# ECSA Data Analysis Project

## Quality Assurance 344

Simanye Dyasi

22574115

21 October 2022

# Abstract

Data analysis is conducted on sales data provided by some online company. In order to clean the data and easier to process the data is first wrangled after which it undergoes statistical analysis in order to establish a better understanding of the data as well as the company behind it. After this the delivery process is evaluated by SPC and after all this information is combined it becomes evident that the performance of this business is Good.

# Table of Contents

Abstract ..... 2
Table of Figures ..... 3
Introduction ..... 5
Data Analysis ..... 5
Part 1: Data Wrangling ..... 5
Part 2: Descriptive statistics ..... 5
Process Capability Indices: ..... 7
Part 3: Statistical Process Control (SPC) ..... 8
3.1 S Charts ..... 8
3.2 X bar charts ..... 12
Part 4: Optimizing the Delivery process ..... 15
Part 6:Reliability of the Service and Products ..... 16
Taguchi loss function ..... 16
Chapter 7, problem 6 ..... 16
Chapter 7, problem 7 ..... 16
Problem 27: System Reliability ..... 16
Conclusion ..... 17
References ..... 17

# Table of Figures

Figure 1:S chart for Technology ..... 8
Figure 2: S chart for Clothing ..... 9
Figure 3: S chart for Household ..... 9
Figure 4: S chart for Luxury ..... 10
Figure 5: S chart for Food ..... 10
Figure 6: S chart for Gifts ..... 11
Figure 7: S chart for Sweets ..... 11
Figure 8: X-bar chart for Technology ..... 12
Figure 9: X-bar chart for Clothing ..... 12
Figure 10: X-bar chart for Household ..... 13
Figure 11:X-bar chart for Luxury ..... 13
Figure 12: X-bar chart for Food ..... 14

Figure 13: X-bar chart for Gifts ..... 14
Figure 14: X-bar chart for Sweets ..... 15

# Introduction

Given the client data from an online business and analysis of the data is done using data analysis techniques in R/R studio. The project is subdivided into 3 parts namely, part 1: data wrangling, part 2: descriptive statistics, part 3: statistical process control. Then we draw a conclusion based on the data analysis techniques used.

# Data Analysis

## Part 1: Data Wrangling

Data wrangling can be considered as improving the quality of the data by removing data instances that have data quality issues for example missing values and invalid entries. This is done by reading all the sales data into R. After that the two main data quality issues with this data was identified to be missing values and negative values. These values are then Isolated and removed from the rest of the data leaving only the valid data instances. Having initially had 180000 instances of data in the original data set, 179978 instances remained as the valid data instances.

Furthermore, the invalid data instances are also grouped separately from the valid instances. A primary key is added to the valid data and thereafter to the invalid data. This primary key is used as a type of index since the data was separated and the existing index is no longer accurate.

## Part 2: Descriptive statistics

During this section the data is analyzed to identify any patterns and trends that exist in the data, using statistical metrics. An efficient and effective way of doing this is formulating a data report table, which provides a statistical summary of each feature of the data.

In order to formulate this data report cardinality is used to identify and separate the features of that are categorical and those that are continuous, the logic being that the categorical feature have low cardinality due to having a fixed number of possible values, and the continuous features have higher cardinalities as the is an almost infinite possible number of values for them.

*Table 1: Data Report all features*

| | Feature | Count | % missing | Cardinality | Mode | Mode freq. | Mode % | 2nd Mode | 2nd Mode freq. | 2nd mode % |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Class | 179978 | 0 | 7 | Gifts | 39149 | 21.75 | Technology | 36347 | 20.20 |
| 2 | Year | 179978 | 0 | 9 | 2021 | 33443 | 18.58 | 2029 | 22475 | 12.49 |
| 3 | Month | 179978 | 0 | 12 | 12 | 15225 | 8.46 | 10 | 15221 | 8.46 |
| 4 | Day | 179978 | 0 | 30 | 17 | 6126 | 3.40 | 25 | 6122 | 3.40 |
| 5 | Why.Bought | 179978 | 0 | 6 | Recommended | 106985 | 59.44 | Website | 29447 | 16.36 |

Table 1 gives a statistical summary of the features it includes the count which is the total number of instances that have a value for the feature. The percentage missing values in each feature gives the percentage of data instances that don't have a value for the feature. The cardinality of the feature, the modal value for that feature, the frequency of the mode, percentage mode in the data, the second mode, second mode frequency and percentage second mode are also included.

*Table 2: Data report for Continuous features*

| | Feature | Count | % missing | Cardinality | Min | 1st Qrt | Mean | Median | 3rd Qrt | Max | Std Dev. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | AGE | 179978 | 0 | 91 | 18.00 | 38.00 | 54.56552 | 53.00 | 70.00 | 108 | 20.38881 |
| 2 | Price | 179978 | 0 | 78832 | 35.65 | 482.31 | 12294.09837 | 2259.63 | 15270.97 | 116619 | 20889.15025 |
| 3 | Delivery.time | 179978 | 0 | 148 | 0.50 | 3.00 | 14.50031 | 10.00 | 18.50 | 75 | 13.95578 |

Table 2 summarizes the features Age, Price and Delivery time which are the continuous. The table includes some of the information from Table 1 namely count, percentage missing and cardinality but the Five number summary is included (Minimum, first quartile, median, third quartile and maximum), along with the mean and standard deviation of each feature.

