ECSA Statistical Analysis Report

Prepared for: Quality Assurance 344

Prepared by: E. Ferreira

(Student number: 23620048)

Date: October 2022

Abstract

Sales data for a company was collected over several years and analyzed to determine the reliability of their delivery times. The analyzed data (obtained using descriptive statistics) provides a general understanding of the data and processes in the business. Findings were used in correlation with statistical process control values and stability of the delivery process was determined.

Furthermore, options for optimization of the delivery process were discussed. MANOVA test and other quality assurance calculations helped determine the next steps towards improvement for the business.

Table of content

1. Introduction

This report will be discussing the statistical analysis of sales data for a company towards improving their processes for higher service quality.

Sales data for a company was collected over several years and analyzed to determine the reliability of their delivery times. The report starts by analyzing the data to get a general understanding and to identify which features influence each other. This is done by calculating some descriptive statistical values. These findings are used in correlation with the statistical process control values to determine the stability of the delivery process.

Furthermore, options for optimization of the delivery process is discussed alongside a MANOVA table and other quality assurance calculations to determine the next steps towards improvement for the business.

2. PART 1: Data wrangling

The statistical analysis in conducted using R-software. The relevant packages and libraries required to for this study are included in the software and need to be installed and loaded in R.Markdown as the first step – these will be used to perform analysis on the data and identify trends that can assist the business in their decisions going forward.

Before statistical analysis can be performed on the data, the data must be pre-processed to remove all instances that could result in errors in predictions or will interfere with the code later.

Data is imported using the *read.csv()* function and the imported data set is arranges in the following order: “ID”, “X”, “Year”, “Month”, “Day”.

3. PART 2: Descriptive statistics

To aid in the understanding of the data used for analysis, several graphs were created to visualize the data. The results are discussed below.

3.1. Process capability indices

Capability potential (CP) and Capability performance (CPK) are calculated to measure the implementation of the voice of the customer (VOC) and to illustrate how well the process meets its specifications.

CP ratios represents the measure of how the spread of the process (over six standard deviations) fits into the specification range (VOC). The CPK ratio is like the CP ratio, but additionally measures how close the mean of the process is to the target value. This is obtained by calculating two intermediate values, CPU and CPL. A CPU and CPL measures the VOC over three standard deviations. Generally, larger values are preferred for both the CP and CPK values. A value of 1 is hardly capable, a value >1.3 is acceptable and higher values are desired.

Class_names <chr>	sd <dbl>	Cp <dbl>	Cpu <dbl>	Cpl <dbl>	Cpk <dbl>
Clothing	0.6236642	6.4137077	8.0173877	4.810028	4.8100278
Food	0.2861025	13.9810049	25.0469544	2.915055	2.9150553
Gifts	2.9605905	1.3510818	1.2508151	1.451349	1.2508151
Household	6.2257998	0.6424877	-1.3234907	2.608466	-1.3234907
Luxury	0.9025717	4.4317809	7.3968342	1.466728	1.4667276
Sweets	0.5214166	7.6714096	13.7438233	1.598996	1.5989959
Technology	3.5019927	1.1422068	0.3796933	1.904720	0.3796933

Figure 1: Cp indices (e)

As seen in the table above, Food (CP=13.98), Sweets (7.67), Clothing (6.41) and Luxury (4.43) have the highest CP values, indicating that these items satisfy customer requirements. This could be due to shorter delivery time, lower prices or more advertisements. The reasons for differences in indices will be further explored below.

Gifts has a CP=1.35 which is low compared to the classes listed before, but the value is still acceptable, and the process capability is satisfactory. Technology-class has a CP=1.14 and Household has a CP=0.64. These values indicate that the process is not capable and the do not meet the requirements and VOC.

The CPK values confirm the observations made about the CP values. Clothing, food, and sweets have the highest CPK values, indicating that the average delivery times meet the target value (VOC specifications). Following the same trend as above, luxury- and gift- products have acceptable CPK values, but these items run the risk of not meeting delivery time targets on several deliveries. Again, the technology and household class underperform and does not meet customer requirements.

3.2. Price distribution

As seen in the graph below, the prices of items are spread out significantly, with the majority items having lower prices. This is potentially insightful, and the graph is worth exploring in further detail.

