deliveries at a cost of R2.5/item/hour, and similarly slow down the process at a cost of -R-2.5/item/hour. By using a brute-force method of calculating how much each hour would cost in penalties and in speeding up deliveries, the number hours which need to be shifted can be calculated.

The business currently pays R758674 in penalties, and this will be a starting point to compare costs with. Shifting the process to be 1-hour faster costs R 460992.5, 2 hours faster costs R 349689.5, 3 hours faster costs R 340870 and 4 hours faster costs R387487. As can be seen on the figure below, the optimal shift is 3 hours faster.

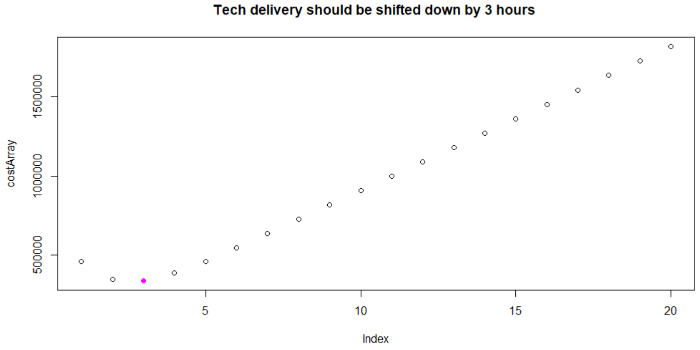


*Figure 27: Optimal delivery time*

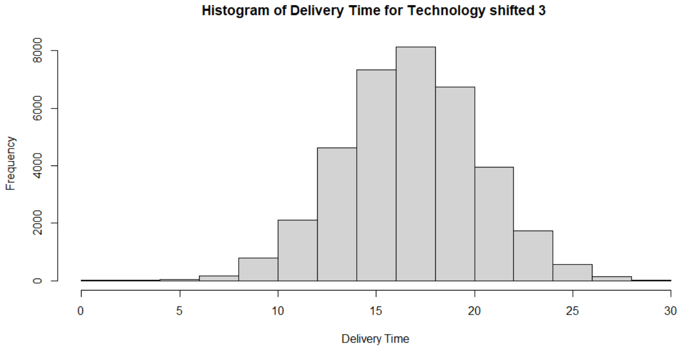


*Figure 28: Shifted histogram of technology delivery times*

## 4.4

A type II error, also known as consumer's error, happens when a process is incorrectly deemed in control, when it's in fact no longer in control. In this application, the mean delivery time for the technology class moved to 23 hours, so the process is no longer in control, but this fact is unknown.

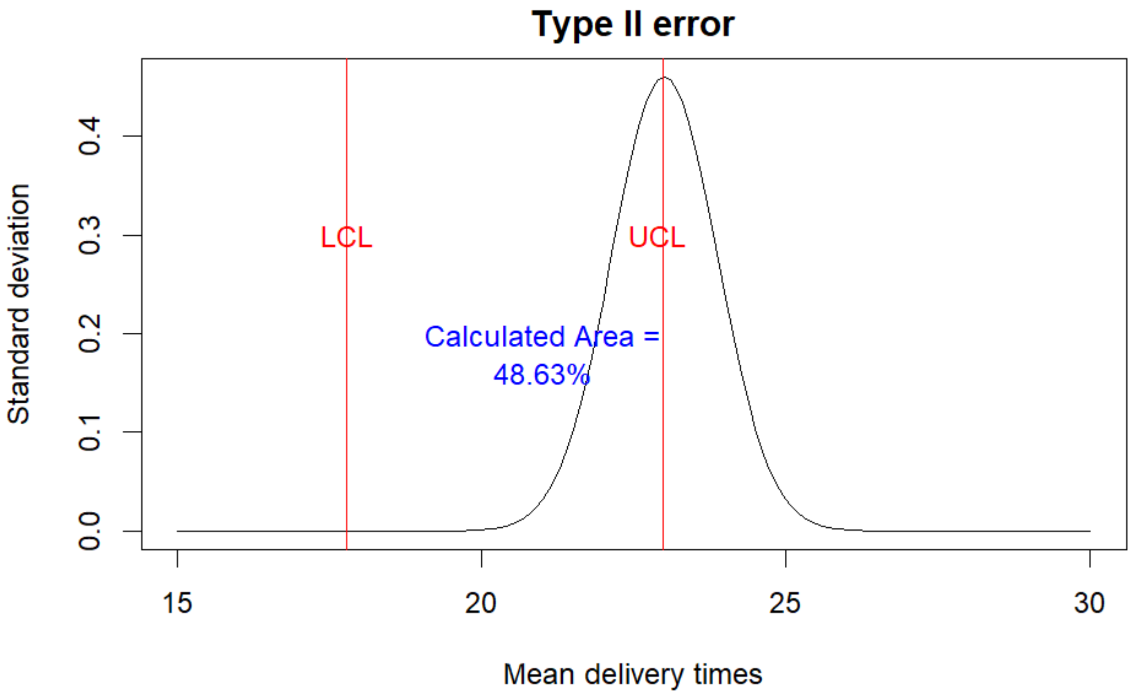
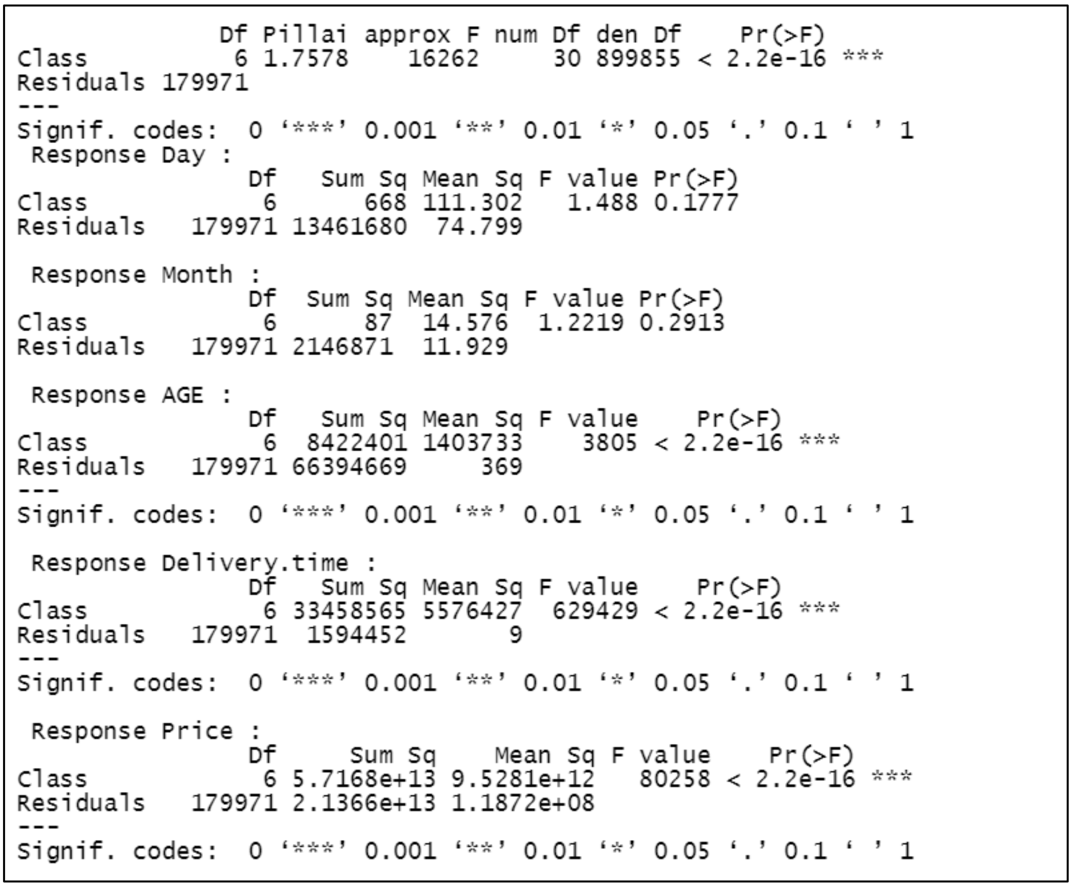


*Figure 29: Type II error*

# Part 5

A MANOVA test can be used to determine whether various levels of independent variables influence dependent variables. The independent variables can be separate or in combination with one another. Taking the results of part 2,3 and 4 into consideration, it becomes apparent that class has a strong relationship with some variables. The relationships to be investigated will be with Day, Month, Age, Delivery time and Price. The following results can be obtained from R code:

*Table 10: MANOVA*



```
               Df Pillai approx F num Df den Df    Pr(>F)    
Class           6 1.7578    16262     30 899855 < 2.2e-16 ***
Residuals 179971                                             
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
 Response Day :
                Df   Sum Sq Mean Sq F value Pr(>F)
Class            6      668 111.302   1.488 0.1777
Residuals   179971 13461680  74.799               

 Response Month :
                Df  Sum Sq Mean Sq F value Pr(>F)
Class            6      87  14.576  1.2219 0.2913
Residuals   179971 2146871  11.929               

 Response AGE :
                Df   Sum Sq Mean Sq F value    Pr(>F)    
Class            6  8422401 1403733    3805 < 2.2e-16 ***
Residuals   179971 66394669     369                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

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

Using an alpha value of 0.05, because it is a common alpha value to be used and makes the critical region 5%, the results can be summarised as follows:

| $H_o$: | There is no relationship between Class and average Age | | | |
|---|---|---|---|---|
| | Alpha | Mean sq | F-value | Pr(>F) |
| | 0.05 | 1403733 | 3805 | <2.2e-16 |
| Comment: | The P value is smaller than alpha so $H_o$ can be rejected | | | |

| $H_o$: | There is no relationship between Class and average Price | | | |
|---|---|---|---|---|
| | Alpha | Mean sq | F-value | Pr(>F) |
| | 0.05 | 9.5281e+12 | 80258 | <2.2e-16 |
| Comment: | The P value is smaller than alpha so $H_o$ can be rejected | | | |

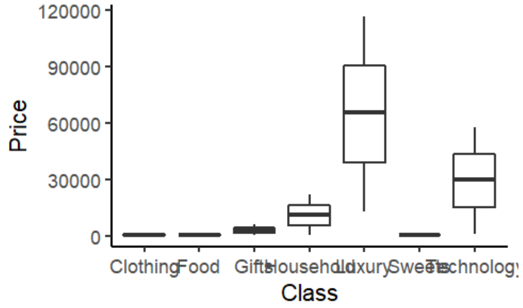
| $H_o$: | There is no relationship between Class and average Delivery time | | | |
|---|---|---|---|---|
| | Alpha | Mean sq | F-value | Pr(>F) |
| | 0.05 | 5576427 | 629429 | <2.2e-16 |
| Comment: | The P value is smaller than alpha so $H_o$ can be rejected | | | |

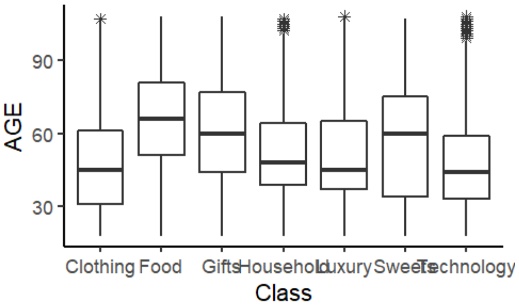
| $H_o$: | There is no relationship between Class and average Day | | | |
|---|---|---|---|---|
| | Alpha | Mean sq | F-value | Pr(>F) |

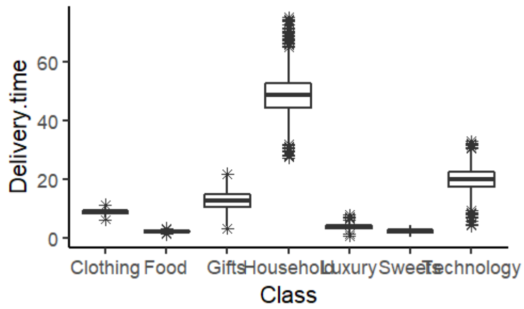
| | | | | |
|---|---|---|---|---|
| | 0.05 | 111.302 | 1.488 | 0.1777 |
| Comment: | The P value is not smaller than alpha so $H_o$ can't be rejected | | | |

| | | | | |
|---|---|---|---|---|
| $H_o$: | There is no relationship between Class and average Month | | | |
| | Alpha | Mean sq | F-value | Pr(>F) |
| | 0.05 | 14.576 | 1.2219 | 0.2913 |
| Comment: | The P value is not smaller than alpha so $H_o$ can't be rejected | | | |

From these tests it can be concluded that the day and month a product was bought has no significance on the class of the product that was bought, because of the low F values, and those $H_o$'s being rejected. On the other hand, age, delivery time and price have a strong relationship between the class of product bought, because of the high F values. Delivery time has the largest F, suggesting that class and delivery time has the strongest relationship.