*Table 3: The table of cut offs, min and max*

| | Feature | Lower cut off | Min | Lower outliers? | Higher cut off | Max | Upper outliers? |
|---|---|---|---|---|---|---|---|
| 1 | AGE | -10.00 | 18.00 | no | 118.00 | 108 | no |
| 2 | Price | -21700.68 | 35.65 | no | 37453.96 | 116619 | yes |
| 3 | Delivery.time | -20.25 | 0.50 | no | 41.75 | 75 | yes |

Table 3 is a type of distribution summary and gives a better idea of outliers in the data, minimums and maximums, lower and upper outliers.

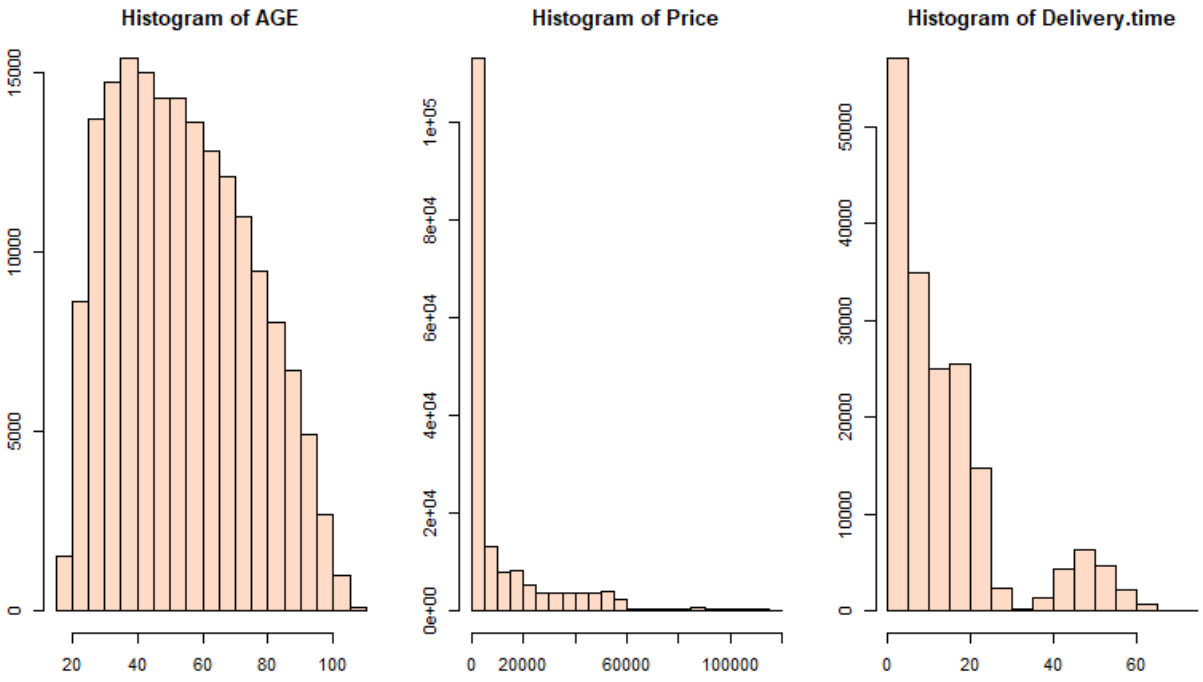


*Figure 1:The sales distributions for AGE, Price and Delivery time*

Figure 1, shows the distribution of the sales for each of the continuous features. Age is left skewed with a fairly normal distribution this shows that the younger generation makes up a larger amount of their customers. The price has a negatively exponential distribution that's skewed left, it illustrates that the

products that are priced lowest contribute the most to the total sales made by the company. The Delivery time has a unique distribution that looks like a left skewed distribution with a smaller normal distribution on its right-hand tail and this illustrates that most of their sales are made of quickly delivered products, the smaller normal distribution could indicate a niche market that buys products that take a long time to ship.