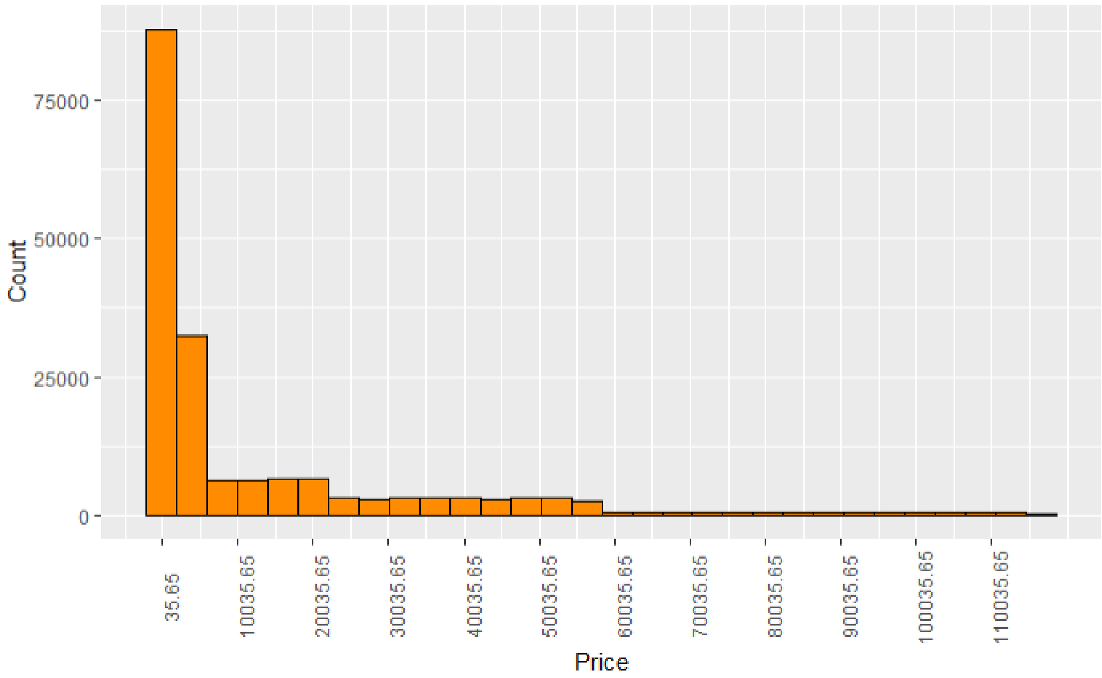


Figure 2 Price distribution graph

3.3. Price distribution described statistically

RStudio: Notebook Output

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
  35.65   482.31  2259.63 12294.10 15270.97 116618.97 
```

Figure 3 Summary of price data

The above image describes the statistical spread of the price-data. There is a major gap between the minimum and maximum cost of a product and most of the data is in a small range – the maximum price is an outlier.

RStudio: Notebook Output

Description: df [7 × 2]

Class <chr>	Value <dbl>
Sweets	6556385
Household	220899901
Gifts	115953130
Technology	1072529553
Luxury	769789207
Food	10024326
Clothing	16911791

7 rows

Figure 4 More Price data (b)

This image represents the total value of sales made in each class. Technology has the highest value, followed by Luxury, Household, Gifts, Clothing, Food and Sweets. This order is interesting when compared to the ranking of highest to lowest CP value. Food, sweets, and clothing had the highest

CP values but has the lowest total sales values. This could indicate that lower prices tend to satisfy the VOC more compared to higher prices.

3.4.Price distribution by class

The figure below highlights the point made above as it clearly shows the prices for food, sweets, and clothing are lower than prices for technology, luxury, and household items. These products also attract significantly more sales.

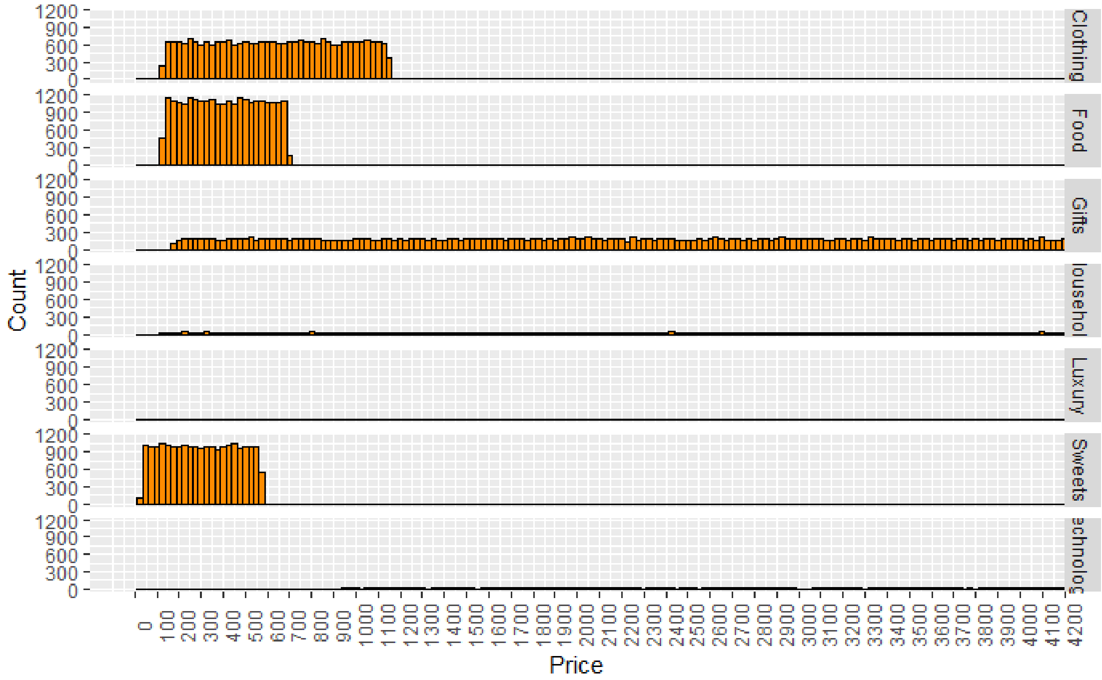


Figure 5 Price distribution by class

3.5.Delivery time distribution

Another potential reason for the difference in CP values is delivery time. The faster the delivery time, the higher the potential of satisfying VOC. The graph below shows that the delivery time is relatively spread out, with majority items having a fast delivery time.

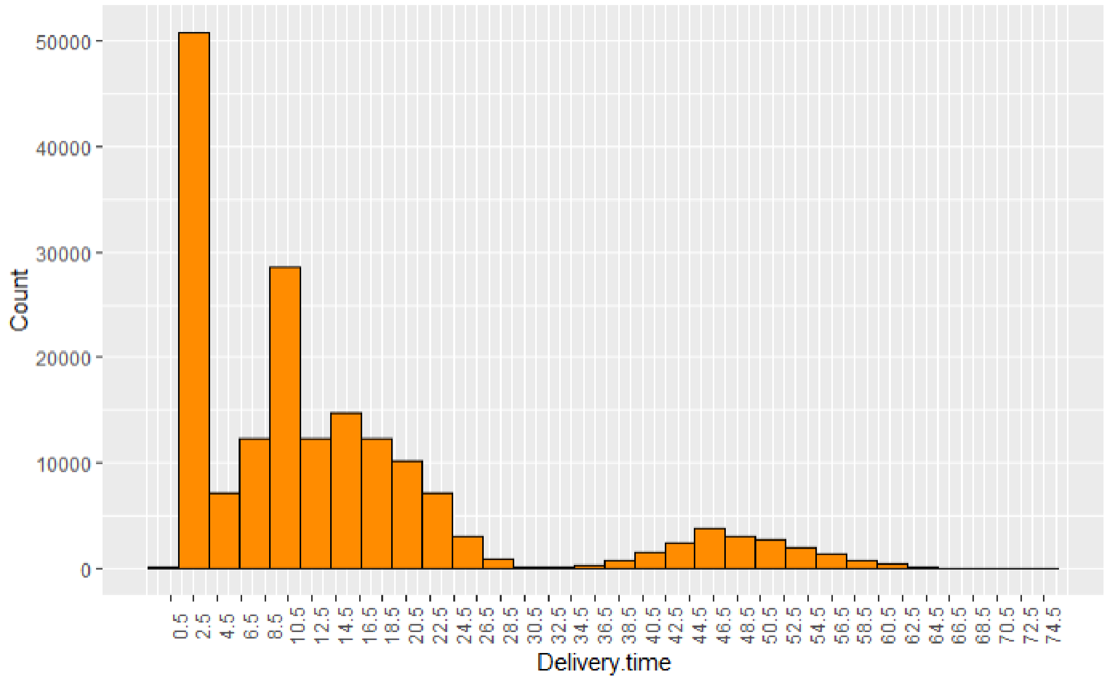


Figure 6 Delivery time distribution

3.6.Delivery time distribution by class

The delivery time distribution by class indicates that sweets, luxury, food, and clothing items generally have the fastest delivery times. Gifts and technology have longer delivery times, but they are still acceptable and household items have the longest delivery times. Comparing these results to the CP values, it is important to realize that food, sweets, and clothing feature amongst the best performing classes once again. This indicates that faster delivery times are important to satisfy VOC.

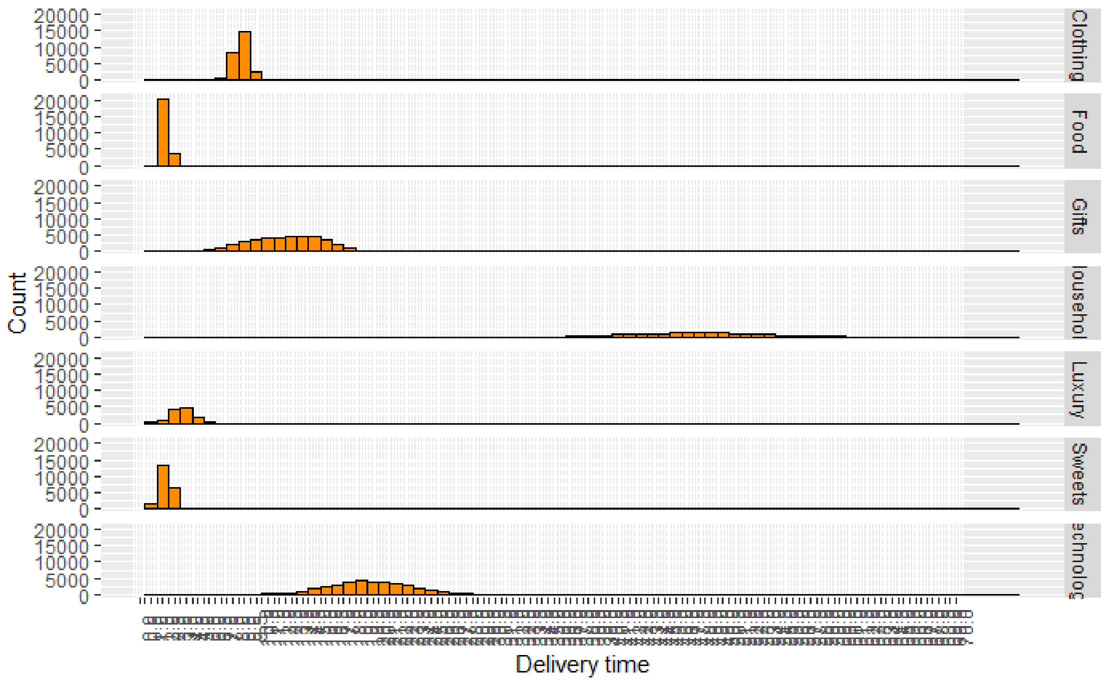


Figure 7 Delivery tie distribution by class

3.7. Distribution of types of products sold

The following figures are visual representations of the distribution of items sold per class and reasons for purchase. These graphs help to understand the data and to indicate which classes are responsible for majority sales and why.

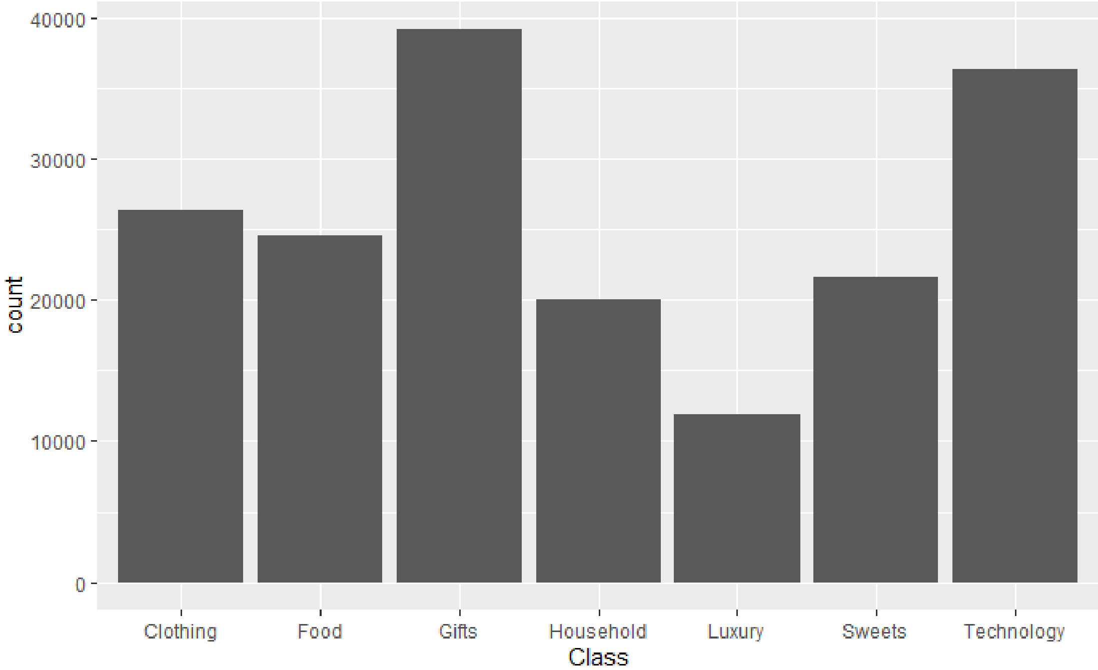


Figure 8 Distribution of types of products sold

As indicated above, household and luxury items make up majority of sales. This is alarming considering that these two items have the lowest CP values. As the price of a product is relatively fixed, the company should aim to shorten delivery times to improve the satisfaction of VOC (and CP) to acquire a better customer relationship and potentially increase revenue.

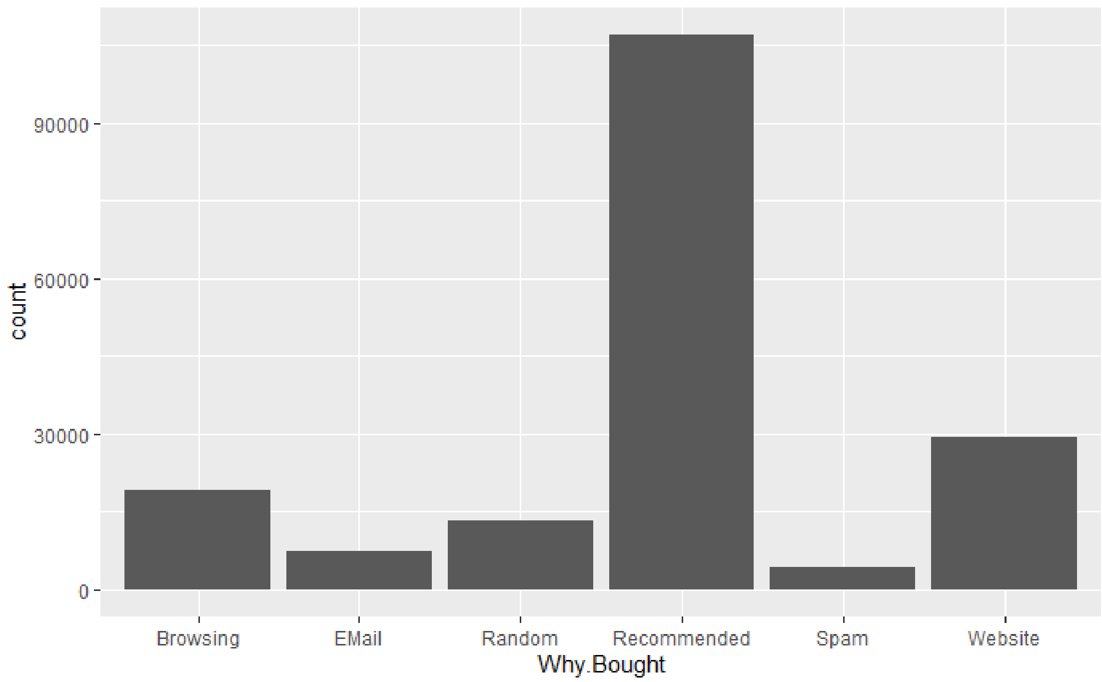


Figure 9 Reasons for purchase

4. PART 3: Statistical process control (SPC)

4.1. Summary of process control parameters

Thirty samples, consisting of 15 instances each, of each class was taken to calculate the statistical process control parameters. These parameters provide the upper and lower bounds where the process is allowed to operate in to be considered within the acceptable range. The complete set of x-bar and s charts can be found in Appendix A: Statistical Process Control Charts.

X-bar-S charts are used in combination to keep track of process variability. X-bar charts make use of the mean of each subgroup (sample) taken at regular intervals (constant number of instances per sample) – they monitor the consistency of the process mean. S-charts make use of subgroup standard deviations. The information from these charts is used to monitor the process variability.

The image below shows the formulas used and the steps followed to calculate the center lines, and upper- and lower control limits for all charts.

