Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

forward together
sonke siya phambili
saam vorentoe

ECSA Graduate Attributes Project

by

M.S.V. Elliott
23601825

Technical Report
Quality Assurance 344
Faculty of Engineering, Stellenbosch University and ECSA

10 October 2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The client data of a company was provided for analysis. The data set contained 180 000 instances, and eight input features. These features are: Age (of client) , Class (the type of product purchased), Price(of the item), Year, Month and Day (of the purchase), Delivery.time and finally Why.Bought (the reason for the purchase).

In this report, statistical analysis will be performed on this data set provided. The data will first be cleaned. Next, Descriptive statistics will show various visualizations of the data, in order to provide further information about the data and relationships that may exist between different features. In Statistical Process Control, Charts (s and x-bar) will be initialized to determine whether a process is in or out of control. Thirty samples will be used to create s-charts to see if the limits for the x-bar chart can be calculated correctly. The x-bar chart can then be used to determine whether the process is in or out of control.

PART 1 DATA WRANGLING

The data in SalesTable2022,comprisizing of 180 000 instances and 9 input features, was analysed and then separated into a valid data set and an invalid data set. The data instance was considered invalid if it was listed as NA, i.e. missing values were removed from the data set. Any negative values in the Price column were also removed. The data instances removed from the original data set formed the invalid data set. The remainder of the data instances formed the valid data set.
The results were as follows:
The valid data set (ValidBest) contained 179 978 valid observations
The invalid data set (InvalidBest) contained 22 observations

The Primary key simply lists the number of the new recorded invalid instance in the data frame. The secondary key in the data frame is the original index of the instance, meaning the original number of where this instance was contained in the original excel file SalesTable2022.

Table 1 Head of Invalid Dataset

PrimaryKeyIV <int>	X <int>	ID <int>	AGE <int>	Class <chr>	Price <dbl>	Year <int>	Month <int>	Day <int>
1	12345	18973	93	Gifts	NA	2026	6	11
2	16321	81959	43	Technology	NA	2029	9	6
3	19541	71169	42	Technology	NA	2025	1	19
4	19999	67228	89	Gifts	NA	2026	2	4
5	23456	88622	71	Food	NA	2027	4	18
6	34567	18748	48	Clothing	NA	2021	4	9

The 9 input features contained within the data set:

1. ID: index for identifying purchase entry
2. Age: age of the customer who made the purchase
3. Class: The type of product purchased. The products purchased were grouped into 7 classes: Gifts, Technology, Clothing, Household, Sweets, Food and Luxury
4. Price: the price of the purchase
5. Year: year of purchase
6. Day: day of purchase
7. Month: month of purchase
8. Delivery.time: time taken to delivery the purchase
9. Why.Bought: reason that the customer purchased from this particular company

PART 2 DESCRIPTIVE STATISTICS

GRAPHICAL REPRESENTATION OF DATA

Here, various graphical representations of the input features are provided. This helps to visualise the features of the data, in order to easily gain further insight and understanding of these features.