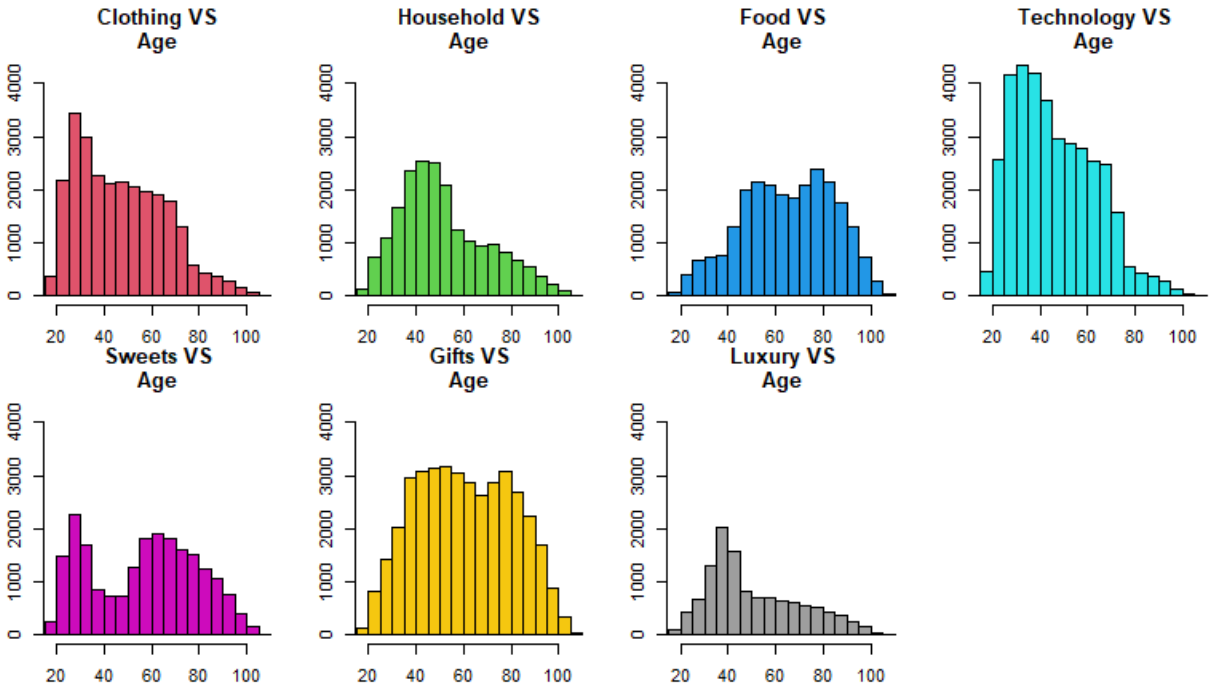


*Figure 2: Age distribution for Each class*

The Age distributions across different product classes in Figure 2 depict a similar trend cross most product classes, the distribution of data is skewed left. This excludes Food and Gifts, Food has a more right skewed distribution and Gifts have a more normal centered distribution. This tell us that in most product classes young people dominate the customer base but in Food the more older customers and that people in the 40 to 80 year range dominate the Gifts customers.

## Process Capability Indices:

Process Capability (Cp) estimates what the process is capable of producing if the mean is centered between upper specification limit USL and lower specification limit LSL. Upper sided index(Cpu) Indicates process capability estimate for specifications that consists of only the USL. R was used to calculate these values as follows;

```
#process capabilities
#The delivery time of tech items isolated
TD = valid[which(valid$Class == "Technology"), 10]
std = sd(TD)
Ex = mean(TD)
Cp = (24-0)/(6*std)
CpU = (24-Ex)/(3*std)
CpL = (Ex - 0)/(3*std)
CpK = min(CpU,CpL)
```

Cp= 1.14  Cpu= 0.380  Cpl= 1.90  Cpk=0.380

It makes sense for the lower specification limit to be 0 in this case as you can deliver a product before its been ordered that's the only way the delivery time could be less than 0, and its possible for it to be 0 if the buyer is buying in person so the LSL makes sense at 0.

## Part 3: Statistical Process Control (SPC).

In this section we use statistical process control to evaluate the process. Statistical process control is a methodology used to control and monitor a process by identifying special causes of variation and implementing corrective actions (J.R. Evans, 2020). The data is thus be used to perform a SPC to demonstrate the quality capability of the delivery process in the online business.

The data is first sorted by Year, Month, Day and X. Then by starting with the oldest valid data and sampling a sample size of 15 sales, the first 30 samples are used from the data to create X-bar and S charts.

### 3.1 S Charts

For the 30 groups of samples the standard deviation are plotted for each class to generate the S charts below. The charts show that the delivery process is low in variability and thus considerably high quality because for most  classes, most of sample standard deviations fall within 2 standard deviations of the average standard deviation of all samples.