```
Creating data samples
```{r}
# order the data
newValid <- validData[order(validData$Year, validData$Month, validData$Day, validData$X),]

# constants used in x-bar and s- chart formulas
a3<- 0.789
b3<- 0.428
b4<- 1.572
```

For each class:
1. Filter by class
2. Create a sample of the first 450 values (per class)
3. Select only the delivery times
4. Create a matrix with 15 columns and 30 rows (15 instances per sample, 30 samples)
5. Create x-bar and s- charts
6. Calculate the values for
  X-BAR CHARTS
    CL (center line) = mean of all row means
    UCL (upper control limit) = CL + A3*(mean of all row standard deviations)
    LCL (lower control limit) = CL - A3*(mean of all row standard deviations)
  S CHARTS
    CL = mean of all row standard deviations
    UCL = CL*B4
    LCL = CL*B3
```

Figure 9: Calculations for control limits

The aim of creating the process control charts is to track the performance of the process and to make sure the process is running withing the desirable boundaries. The system should flag the process if it shows activity outside of the acceptable range for the process.

Concerning performance includes activity as indicated by the red dots in the image below.

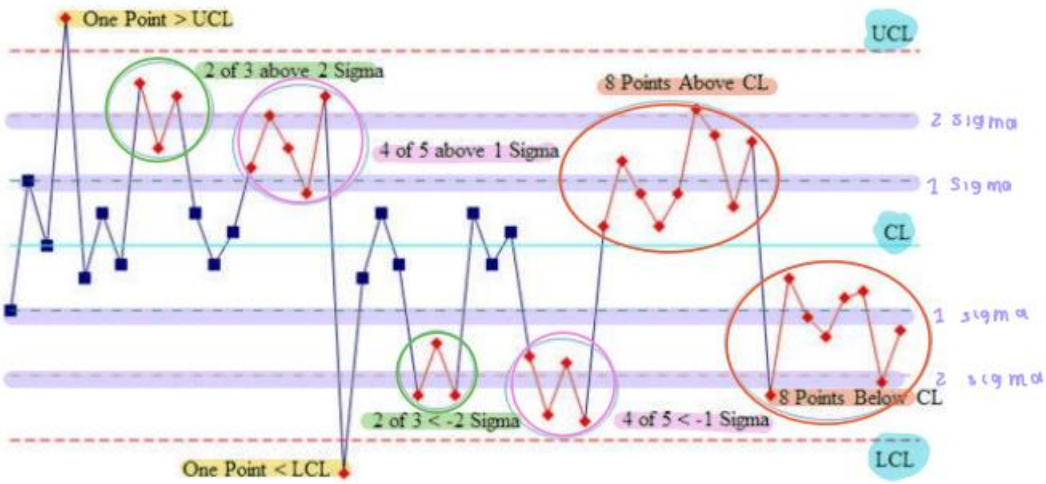


Figure10: Analyze control charts

4.2.X-bar charts

The x-bar chart for technology will be discussed as an example. The same logic can be applied to the charts for the other classes.

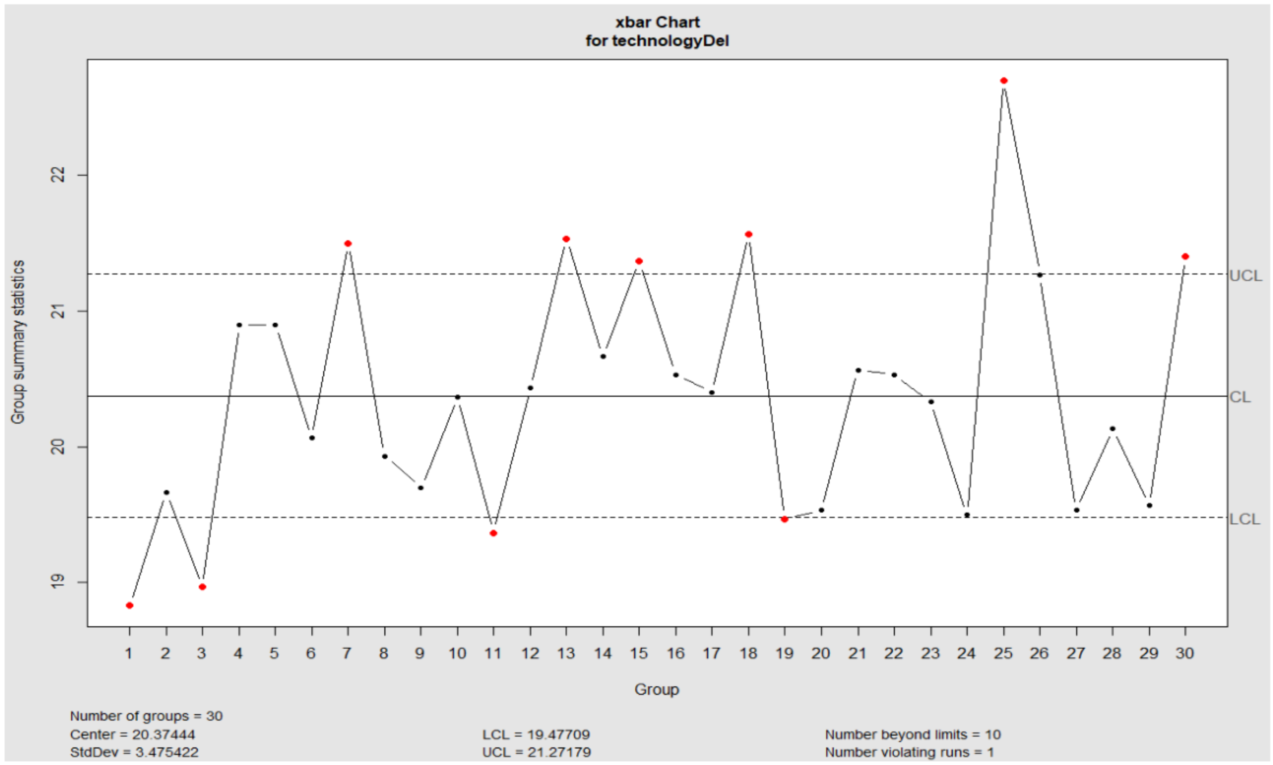


Figure 11: Technology Delivery times x-bar-chart

There are multiple points above (6) and below (4) the upper and lower control limits. These occurrences indicates that the process is out of control and should be investigated. Besides the instances where delivery times fall outside of the limits, the process does not seem to tend towards a faster (chart shifts downwards, moving the center line (mean) to a lower value) or slower (chart shifts upwards, moving the center line (mean) to a higher value) delivery time.

4.2.1. X-bar charts summary

RStudio: Notebook Output

Description: df [7 × 8]

Class <chr>	x.UCL <dbl>	x.2U <dbl>	x.1U <dbl>	x.CL <dbl>	x.1L <dbl>	x.2L <dbl>	x.LCL <dbl>
Technology	22.990345	22.118378	21.246411	20.374444	19.502477	18.630510	17.758543
Clothing	9.406600	9.261067	9.115533	8.970000	8.824467	8.678933	8.533400
Household	50.277816	49.039285	47.800753	46.562222	45.323691	44.085160	42.846629
Luxury	5.511177	5.252637	4.994096	4.735556	4.477015	4.218474	3.959934
Food	2.704311	2.632874	2.561437	2.490000	2.418563	2.347126	2.275689
Gifts	9.491443	9.114666	8.737888	8.361111	7.984334	7.607556	7.230779
Sweets	2.901900	2.760526	2.619152	2.477778	2.336404	2.195029	2.053655

7 rows

Figure 12: x-bar chart values summary

The above image provides a summary of the values obtained from the x-bar graphs for all classes (for the first 30 samples). The data shows that household items have the slowest delivery time, followed by technology, clothing and gifts.

RStudio: Notebook Output

Description: df [7 × 4]

Class <chr>	x.UCL <dbl>	x.LCL <dbl>	difference <dbl>
Technology	22.990345	17.758543	5.2318020
Clothing	9.406600	8.533400	0.8732007
Household	50.277816	42.846629	7.4311871
Luxury	5.511177	3.959934	1.5512435
Food	2.704311	2.275689	0.4286227
Gifts	9.491443	7.230779	2.2606642
Sweets	2.901900	2.053655	0.8482454

7 rows

Figure 13: x-bar Values analyzed

Considering the UCL and LCL and the difference between the bounds, food-items have the most constant delivery times. Sweets and clothing-items also show stable behavior with a small difference between "longest" and "shortest" delivery time for each process. This is due to a small deviation from the mean delivery time. These processes are likely to perform within process bounds without much supervision.

Customer satisfaction on these products is high as seen by the Cp and Cpk values (3.1 Process capability indices). Food, clothing, and sweets had the highest process capability indices and constant, dependable delivery times could be related to this.

4.3.S-charts

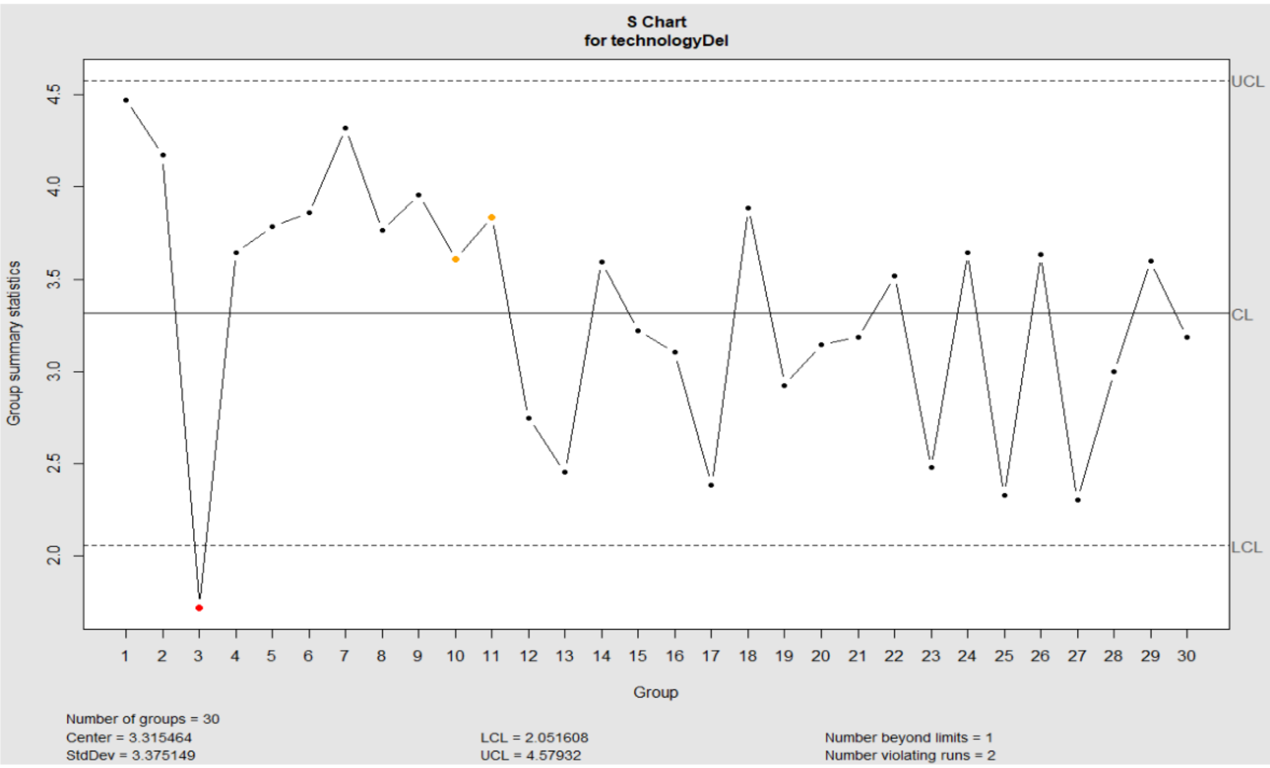


Figure 14: x-bar chart values summary

The s-chart for technology indicates that the process is unstable and requires investigation. Group 3 has an instance that falls below the LCL; this process behavior is concerning and should be resolved. Groups 10 and 11 will also cause the process to be flagged. Here, more than 8 consecutive points fall above the center line. This could indicate an upwards shift of the process mean (the process is moving towards a higher mean) – meaning the delivery times for technology could potentially be becoming slower. If this is the case, it will result in an increased number of late deliveries and a decrease in customer satisfaction. The result of this shift can be harmful towards the company and the process should therefore be investigated.

4.3.1. S-charts summary

RStudio: Notebook Output

Description: df [7 × 8]

Class <chr>	s.UCL <dbl>	s.2U <dbl>	s.1U <dbl>	s.CL <dbl>	s.1L <dbl>	s.2L <dbl>	s.LCL <dbl>
Technology	5.2119092	4.5797607	3.9476123	3.3154639	2.6833154	2.0511670	1.4190185
Clothing	0.8698805	0.7643734	0.6588662	0.5533591	0.4478520	0.3423448	0.2368377
Household	7.4029317	6.5050358	5.6071399	4.7092441	3.8113482	2.9134523	2.0155565
Luxury	1.5453452	1.3579115	1.1704778	0.9830440	0.7956103	0.6081766	0.4207428
Food	0.4269929	0.3752033	0.3234137	0.2716240	0.2198344	0.1680447	0.1162551
Gifts	2.2520685	1.9789169	1.7057652	1.4326136	1.1594619	0.8863103	0.6131586
Sweets	0.8450201	0.7425283	0.6400365	0.5375446	0.4350528	0.3325609	0.2300691

7 rows

Figure 15: S chart values summary

RStudio: Notebook Output

Description: df [7 × 4]

Class <chr>	s.UCL <dbl>	s.LCL <dbl>	s.difference <dbl>
Technology	5.2119092	1.4190185	3.7928907
Clothing	0.8698805	0.2368377	0.6330428
Household	7.4029317	2.0155565	5.3873752
Luxury	1.5453452	0.4207428	1.1246024
Food	0.4269929	0.1162551	0.3107379
Gifts	2.2520685	0.6131586	1.6389099
Sweets	0.8450201	0.2300691	0.6149510

7 rows

Figure 16: S chart values analyzed

S-charts represent the variability of the process. As seen in the data above, household delivery times are the most inconsistent, resulting in high variability. The process for technology delivery is also high in comparison to the rest. These two products had the lowest Cp values, translating to low VOC satisfaction.

These two products had the highest standard deviation but also the highest mean delivery time (delivery time is the longest). Items belonging to these classes are high-valued – the company should aim to improve the Cp values on these items for increased customer satisfaction leading to potential increase in sales and revenue. These processes are currently considered to be unstable and require investigation and monitoring for improvement.

5. PART 4: Optimization of the delivery process

5.1. A: Process performance relative to control limits

Considering the remaining delivery time data for each class, samples (of 15 instances) were compared to the original 30 (15x30) samples. The data below indicates whether the process remained stable or tended towards a faster/slower delivery time over time.

Class_names	Class.lengths	Over	Under	Over.percentage	Under.percentage
<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Clothing	1730	0	0	0.00000000	0.00000000
Food	1608	1	6	0.06218905	0.37313433
Gifts	2579	0	0	0.00000000	0.00000000
Household	1307	172	0	13.15990819	0.00000000
Luxury	761	0	389	0.00000000	51.11695138
Sweets	1407	3	1	0.21321962	0.07107321
Technology	2393	0	19	0.00000000	0.79398245

7 rows

Figure 17: Performance relative to control limits

The values were obtained by taking the mean of each sample and comparing it to the upper and lower bounds created by the first 30 instances. If the new sample mean falls outside of the acceptable range, it is counted as “Over” or “Under” respectively.

Clothing, food, gifts, sweets, and technology seems to have stable processes for delivery with little to no instances falling outside of the boundary area. Household and luxury items, however, show interesting results. The household process seems to have shifted upwards, tending towards a slower delivery time. This will be problematic if it is not monitored, because delayed deliveries come at a cost. On the contrary, luxury items seem to be delivered sooner than expected, having more than 50% items being delivered faster than the LCL. This could either mean that the process is running more efficiently, and the mean has shifted downwards, or the data is being collected incorrectly or inaccurately, falsely indicating faster delivery times. This process should also be examined and adjusted accordingly.

5.1.1. Household

Taking a closer look at household delivery times data, the first three and last three instances that exceeded the UCL are inspected.

Index (outside of bounds)	Index (luxury data)	Value	Distance from UCL	Distance as percentage of UCL
1	2	50.30000	0.025000	0.0497265
2	14	52.36667	2.091667	4.1604509
3	25	51.30000	1.025000	2.0387867
170	1304	50.56667	0.29166667	0.5801425
171	1305	50.33333	0.05833333	0.1160285
172	1307	50.46667	0.19166667	0.3812365

The values [2] and [3] seem to lie relatively far above the UCL. When looking at the distance as a percentage of the UCL, these values are significantly higher than the other instances examined. Other instances like this could easily be in the data and could result in unnecessary high delivery costs. If other instances like these exist, it could reflect badly on the company and customers can be left unsatisfied. These instances should be followed up and inspected. The relevant changes should be made to the system. Failure to do so could lead to increased costs or loss of customers – both being detrimental to the company.

5.1.2. Luxury

Luxury items falling outside of bounds totaled to "l.under" = 389. Extracting the first and last 3 values to fall under the LCL resulted in the following:

Index (outside of bounds)	Index (luxury data)	Value	Distance from LCL	Distance as percentage of LCL
1	6	3.866667	0.093233	2.3544281
2	7	3.600000	0.359900	9.0886133
3	8	3.800000	0.159900	4.0379808
387	751	3.833333	0.126567	3.1962171
388	753	3.933333	0.026567	0.6709008
389	759	3.933333	0.026567	0.6709008

The values under consideration do not seem to lie far from the LCL, but when taken as a percentage (distance/LCL) the numbers are a little more concerning. The instance with index 7 lies 9% from the boundary. The process should be inspected and monitored to determine the cause of variability.

Inconsistencies should be resolved so that the company does not overspend on delivery. If it is found that the process mean has shifted, the system should be updated accordingly. Faster delivery times can be used as advertisement or can lead to an increase is sales price (for better customer service and service quality). Both possibilities have the potential to positively influence revenue.

5.2. B: Process performance based on variability

Class.names <chr>	stable <dbl>	total.cases <int>	percentage <dbl>
Clothing	414	1730	23.930636
Household	120	1307	9.181331
Food	387	1608	24.067164
Technology	663	2393	27.705809
Sweets	392	1407	27.860697
Gifts	0	2579	0.000000
Luxury	173	761	22.733246

7 rows

Figure 18: Performance relative to variability

The data above shows the stability of each of the processes. The values in the column “stable” represents the number of values for that class, falling between 0.4 and -0.3 sigma values of the mean. The more values in the given range, the more reliable the delivery process for that class. Reliable delivery times results in higher customer satisfaction and attracts more customers. Currently, technology items have the most reliable delivery time based off the analysis above. Technology items also have a high mean delivery time value – this could potentially be the reason for less variability. Household items seem to have less constant delivery times (times deviate from the mean value). Inconsistency reflects badly on the company and generally don’t attract business.

5.3. Type I error probability

A type I error occurs when the test indicates that the process is unstable/out of control bounds, given that the system is stable. This is also referred to as a false positive.

To calculate the probability of making a type I error, the normal distribution is considered. The probability of a value falling three sigma values above or below the center line is equal to the probability of making a type I error. This percentage is the same for every process: 0.1349898%. The method is illustrated below.