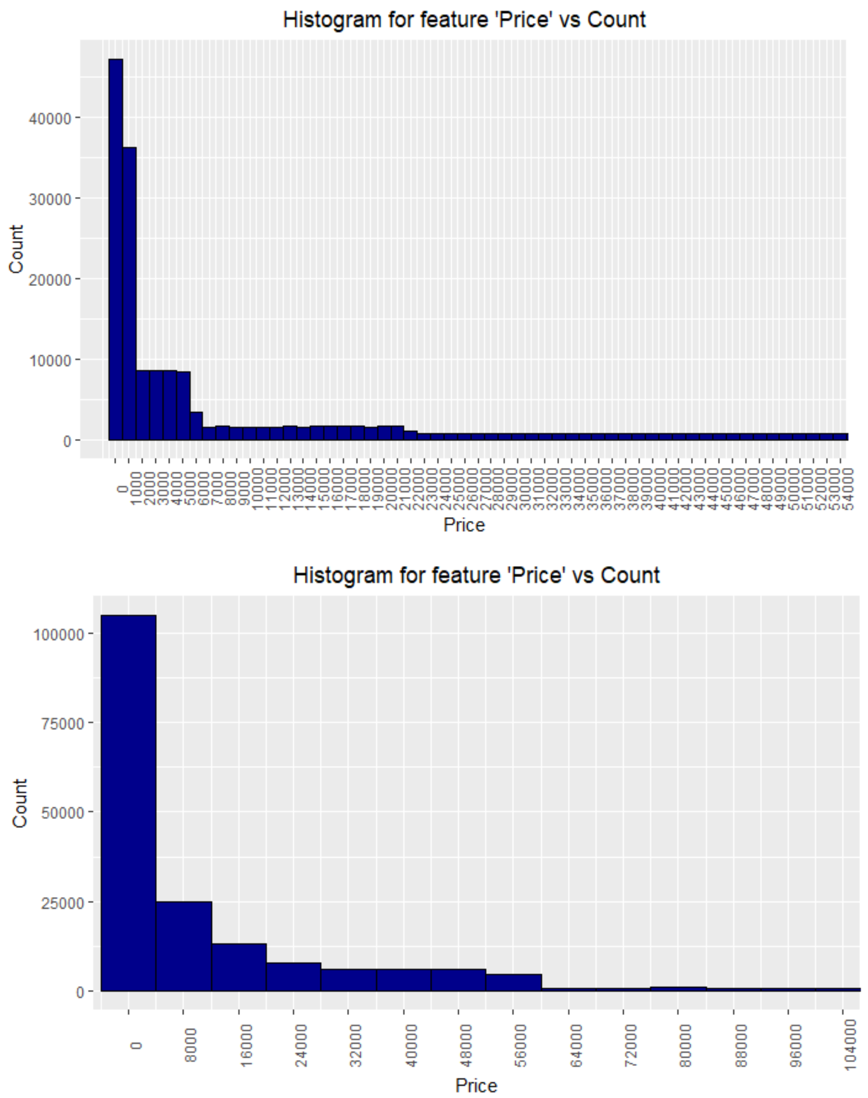


Figure 1 Histograms depicting the number of items purchased at a particular price, on two different scales

Figure 1-a and 1-b are histograms depicting the number of items bought for a particular price. These graphs are the same, simply represented on different scales.

These graphs show an exponential relationship between price amount and count (number of items bought).

This shows that more items are purchased for a lower price than a higher one. The number of items purchased increases exponentially as the price decreases.

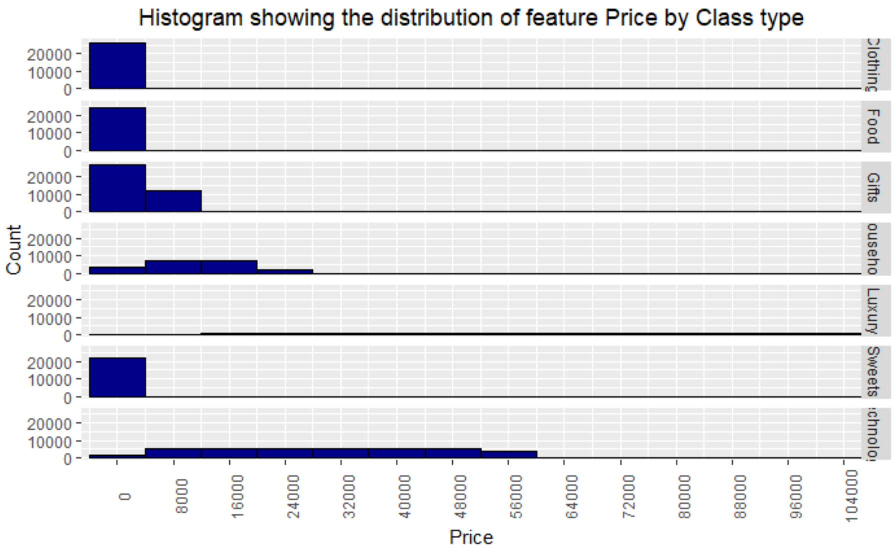


Figure 2 Histogram of feature Price vs Class

The figures here provide a more detailed depiction of price vs count , by including the Class Type (shown on the right of the tables).

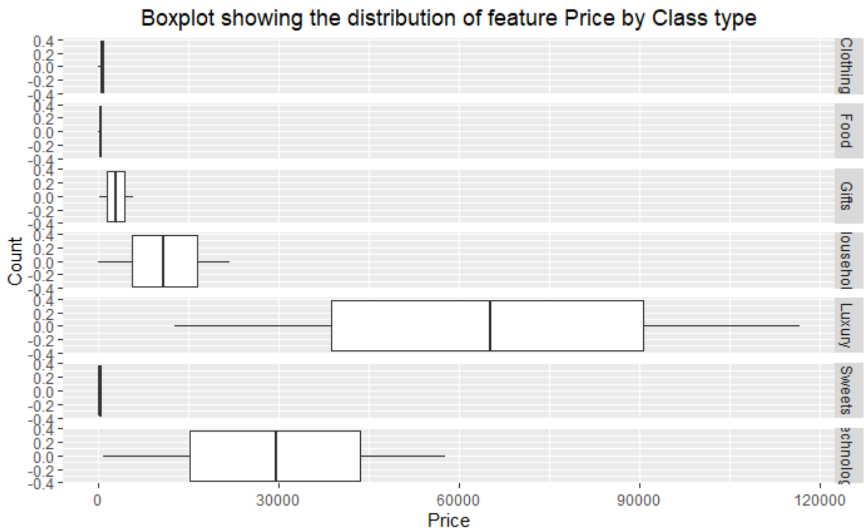


Figure 3 Boxplot of feature of Price vs Class

In Figure 3, the histogram shows how the number of items purchased for each particular price is distributed between the different classes. It can been seen that the class type Luxury is responsible for all purchases made for over 57 000. The number of items purchased for between 0-4000, items in Class Types Clothing, Food and Sweets were purchased in similar quantities.