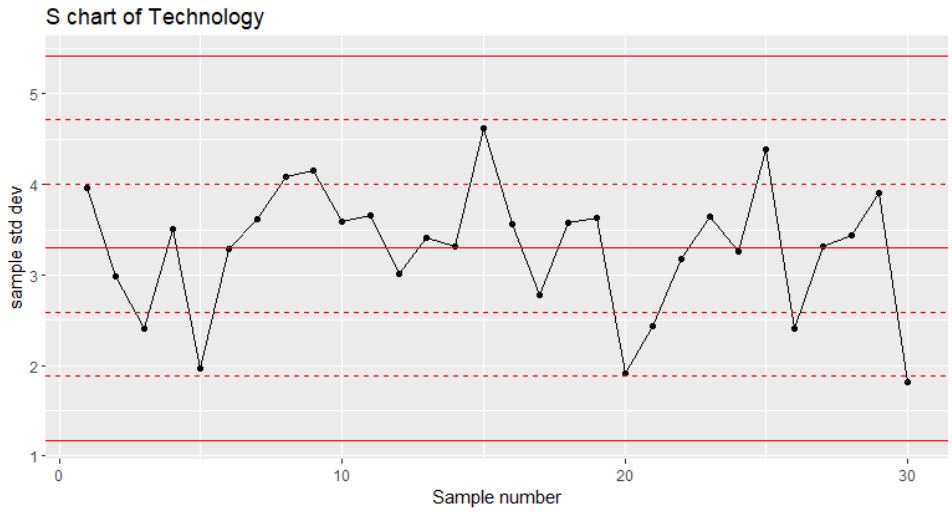


*Figure 3:S chart for Technology*

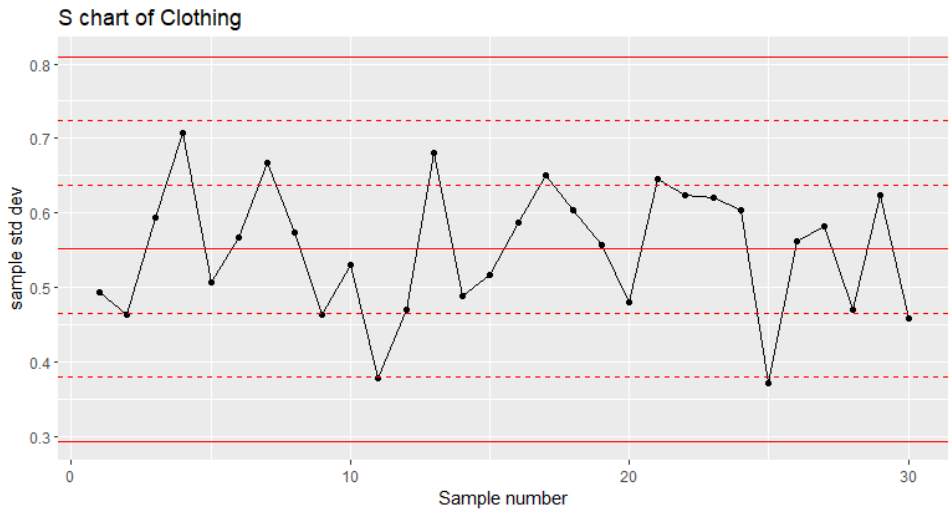


*Figure 4: S chart for Clothing*

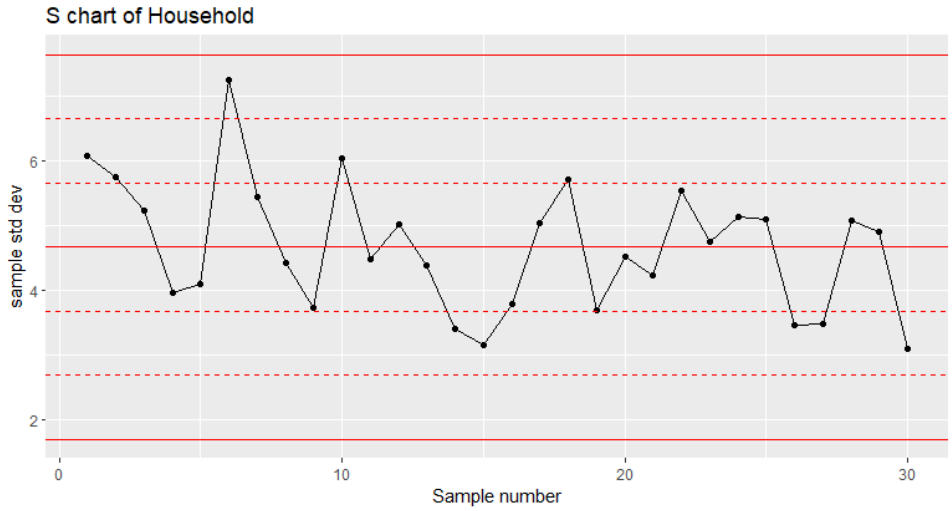


*Figure 5: S chart for Household*

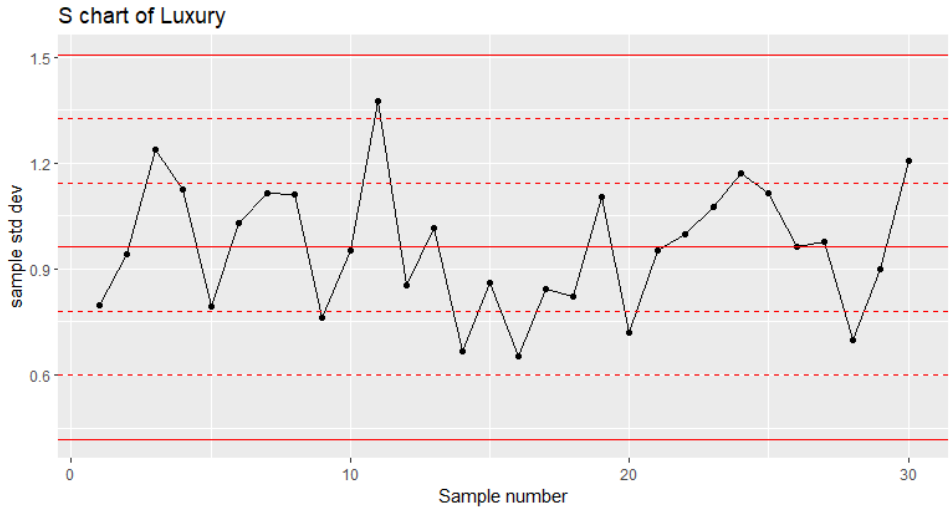


*Figure 6: S chart for Luxury*

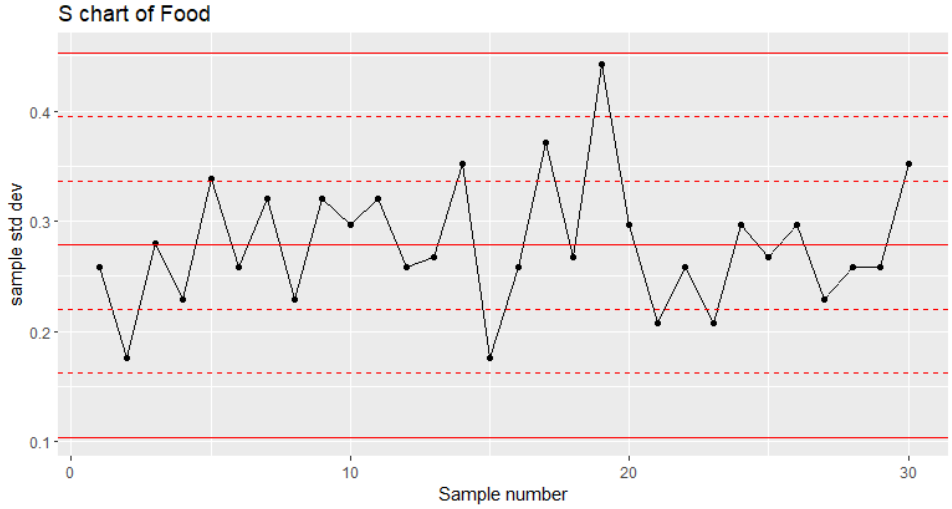


*Figure 7: S chart for Food*

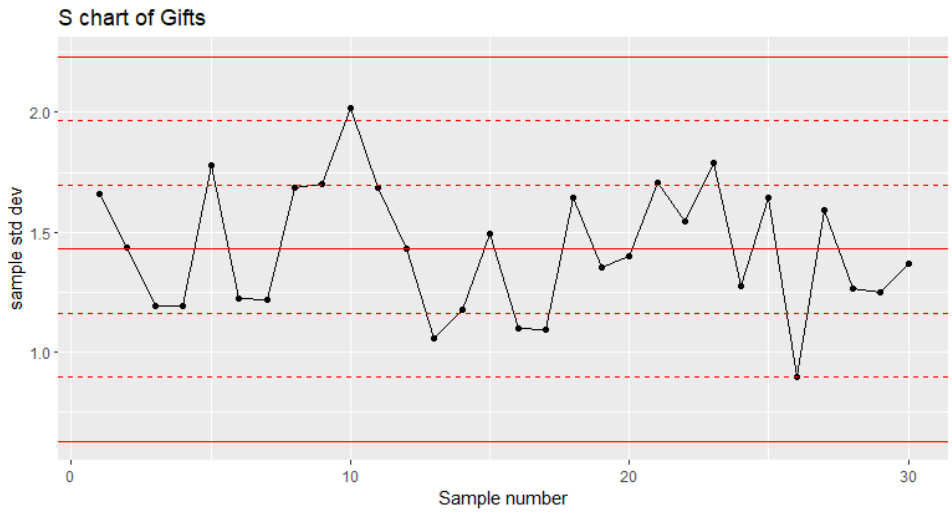


*Figure 8: S chart for Gifts*

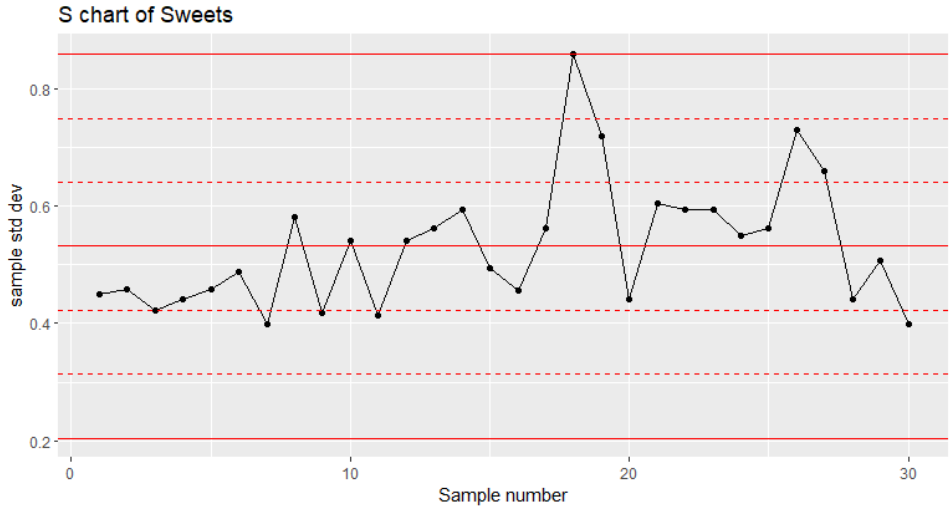


*Figure 9: S chart for Sweets*

*Table 4: S chart limits*

| | Class | UCL | U2Sigma | U1Sigma | CL | L1Sigma | L2Sigma | LCL |
|---|---|---|---|---|---|---|---|---|
| 1 | Technology | 4.00 | 4.71 | 5.42 | 3.30 | 2.59 | 1.88 | 1.17 |
| 2 | Clothing | 0.64 | 0.72 | 0.81 | 0.55 | 0.47 | 0.38 | 0.29 |
| 3 | Household | 5.66 | 6.65 | 7.64 | 4.67 | 3.68 | 2.69 | 1.70 |
| 4 | Luxury | 1.14 | 1.32 | 1.51 | 0.96 | 0.78 | 0.60 | 0.42 |
| 5 | Food | 0.34 | 0.39 | 0.45 | 0.28 | 0.22 | 0.16 | 0.10 |
| 6 | Gifts | 1.70 | 1.96 | 2.23 | 1.43 | 1.16 | 0.89 | 0.63 |
| 7 | Sweets | 0.64 | 0.75 | 0.86 | 0.53 | 0.42 | 0.31 | 0.20 |

From the values in Table 4 and the graphs in Figure 3 to Figure 9 it is evident that all the processes for every category of items sold in in control. None of the samples have a distribution spread that is too large, therefore the s-charts for all the classes are in control. Common causes in the processes create variation in the spread of distributions, therefore, the process of each chart can be predicted within certain limits for different processes of the different classes. Since none of the s-charts are out of control the x-charts, can be plotted and evaluated.