The box plots above verify the conclusion that the average price, delivery time and age have a strong relationship with Class. The average price is highest with luxury and technology products, and differ with other classes, which is evidence that class has a relationship with price. Similarly, the average age differs per class, so the conclusion that there's a relationship between age and class makes sense. Lastly, average delivery times are different for every class, with household goods having the longest mean delivery times. This boxplot once again shows that the conclusion that there's a relationship between delivery time and class makes sense.

# Part 6: Reliability of the service and products

## 6.1 Taguchi loss function

“Taguchi measured quality as the variation from the target value of a design specification, and then translated that variation into an economic “loss function” that expresses the cost of variation in monetary terms. In mathematical terms, Taguchi assumes that losses can be approximated by a quadratic function so that larger deviations from target correspond to increasingly larger losses. For the case in which a specific target value, T, is determined to produce the optimum performance, and in which quality deteriorates as the actual value moves away from the target on either side (called “nominal is best”), the loss function is represented by: **$L(x) = k\ (x\text{-}T)^2$**” (Evans, 2010)

**FIGURE 7.12**
Nominal-Is-Best
Loss Function

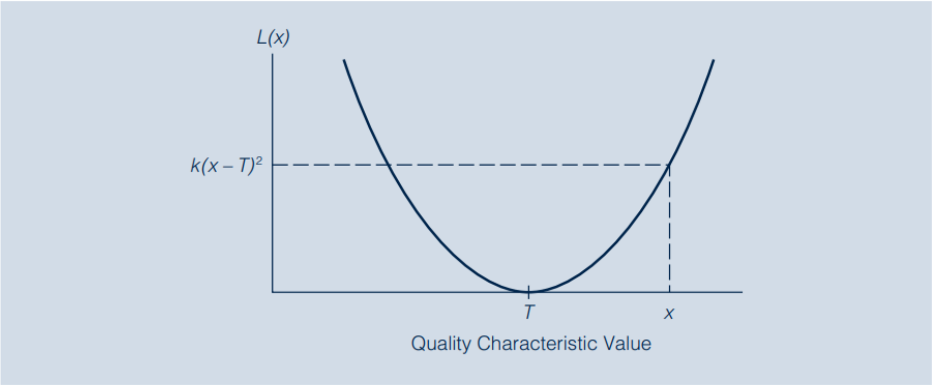


*Figure 30: Figure from textbook*

## Problem 6

$L= k(y\text{-}m)^2 = 45$ when $(y\text{-}m) = 0.04$

$k = 45/ (0.04)^2 = 28125$

$L = 28125\ (y\text{-}0.06)^2$

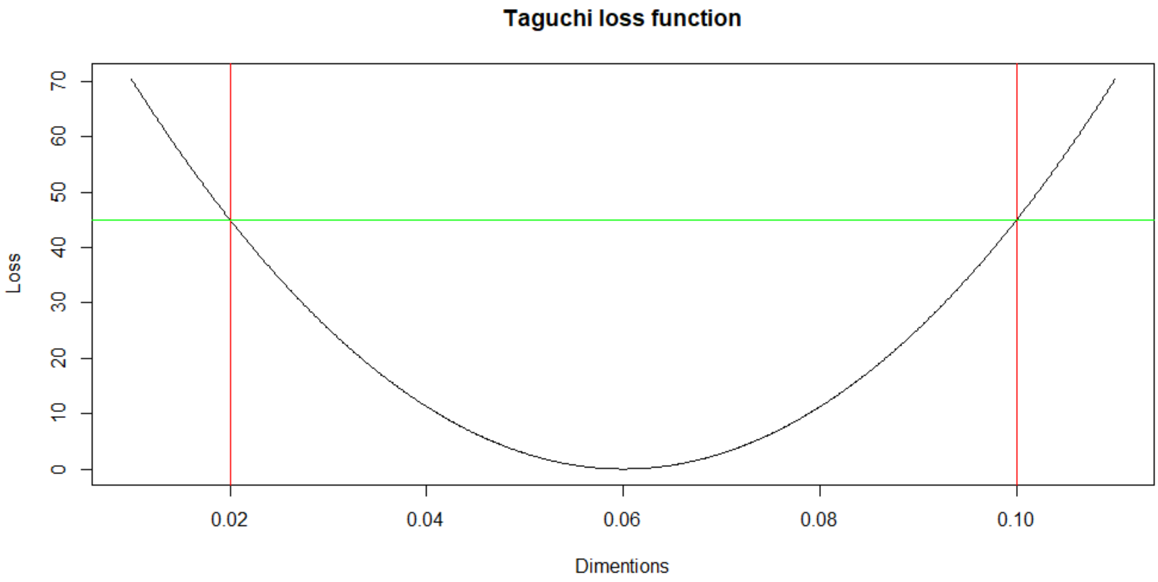


*Figure 31: Taguchi loss function*

From the figure above it's observed that the costs follow a quadratic function. The red lines are at 0.02 cm and 0.1 cm and the green line represents the cost of scrap. At 0.06 cm there is no cost of scrap but moving further away from this point in either direction makes the cost of scrap more expensive. Therefore, it's in the businesses best interest to keep within the specifications, since moving outside the boundaries will be very expensive to the business and ultimately lower profit.

If scarp is reduced to $35 per part, the loss function is as follows

k = 35/ ($0.04^2$) = 21875

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

L = 21875*$0.027^2$ = 15.95

In the figure below it's once again observed that at 0.06 cm there are no cost of scrap but moving further away from this point incurs more costs. Now, at +- 0.04 cm the cost is $35, as can be seen by the red and green lines. If the deviation is reduced to 0.027 cm, the loss will cost $15.95, which is represented by the red crosses.