```
4.2
# calculate the probability of making a type 1 error using the pnorm() function
```{r}
#----------------
# TYPE 1 ERROR
# ----------------

pnorm(17.758543, mean=20.374444, sd=(22.990345-17.758543)/6)
pnorm(-3)

```

 [1] 0.001349898
 [1] 0.001349898
```

Figure 19: Tyoe I error

These errors can cost the company money – spending resources on “fixing” a process that was not flawed is unnecessary. These avoidable costs decrease revenue.

5.4.Cost of reduced delivery time

Late delivery of items cannot be left undealt with as this can accumulate to a significant cost, reducing the revenue of the company remarkably.

As identified by the x-bar and s-chart analysis, Technology delivery times extremely variable. Delivery times for these items are also amongst the slowest of all data instances. This can result in unnecessary big expenses; thus, technology delivery times were considered for the following study.

Technology items delivered slower than 26hrs translates to R329/item-late-hour in lost sales. The total cost of lateness was calculated by first determining which deliveries took longer than 26hrs. The total “hours late” was determined by subtracting 26hrs from all these instances and taking the sum of the remainder. This value was multiplied by R329, calculating to a total cost of R758 674.00. By reducing the process variability and delivery time for technology items, a weighty amount of money can be saved.

The cost of reducing the delivery time is R2.5/item/hour. Currently, the average delivery time for technology items is 20.01095hrs. Reducing the delivery time for each instance by x-hours, will reduce the mean by the same value. This will result in fewer instances being delivered after 26hrs. The tradeoff between cost of reducing delivery time and cost of late delivery is shown in the table below.

RStudio: Notebook Output

Description: df [9 × 5]

reducedHrs <int>	out.of.bounds <dbl>	hrs.late <dbl>	cost <dbl>	save <dbl>
0	724	2306.0	758674.0	0.0
1	354	1125.0	460992.5	297681.5
2	159	510.5	349689.5	408984.5
3	66	207.5	340870.0	417804.0
4	22	73.0	387487.0	371187.0
5	5	19.5	460753.0	297921.0
6	0	4.0	546521.0	212153.0
7	0	0.0	636072.5	122601.5
8	0	0.0	726940.0	31734.0

9 rows

Figure 20: Cost of reduced delivery times

The data above represents the number of hours reduced per item (faster average delivery time), the total instances that exceed 26hrs, the total hours late, and the corresponding cost of lateness and the total value saved by reduced hours.

Reducing the delivery time by 3hrs results in the lowest cost (R340 870.00), saving R417 804.00 in cost of lateness.

5.5. Type II error probability

A type II error occurs when the test indicates that the process is stable, given that the system is unstable/outside of control boundaries. This is also referred to as a false negative.

This will likely occur when the process mean (unknowingly to management) shifts upwards or downwards. Calculating the probability of making a type II error in the case where the mean shifted to 23hrs can be seen in the code below. The same logic is applied to calculate the type II error than the type I error, following a normal distribution. The likelihood of a type II error occurring is 49.55827%. These errors result in wrong delivery times, resulting in a low-quality service being delivered to customers. This could drive customers away and reflects badly on the company, and ultimately cause the company to loose money.

```
4.4
# calculate the probability of making a type 2 error using the pnorm() function
```{r}
#-----------------
# TYPE 2 ERROR
# ---------------

pnorm(22.990345, mean=23, sd=(22.990345-17.758543)/6)

```

 [1] 0.4955827
```

Figure 21: Type II error

6. PART 5: DOE and MANOVA

The MANOVA (Multivariate Analysis of Variance) test compares and analyzes the differences between two dependent variables (in this case, price, and delivery time) and the independent variable (class).

MANOVA tests avoid making type I errors (frequently encountered by ANOVA tests) by combining the information from the dependent variables and calculates group differences. The hypotheses for MANOVA are as follow:

HO: Class means are do not differ (significantly) from each other.
H1: At least one class mean is different.

Here, the results for the MANOVA test can be seen.

```
 Response Price :
                Df     Sum Sq    Mean Sq F value    Pr(>F)    
Class            6 5.7165e+13 9.5275e+12   80238 < 2.2e-16 ***
Residuals   179976 2.1370e+13 1.1874e+08                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Delivery.time :
                Df   Sum Sq Mean Sq F value    Pr(>F)    
Class            6 33461034 5576839  629489 < 2.2e-16 ***
Residuals   179976  1594464       9                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 22: MANOVA test results

R’s default for MANOVA test is to use the F-statistic. In the test results above, the p-value is insignificant (practically zero), and the null hypothesis can be rejected. This means that at least one class mean value differs from the rest.

The box plots below illustrate the results above. The distribution of delivery times for each class can be seen in the first image. The representation makes it clear that delivery times are not constant but differ for each type of product (class). The graphs also indicate that the most variation exists for Household, Gifts and Technology items.