## 3.2 X bar charts

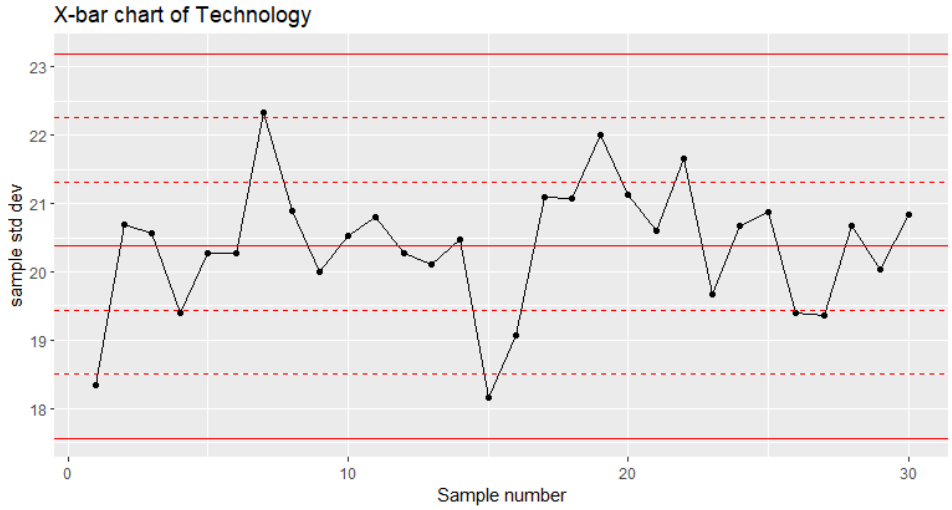


*Figure 10: X-bar chart for Technology*

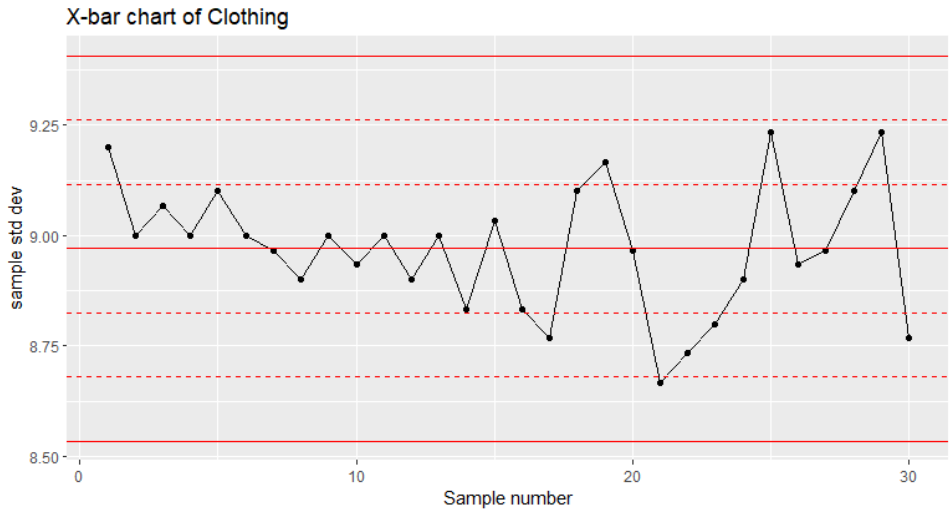


*Figure 11: X-bar chart for Clothing*

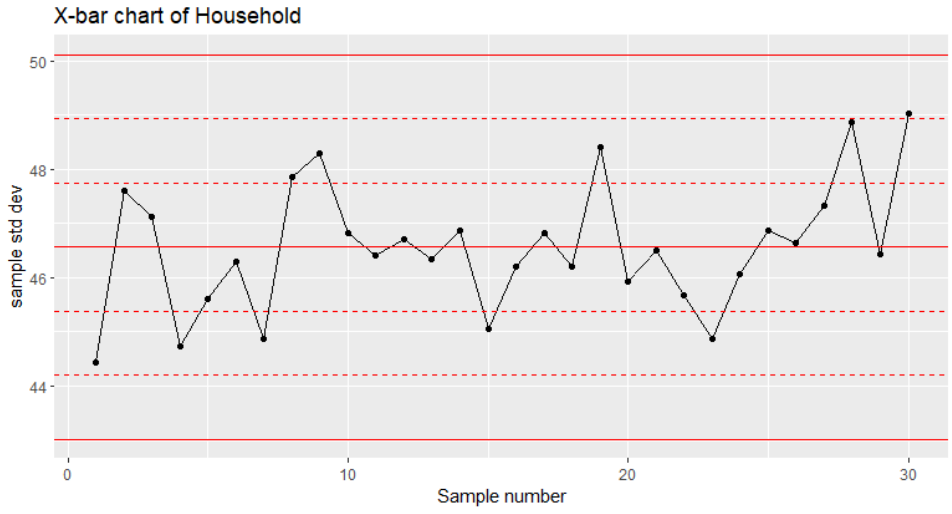


*Figure 12: X-bar chart for Household*

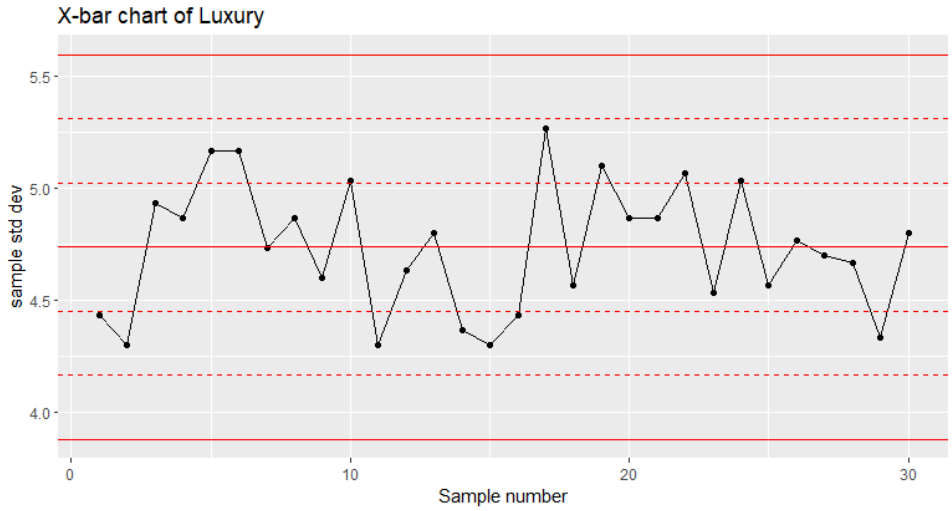


*Figure 13:X-bar chart for Luxury*

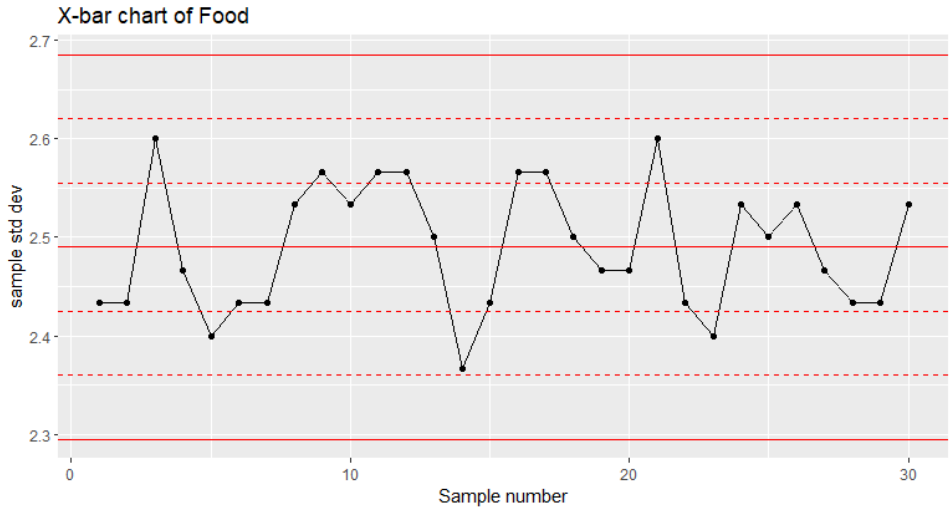


*Figure 14: X-bar chart for Food*

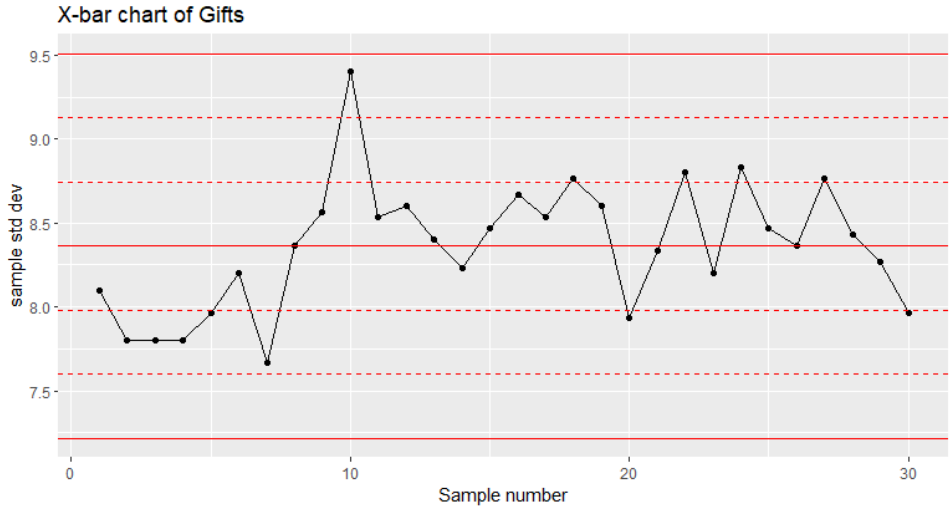


*Figure 15: X-bar chart for Gifts*

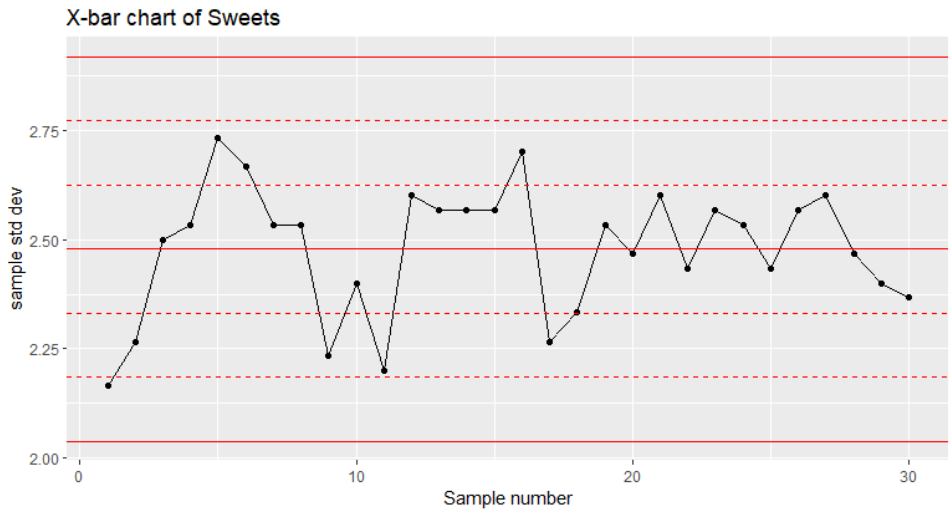


*Figure 16: X-bar chart for Sweets*

*Table 5: X-bar Chart Limits*

| | Class | UCL | U2Sigma | U1Sigma | CL | L1Sigma | L2Sigma | LCL |
|---|---|---|---|---|---|---|---|---|
| 1 | Technology | 23.19 | 22.25 | 21.31 | 20.37 | 19.44 | 18.50 | 17.56 |
| 2 | Clothing | 9.41 | 9.26 | 9.12 | 8.97 | 8.82 | 8.68 | 8.53 |
| 3 | Household | 50.11 | 48.93 | 47.74 | 46.56 | 45.38 | 44.20 | 43.02 |
| 4 | Luxury | 5.59 | 5.31 | 5.02 | 4.74 | 4.45 | 4.16 | 3.88 |
| 5 | Food | 2.69 | 2.62 | 2.56 | 2.49 | 2.42 | 2.36 | 2.29 |
| 6 | Gifts | 9.51 | 9.12 | 8.74 | 8.36 | 7.98 | 7.60 | 7.22 |
| 7 | Sweets | 2.92 | 2.77 | 2.62 | 2.48 | 2.33 | 2.18 | 2.04 |

The center line (CL) in the tables above is the mean of the Delivery Time of the 30 samples used. The upper control limit (UCL) is three sigma standard deviations below the mean and the lower control limit (LCL) is three sigma standard deviations above the mean. UCL2 and UCL1 are two and one sigma standard deviations below the mean respectively. Similarly, LCL1 and LCL2 are one and two sigma standard deviations above the mean, respectively. This counts for both the s-charts and the x-charts.