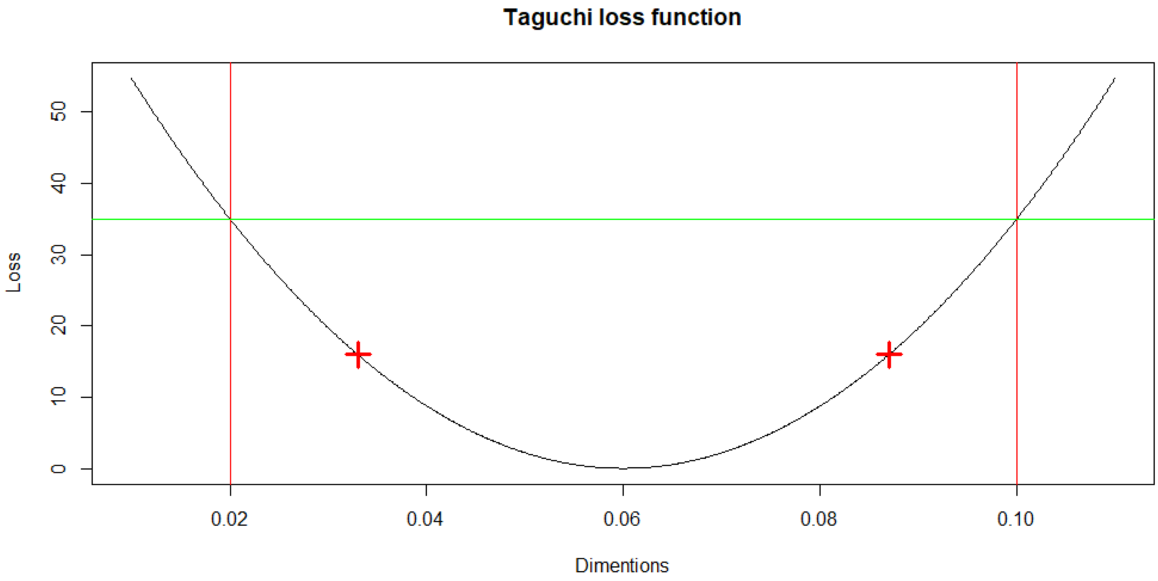


*Figure 32: Taguchi loss function II*

## 6.2 problem 27

If only one machine in each A, B and C is working, the reliability is:

$$R_A*R_B*R_C = 0.85*0.92*0.9 = 0.7038$$

For parallel connected components the combined reliability is 1- (both fail). The parrel components will work when both are working and when only one is working, so its only out of order when both are failing.

Both A's fail with a probability $(1-0.85)^2$ = 0.0225 and their combined reliability is 1-0.0225 = 0.9775

Both B's fail with a probability $(1-0.92)^2$ = 0.0064 and their combined reliability is 1-0.0064 = 0.9936

Both C's fail with a probability $(1-0.90)^2$ = 0.01 and their combined reliability is 1-0.01 = 0.99

The combined reliability is thus:

**$R_{AA}$*$R_{BB}$*$R_{CC}$ = 0.9775*0.9936*0.99 = 0.9615**

When two components are connected in parallel then it's a 26% improvement in reliability.

## 6.3 Reliable delivery

The binomial distribution describes the probability of obtaining exactly x "successes" in a sequence of an identical experiments, called trials. A success can be any one of two possible outcomes of each experiment. In some situations, it might represent a defective item, in others, a good item. The probability of success in each trial is a constant value p. The binomial probability function is given by the following formula, where p is the probability of a success, n is the number of items in the sample, and x is the number of items for which the probability is desired:

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

In order to answer this question, the value of p needs to be calculated. P will be the vehicle's reliability probability and the driver reliability probability.