The second graph shows the price distribution by class. Household, Luxury, and technology items have the biggest variety in these graphs. The differences in these graphs substantiate the results of the MANOVA test – although there is somewhat of a correlation between the delivery time and the price of a product, not all classes follow the same trend; with other words, they don’t all fit the correlation. It can therefore not be assumed that for higher priced items, or lager variety in prices, the delivery times will also be variable.

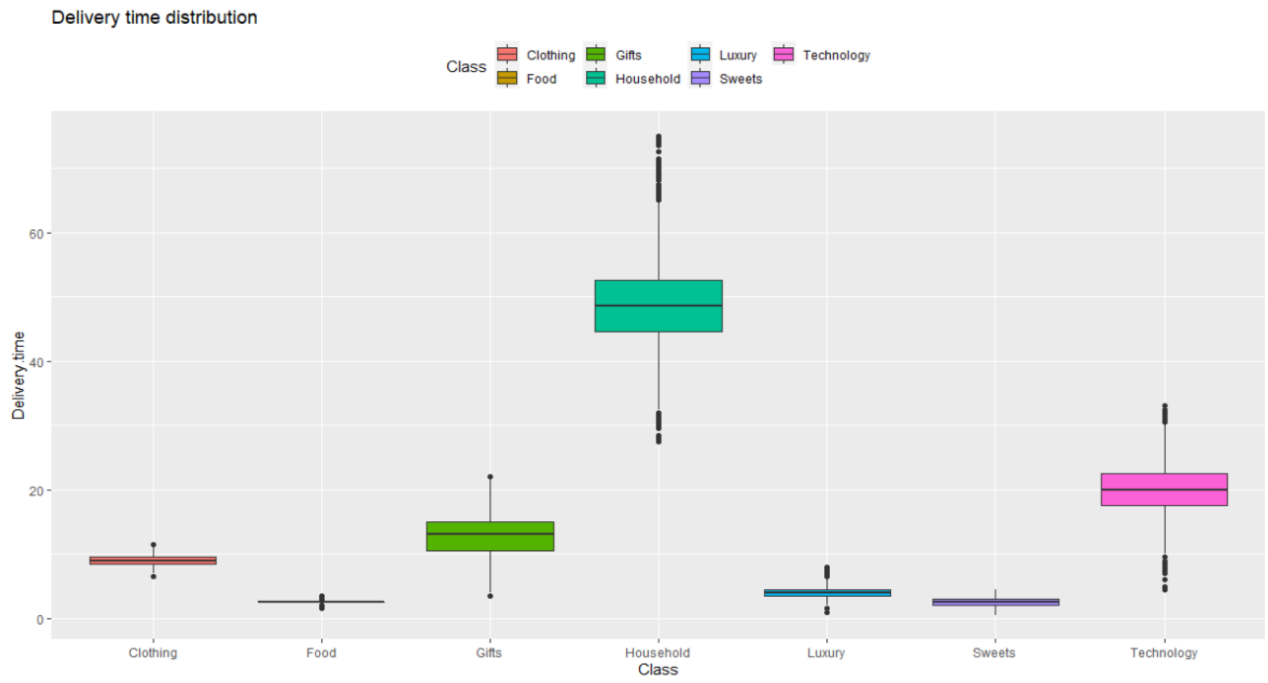


Figure 23: Delivery time distribution

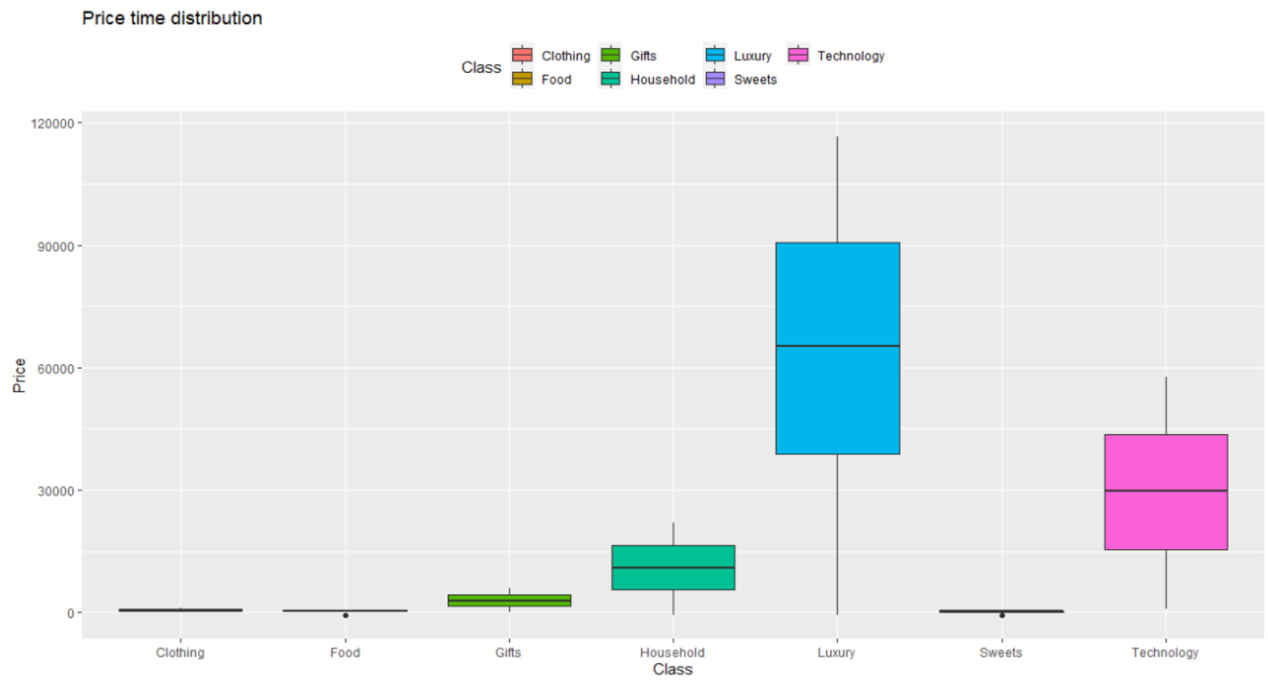


Figure 24: Price distribution

7. PART 6: Reliability of service and components

7.1. Taguchi loss function

The Taguchi loss function is a parabolic function that describes the way a customer experiences a drop in quality. Loss (of quality) is experienced the moment the product differs from the target, no matter how big or small the deviation. The function is mathematically described by $L=k(y-m)^2$. L is the loss, k is a constant (calculated: (cost to destroy defect)/(tolerance)^2), and (y-m) is the difference between the actual (y) and the target value (m).

As seen in the image below, L_1, L_2, and L_3 represent the three separate loss functions for respective tolerances and costs. The Taguchi loss for scenario 3 (Q7b) is calculated as 3.44898. This value translates to the loss (in monetary value) that the customer experiences.

```
PART 6
6.1
```{r}
# ----------------------
# Taguchi loss function
# ----------------------

# 6.1: Q6
y_2 <- 0                 # thickness of the part
k_1 <- 30/(0.035^2)      # determine the k-constant value
L_1<- k_1*(y_1-0.04)^2   # taguchi loss function

# 6.1: Q7(a)
y_2 <- 0                 # thickness of the part
k_2 <- 25/(0.035^2)      # determine the k-constant value
L_2<- k_2*(y_2-0.04)^2   # taguchi loss function

# 6.1: Q7 (b)
y_3 <- 0.027             # thickness of the part
k_3<- 25/(0.035^2)       # determine the k-constant value
L_3<- k_3*(y_3-0.04)^2   # taguchi loss function
L_3
```

 [1] 3.44898
```

Figure 25: Taguchi loss

7.2.System reliability

The reliability of a system is directly related to the quality of the product and the service offered to customers. Reliable systems are more likely to produce good quality products on time.

Given the following system:

Machine	Reliability
A	0.85
B	0.92
C	0.90

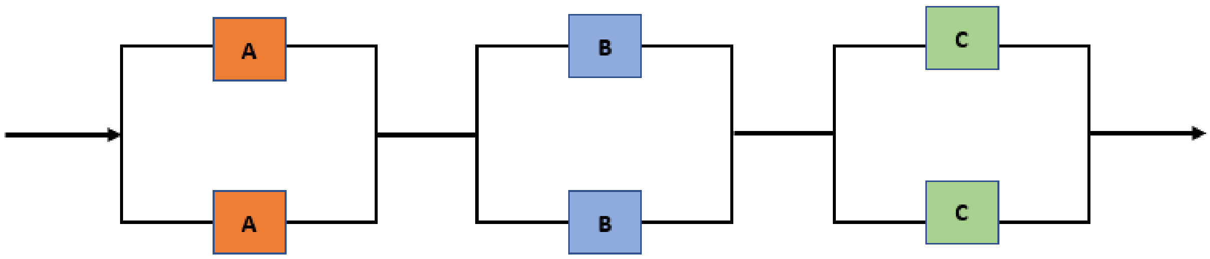


Figure 26: System illustration

Scenario 1: Only one machine in A, B and C is working.

Scenario 2: The machines are connected in parallel and both of each machine, A, B and C, are working.

```
6.2
```{r}
# ---------------------
# System Reliability
# ---------------------

A<- 0.85 # A Reliability
B<- 0.92 # B Reliability
C<- 0.90 # C Reliability

A.fail<- (1-A)^2 # Probability both A's fail
B.fail<- (1-B)^2 # Probability both B's fail
C.fail<- (1-C)^2 # Probability both C's fail

AA<- 1-A.fail # Combined reliability for A
BB<- 1-B.fail # Combined reliability for B
CC<- 1-C.fail # Combined reliability for C

r.a<- A*B*C
r.b<- AA*BB*CC
r.improve<- r.b-r.a

data.frame(r.a, r.b, r.improve)
```
```

Figure 27: System reliability

The code in the image above shows how the reliability of the system is calculated for two different scenarios. Scenario 1 is given by “r.a”, and scenario 2 is given by “r.b”. The difference (improvement) between scenario 1 (r.a) and scenario 2 (r.b) is given by “r.improve”.

RStudio: Notebook Output

Description: df [1 × 3]

r.a <dbl>	r.b <dbl>	r.improve <dbl>
0.7038	0.9615316	0.2577316

1 row

Figure 28: System reliability summary

7.3.Binomial distribution and reliability probability

System reliability is reliant on more than just machines. The reliability of transportation vehicles and employees (drivers) also contribute to whether the product is delivered on time and the quality of service the customer experiences.
Based off given statistics of truck and driver availability, and the availability required to provide reliable services, the probability of on-time deliveries was calculated.