The x-charts that are displayed in Figure 10 to Figure 16 show that the processes for all the classes is in control. None of the samples fall below the LCL or above the UCL. This means the process is working well at high quality.

## Part 4: Optimizing the Delivery process

## Part 6:Reliability of the Service and Products

### Taguchi loss function

The Taguchi Loss Function is used to determine how much the current quality of a product is costing the business. Genichi Taguchi, a Japanese engineer defined quality as the economic value of reducing variation and producing the nominal specifications (J.R. Evans, 2020). The quality is measured as the variation of the target value of a design specification, then it is translated into the monetary loss associated the quality deficit.

The Taguchi Loss Function is given by $L(x)=k(x-T)2$, where

$L(x)$ - $monetary\ loss\ (assumed\ to\ increase\ quadratically)$
$T$- $target\ value\ so\ as\ to\ optimise\ performance$
$x$ - $any\ actual\ value\ of\ the\ quality\ characteristic$
$(x-T)$ - $deviation\ from\ the\ target\ value\ T$
$k$ - $contant,associated\ with\ the\ cost\ of\ a\ cetain\ deviation\ from\ T$

### Chapter 7, problem 6

Given a nominal target of 0.06 cm and tolerance of +-0.04 cm, with a scraping cost of $45 for every part out of spec.

$L(x) = k(x)^2$
k = $45/(0.04)^2$
= 28125
Thus the constant value for k in the Taguchi Loss Function is 28125
$L(x) = 28125(x - T)^2$

### Chapter 7, problem 7

Given that a nominal target and tolerance of 0.06+-0.04 and a scraping cost of $35 for an out of spec part;

$L(x) = k(x)^2$
k = $35/(0.04)^2$
= 21875
Thus the constant value for k in the Taguchi Loss Function is 21875
$L(x) = 21875(x - T)^2$
b) Process deviation from target to be reduced to 0.027 cm?
$L(0.027) = 21875(0.027)^2$