The vehicle's reliability probability is found using the following graph:

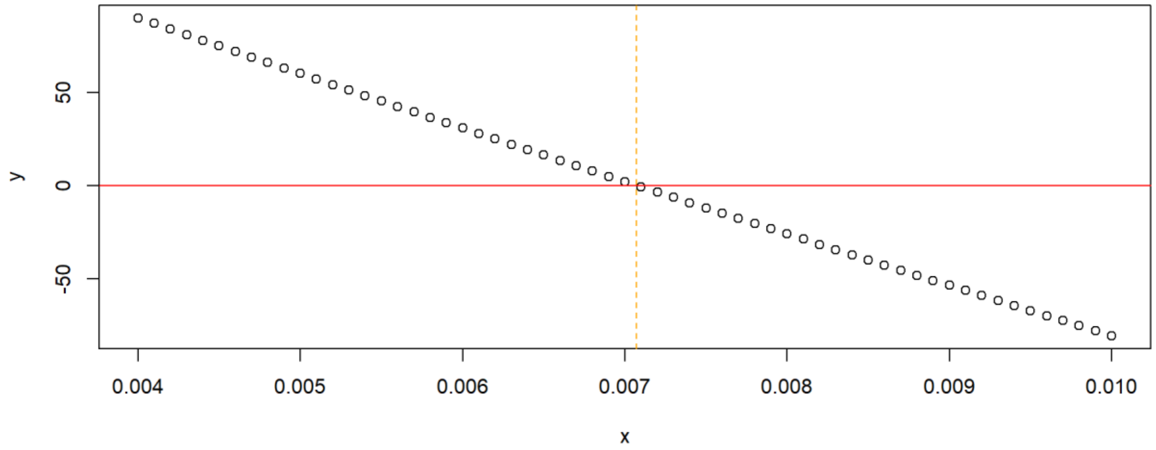


*Figure 33: Vehicle reliability probability*

Vehicle reliability probability = **0.007071661**

Probability of having zero break downs = **0.8615411**

Days with zero breakdowns = 0.8615411*365= **314.4625**

Probability of having only one break down = **0.1288543**

Days with only one break down = 0.1288543*365 **= 47.03181**

Total vehicle reliability = 0.8615411+0.1288543 = **0.9903954**

Days with reliable vehicle per year =0.9903954*365 = **361.4943**

This result means that there are between 361 and 362 days in the year where the business has reliable vehicle service to provide deliveries to customers.

The driver reliability probability is found using the following graph:

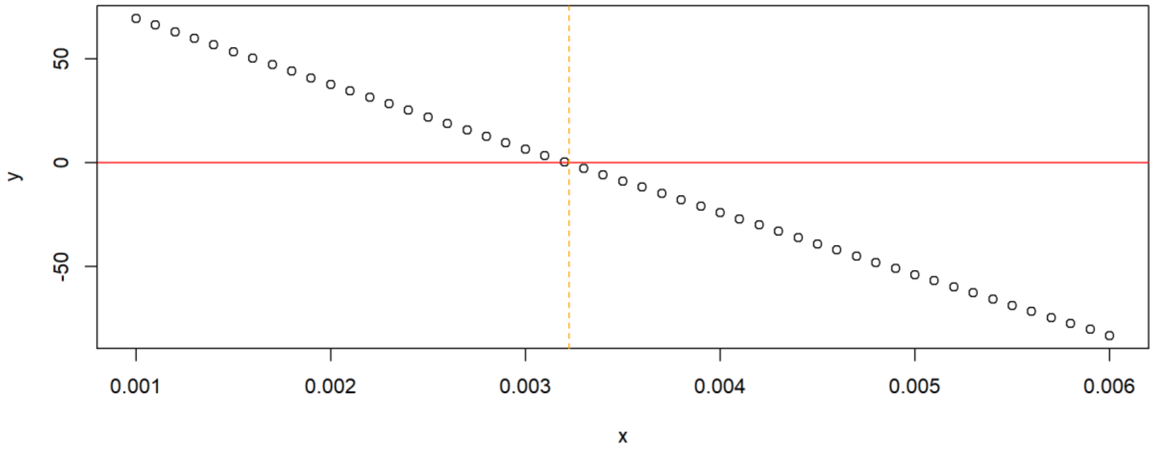


*Figure 34: Driver reliability probability*

Driver reliability probability = **0.003224402**

Probability of having zero drivers off = **0.9344269**

Days with zero drivers off = 0.9344269*365 = **341.0658**

Probability of having only one driver off = **0.06347701**

Days with only one driver off = 0.06347701*365 = **23.16911**

Total driver reliability = 0.9344269 + 0.06347701 = **0.9979039**

Total days with a reliable driver = 0.9979039*365 = **364.2349**

This result means that there are between 364 and 365 days in the year where the business has reliable drivers to provide deliveries to customers.