The probability of being reliable was calculated by taking the complement of the probability of being unreliable. The latter was calculated by taking the probability of not meeting reliability specifications/ not meeting the required number of resources (drivers/delivery vehicles), p, the number of resources available (drivers/delivery vehicles), n, and the minimum number of resources required to be unreliable, x. For example, the company require a minimum of 19 vehicles to provide a reliable service. Following the cumulative binomial distribution, having 18 or less vehicles available will result in unreliable service. Thus, x is equal to 18. Based on the information given, the number of days where 21 vehicles were available was 190 and nineteen vehicles available was 22 days. Of the remaining 1348 days, eighteen or less vehicles were available. Thus, p=1348/1560; the probability of being unreliable. The total number of vehicles available is 21; n=21.

```
6.3
```{r}
# ----------------------
# Binomial distribution
# ----------------------

# Info
# d20<- 190
# d19<- 22
# d18<- 3
# d17<- 1
# days<-1560

# driver20<- 95
# driver19<- 6
# driver18<- 1

#  TRUCKS: 21
x<-18                      # number of vehicles to be considered unreliable
n<-21                      # number of vehicles
p<- (1560-22-190)/1560     # probability of being unreliable

p_21Trucks<- 1-pbinom(x, n, p,  lower.tail = TRUE, log.p = FALSE) # probability of being reliable

#  TRUCKS: 22
xx<-18                      # number of vehicles to be considered unreliable
nn<-22                      # number of vehicles
pp<- (1560-22-190-3)/1560   # probability of being unreliable

p_22Trucks<- 1-pbinom(xx, nn, pp,  lower.tail = TRUE, log.p = FALSE) # probability of being reliable

#  DRIVERS: 21
xxx<-18                         # number of vehicles to be considered unreliable
nnn<-21                         # number of vehicles
ppp<- (1560-95-6)/1560          # probability of being unreliable

p_21Drivers<- 1-pbinom(xxx, nnn, ppp,  lower.tail = TRUE, log.p = FALSE) # probability of being reliable

data.frame(p_21Trucks, p_22Trucks, p_21Drivers)
```
```

Figure 29: Binomial distribution

The same logic was applied to the problem when 22 vehicles were available and when the resource under consideration was the driver availability. The summary of the results can be seen in the image below.

p_21Trucks <dbl>	p_22Trucks <dbl>	p_21Drivers <dbl>
0.4420258	0.6399049	0.8484614

1 row

Figure 30: Reliability summary

8. Conclusion

Sales data for several years was calculated and used for statistical analysis to identify inconsistencies and areas in the business that cause additional costs.

This was achieved by performing basic descriptive statistics to get a general understanding of the data. Statistical process control values were calculated next to determine the stability of each of the processes. Optimization methods are discussed, followed by a MANOVA table and some quality assurance calculations.

To concluding the findings, unstable processes and processes deviating from the mean value calculated initially should be inspected as they could result in extra costs for the business.

9. References

ANOVA in R - how to implement one-way ANOVA from scratch (2022) *ANOVA in R - How To Implement One-Way ANOVA From Scratch*. Available at: https://appsilon.com/anova-in-r/ (Accessed: October 14, 2022).

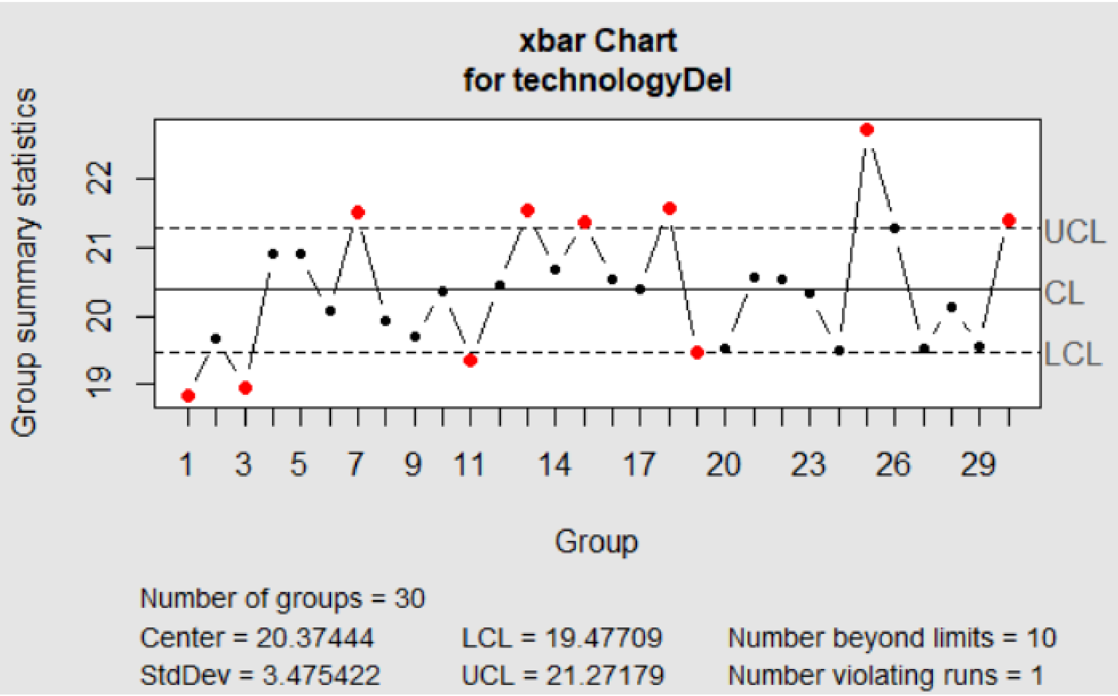
Mosaic plot statistics, *glz.redesse.it*. Available at: https://glz.redesse.it/mosaic-plot-statistics.html

(Accessed: October 14, 2022).

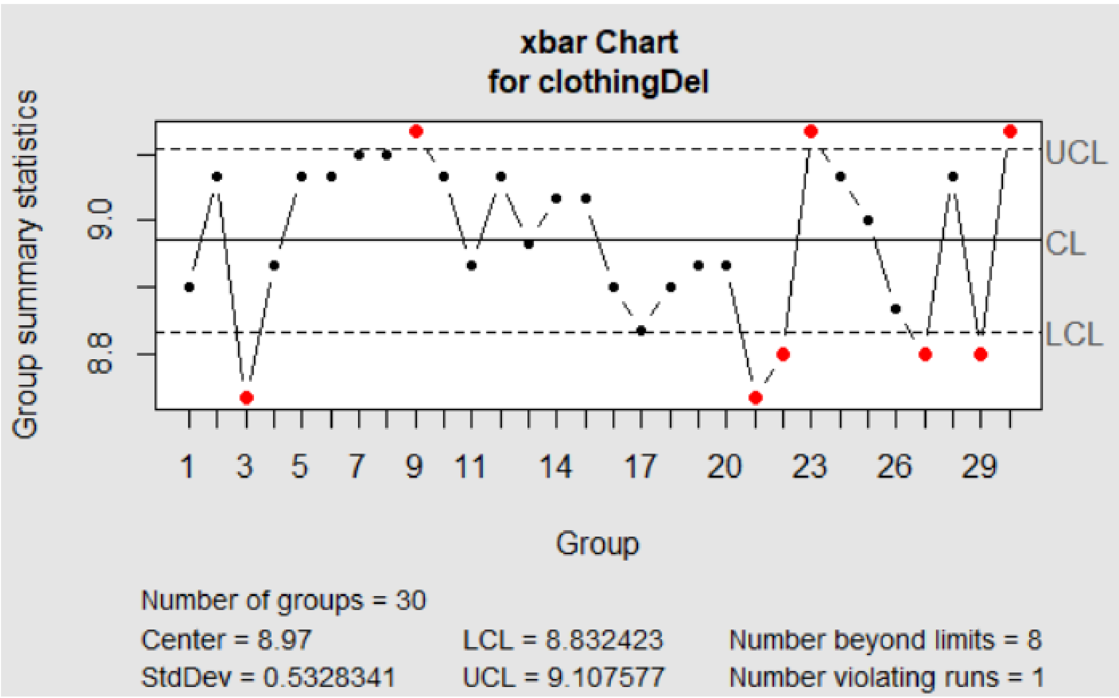
10. Appendix A: Statistical Process Control Charts