Taguchi Loss Function is = $15.95 per part which is a relatively high loss of cost per part this will drive up cost of the company, so the quality of the product must be increased to reduce this high cost. The deviation on the part must be lower even more so that the quality of their promised product or part is better to the customer meaning more customers will buy because it is of great quality parts.

### Problem 27: System Reliability

a) Formula: $R_a \times R_b \times R_c$ , assuming one machine at each stage
$R_aR_bR_c$ = 0.85 x 0.92 x 0.90 = 0.7038 = 0.70 system reliability

This means that the probability that all the system machinery to all be working at a give point in time is 70% and subsequently there a 30% probability that at some point the system will fail i.e breakdowns will occur and need to be fixed thus costing the business money.

b) $R$aa $\times$ $R$bb $\times$ $R$cc = [1-(1 − $R$a )^2] [1-(1 − $R$b )^2] [1-(1 − $R$c )^2] , if there are two machines

at each stage

$R$aa$R$bb$R$cc = [1-(1 − 0.85)^2] [1-(1 − 0.92)^2] [1-(1 − 0.90)^2]

= 0.96153156 = 0.96 system reliability

A 96% reliability indicates that there's only a 4% probability that break downs that result in the system having to stop will occur. This also means that the system will be much more efficient with two machines and will be more effective In achieving the company objective.

## Conclusion

Based on the data analysis techniques and process metrices the business looks to be in Good working order, the quality of the delivery services are confirmed to be high based on the SPC charts. The System reliability has been proven to increase upon more equipment/machinery being made available.

## References

J.R. Evans, W. L. (2020). *Managing for Quality and Peformance Excellence* (11th ed.).