Total reliability = 0.9979039*0.9903954 = **0.9883195**

Total days with reliable delivery= 0.9883195* 365 = **360.7366**

This result means that there are between 360 and 361 days in the year where the business has reliable delivery services to provide deliveries to customers. This result is not too bad because it means there are at most 5 days with unreliable delivery services, but it's important to focus on continuously improving these services. As previously seen, majority of the business's customers come from recommendations, and when a customer has a bad experience with delivery, they might not recommend the business to future client anymore.

## Repeat with 22 vehicles

It must be assumed that the probability of failure per vehicle stays constant, and the new probability of having reliable vehicles can now include the probability of two breakdowns in reliability, because the business will still have reliable vehicles even if there are two breakdowns.

Probability of 2 vehicle breakdowns = **0.01002335**

Days with 2 vehicle breakdowns = 0.01002335* 365 = **3.658524**

Total reliability = 0.01002335+ 0.8615411 + 0.1288543 = **0.9995076**

Days reliable = 0.9995076 * 365 = **364.8203**

With the addition vehicle, the business now has reliable vehicle service 364-365 days of the year. This is an additional 3 days with reliable service.

Overall reliability = 0.9995076*0.9979039 = **0.9974125**

Days with reliable delivery = 0.9974125 * 365 = **364.0556**

With one additional car, the business can have reliable delivery services 364 – 365 days of the year.

# Conclusion

Analysing the valid data set made the relationship between different classes and other variables apparent. Class had a strong relationship between delivery time, age and price. The delivery times for household and gift items are no longer in control and is on an upward trend. Delivery times of luxury items are no longer within the control limits and is on a downwards trend. These instances should be investigated to gain an understanding of a potential error the business might be making.

The business can also consider investing money into either purchasing an additional vehicle or getting resources to speed up delivery by 3 hours. Purchasing another vehicle will result in having reliable delivery 364 days of the year, which will have a positive impact on the business since it will increase customer satisfaction. Speeding up delivery of technology items will cost the business R340870 and keeping the current delivery times will cost the business R758674.

# References

References

Evans, J., 2010. Managing for Quality and Performance Excellence. 11th ed. South Western Educational Publishing, p.363.

Hernandez, F., 2015. Data Analysis with R - Exercises. [online] Fch808.github.io. Available at: <http://fch808.github.io/Data-Analysis-with-R-Exercises.html> [Accessed 7 October 2022].

Hessing, T. and PV, R., 2019. Data Distributions. [online] Six Sigma Study Guide. Available at: <https://sixsigmastudyguide.com/data-distributions/#:~:text=Data%20distribution%20is%20a%20function,has%20a%20scattering%20of%20data.> [Accessed 1 October 2022].

Mcleod, S., 2019. Type I and Type II Errors - Simply Psychology. [online] Simplypsychology.org. Available at: <https://www.simplypsychology.org/type_I_and_type_II_errors.html> [Accessed 10 October 2022].

Steele, C., 2022. Process Capability Statistics: Cp and Cpk, Working Together. [online] Blog.minitab.com. Available at: <https://blog.minitab.com/en/statistics-and-quality-improvement/process-capability-statistics-cp-and-cpk-working-together#:~:text=Definition%20of%20Cpk&text=LSL%20stands%20for%20Lower%20Specification,centered%20between%20the%20specification%20limits.> [Accessed 8 October 2022].

STHDA, 2022. MANOVA Test in R: Multivariate Analysis of Variance - Easy Guides - Wiki - STHDA. [online] Sthda.com. Available at: <http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance> [Accessed 10 October 2022].