10.1. X-bar charts

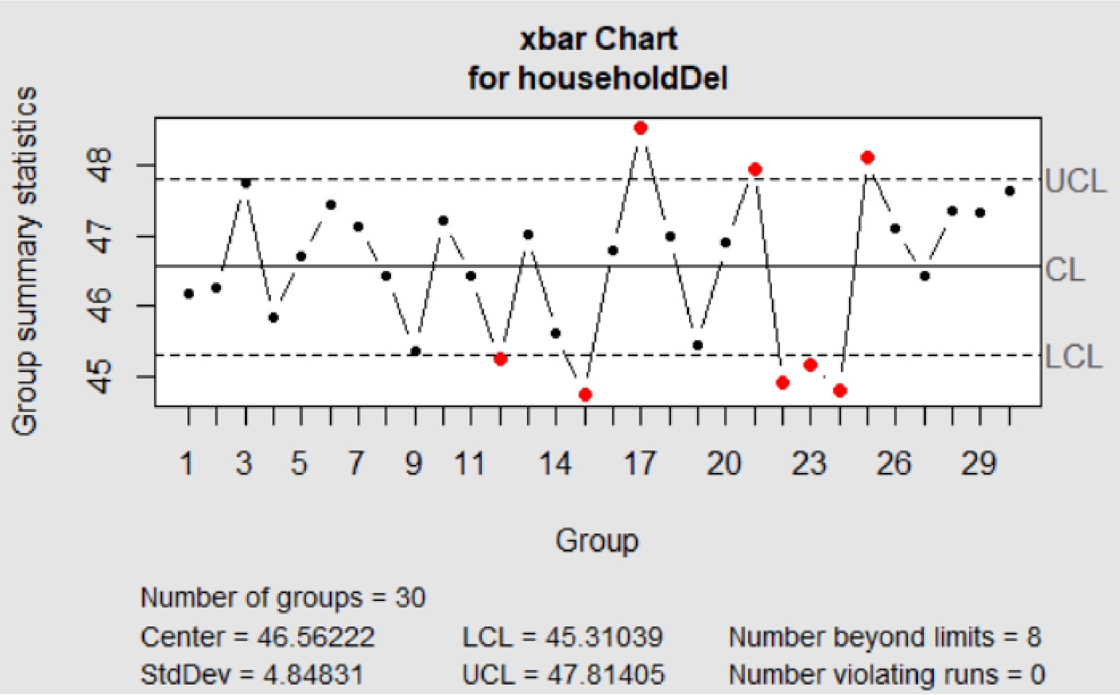
10.1.1. Technology



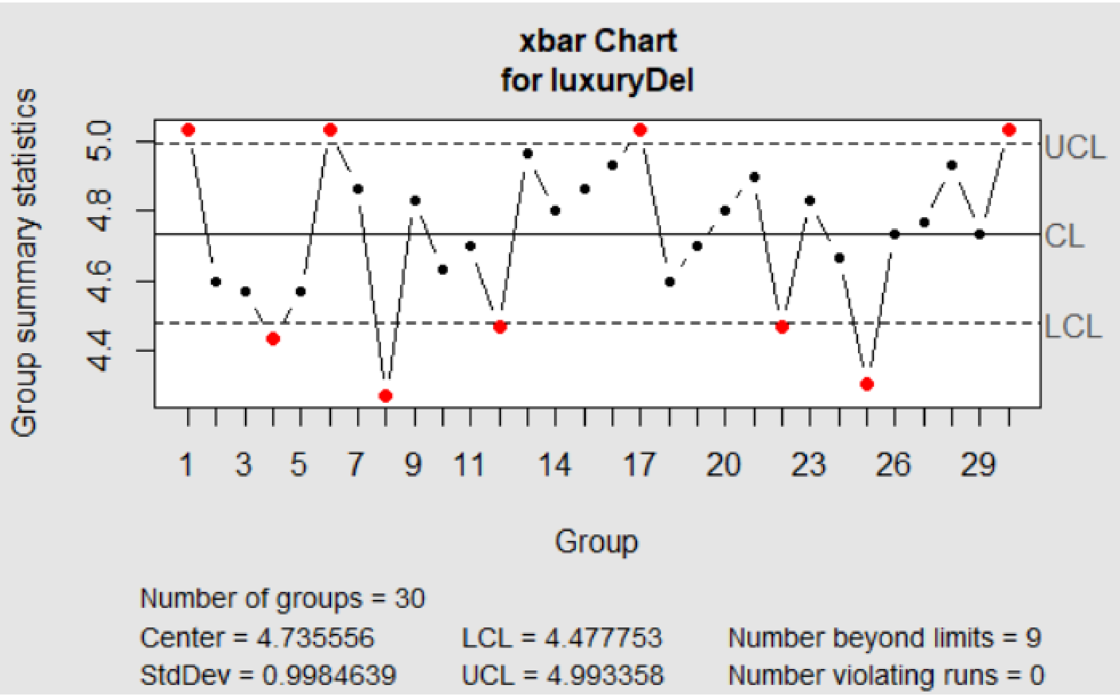
10.1.2. Clothing



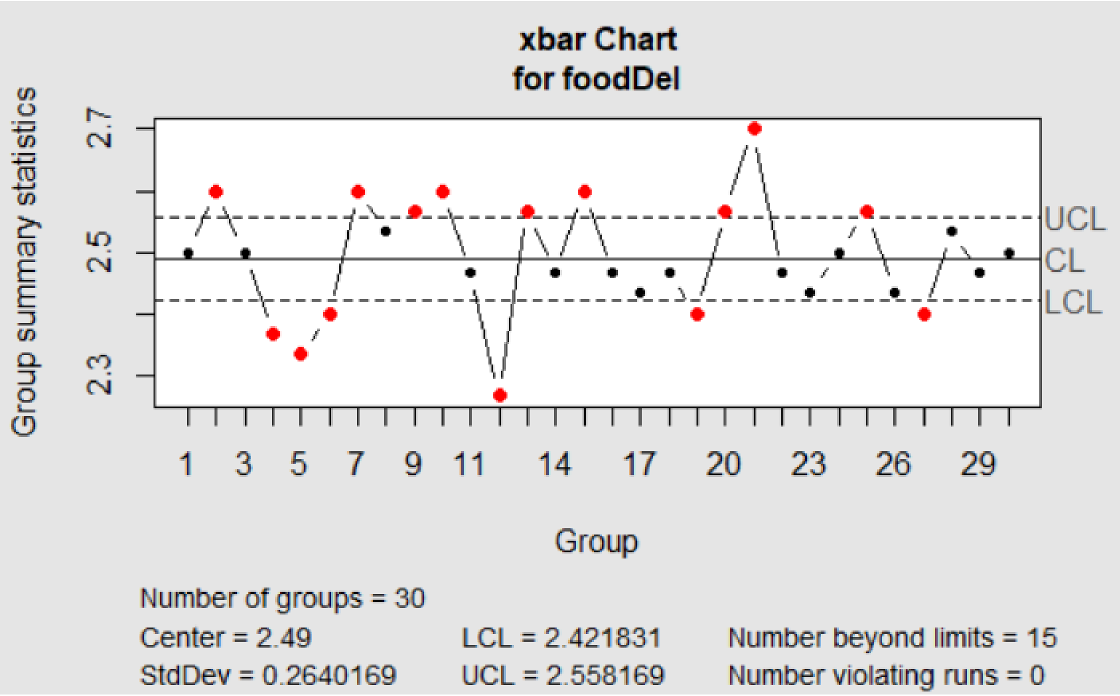
10.1.3. Household



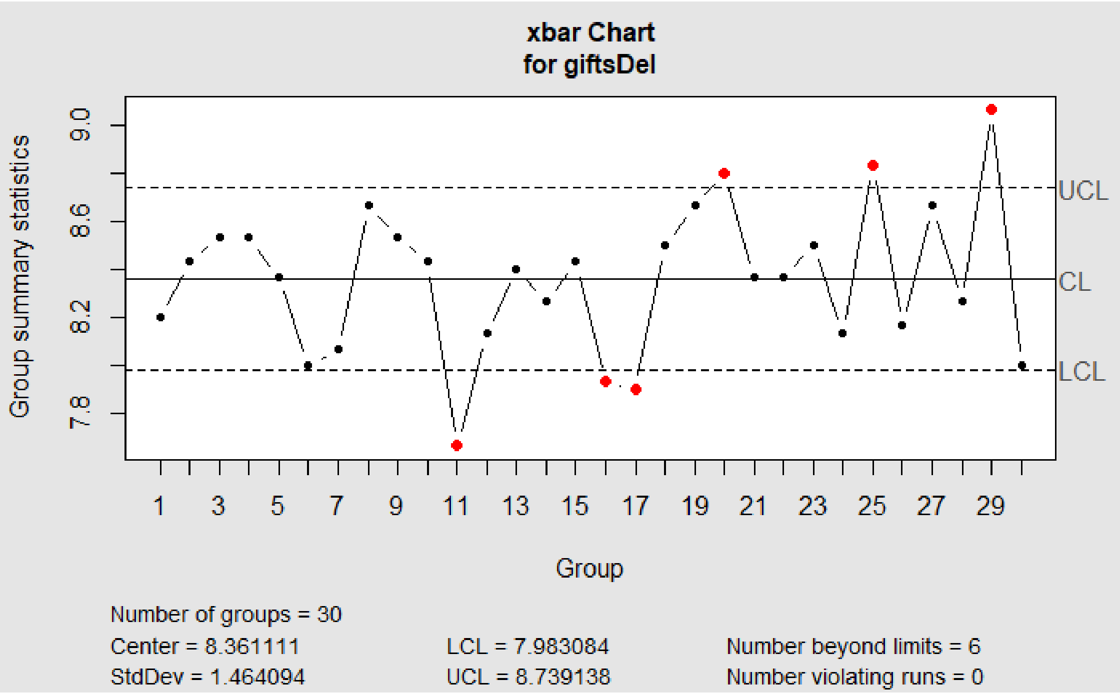
10.1.4. Luxury



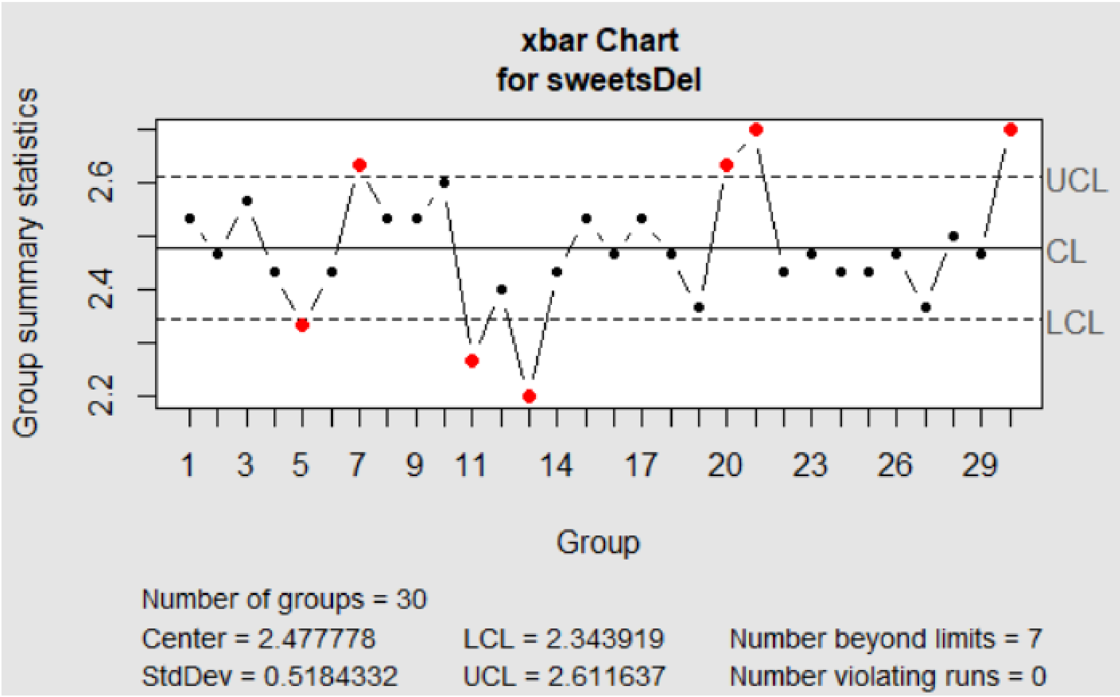
10.1.5. Food



10.1.6. Gifts

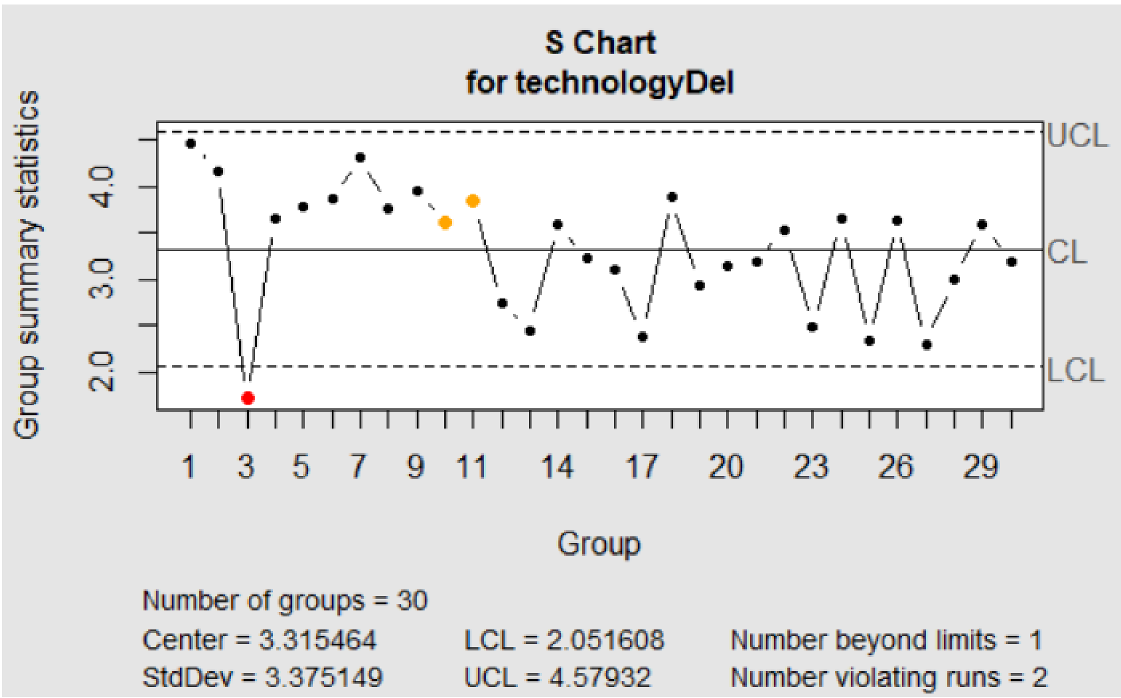


10.1.7. Sweets

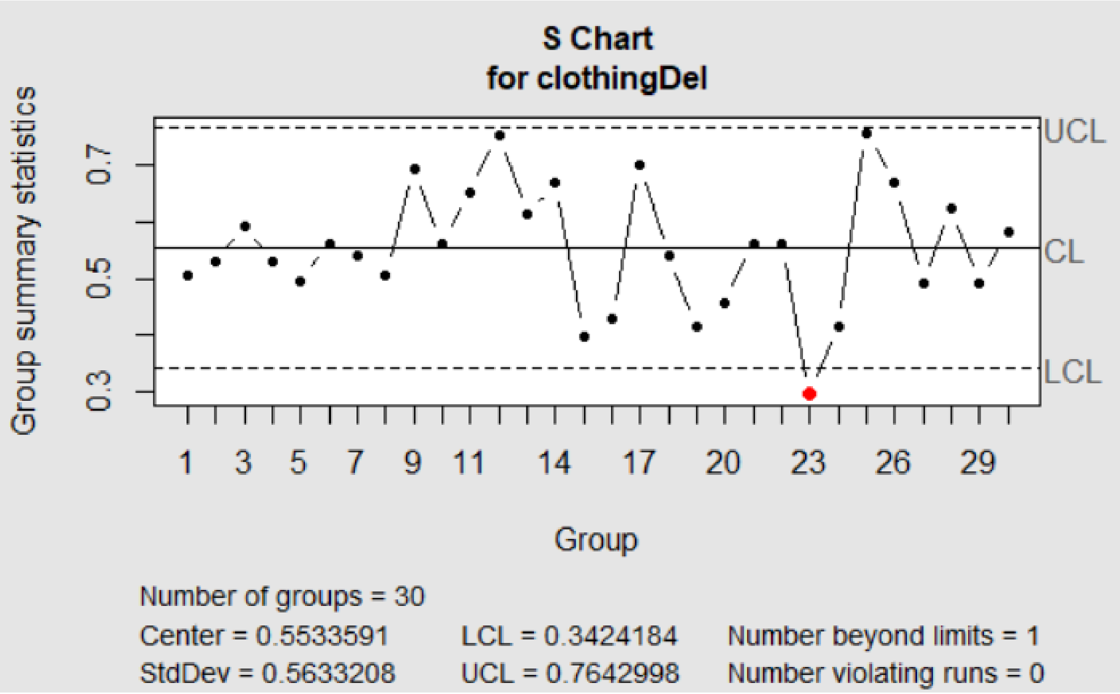


10.2. S charts

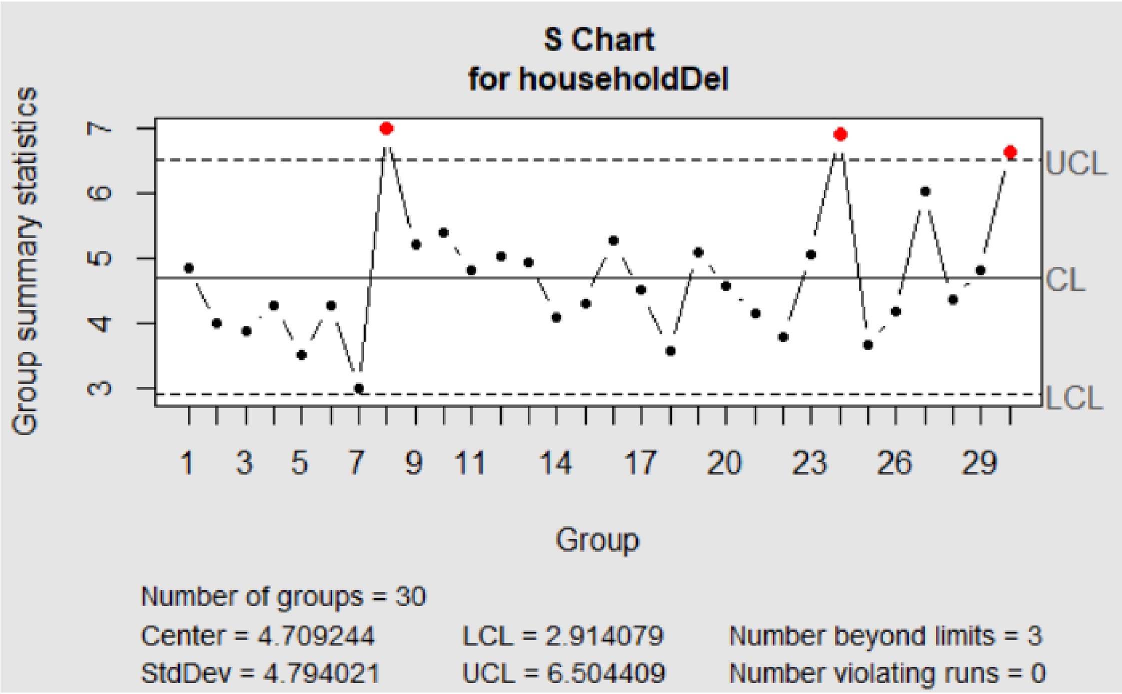
10.2.1. Technology



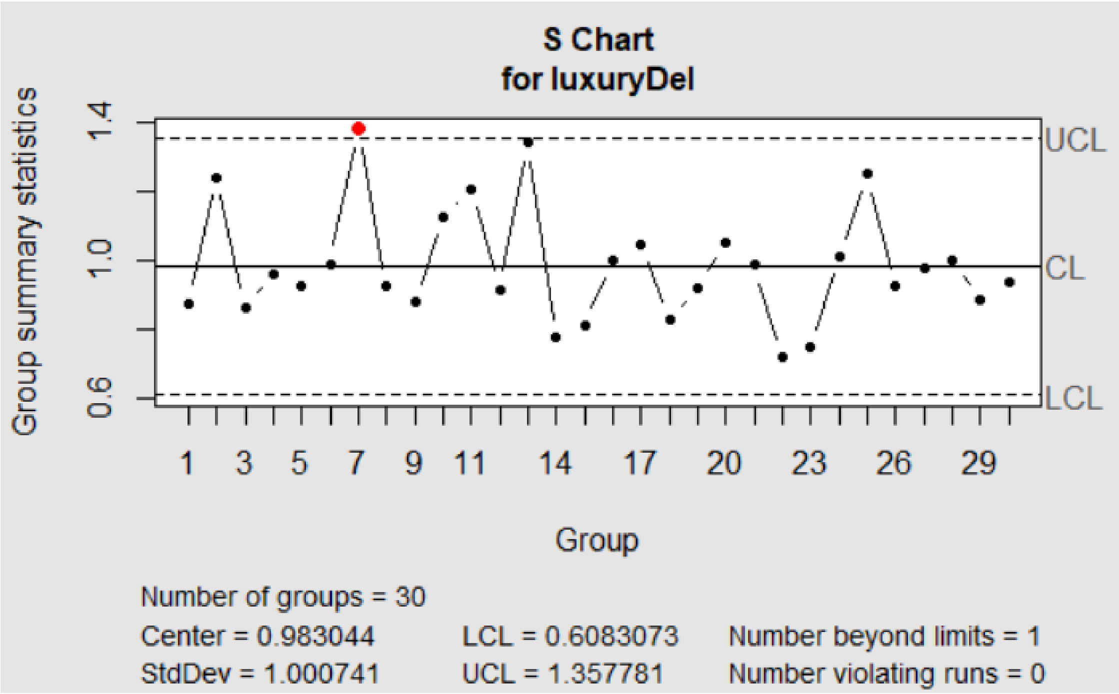
10.2.2. Clothing



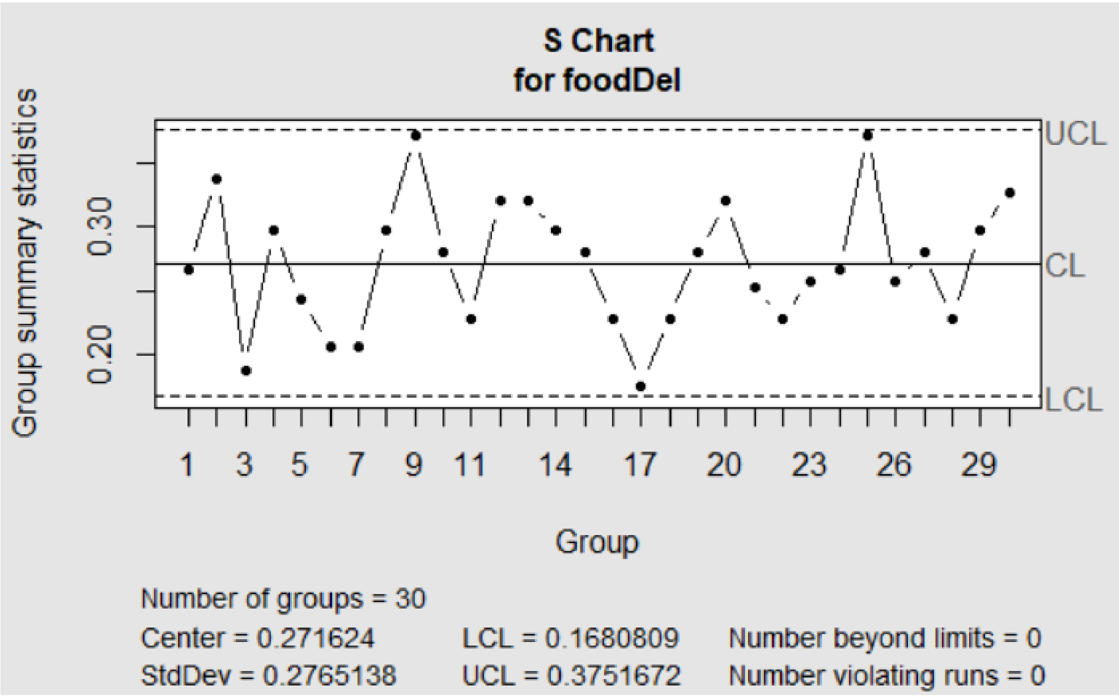
10.2.3. Household



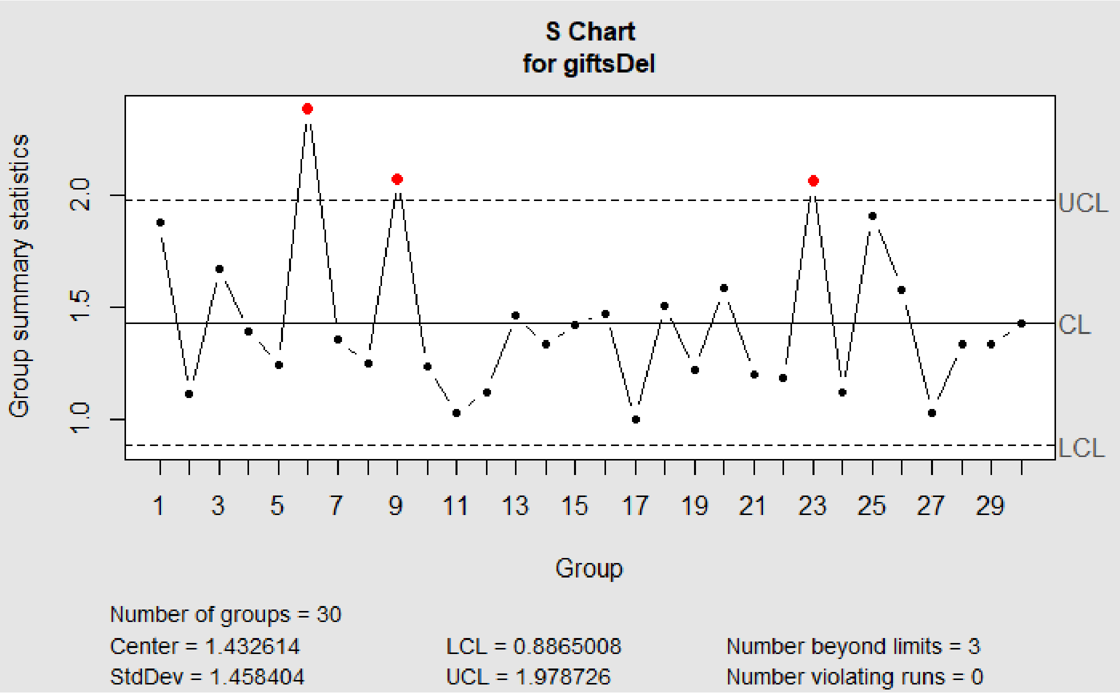
10.2.4. Luxury



10.2.5. Food



10.2.6. Gifts



10.2.7. Sweets