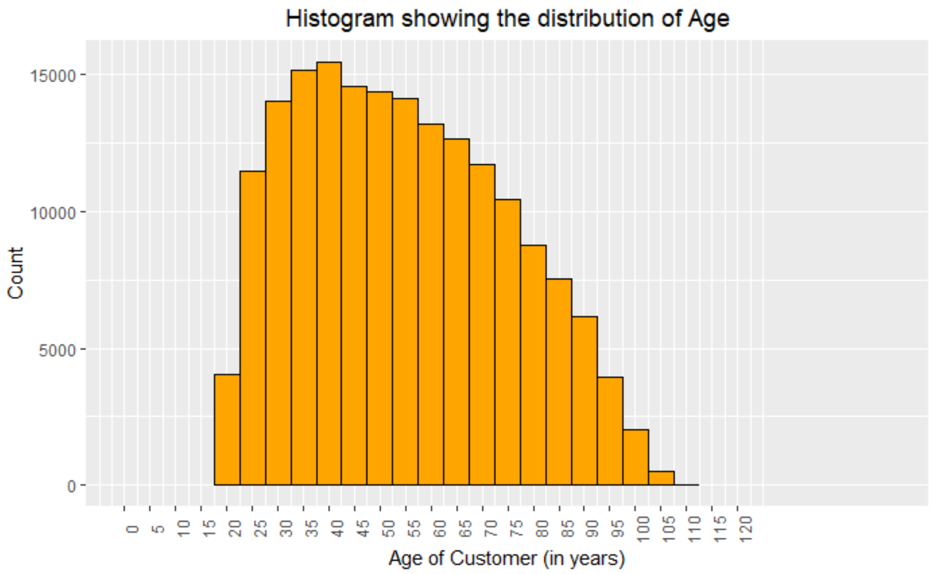


Figure 4 Histogram of distribution of Age

In figure 4, the data is skewed unimodally right, if only slightly. This graph shows that the majority of customers are between 30-55 years of age. This is expected, as this age category is the most economically active in society and generally possess relatively high buying power. No purchases have been recorded that are made by customers younger than around 17 years. This is also expected, as most children do not have access to their own money until they have reached young adulthood (18 years).

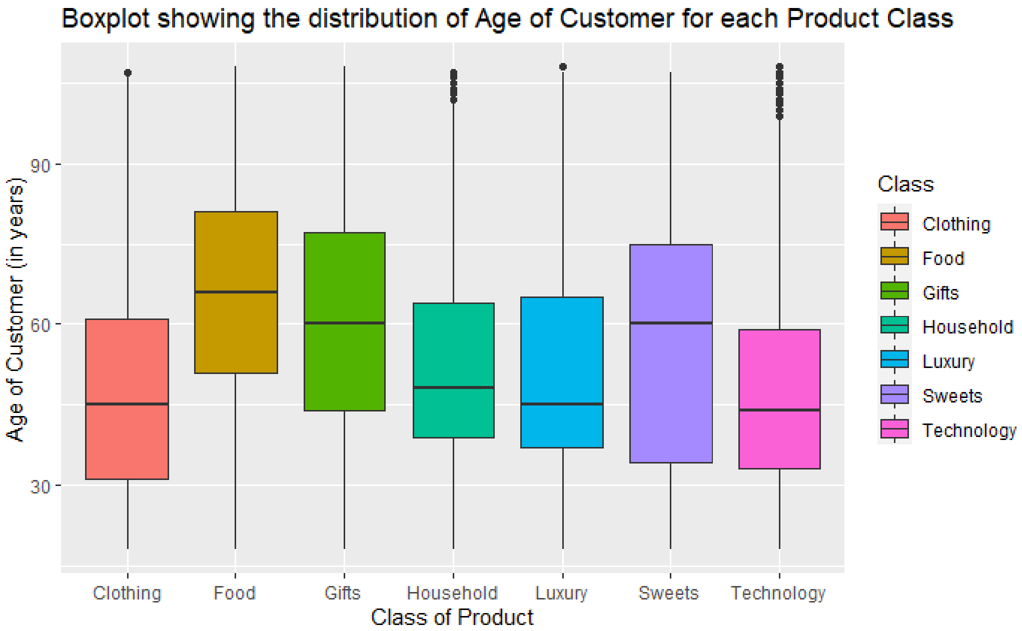


Figure 5 Boxplot for distribution of Age for each Product Class

In figure 5, boxplots representing the distribution of customer age for each product class is given. This graph shows that there is a tendency for the younger generation (around 50 years and below) to purchase items from product classes such Technology, Clothing, Luxury and Household. For product classes Food, Gifts and Sweets, the median is around 70 years. This indicates that the older generation are the predominant buyers of products in this class.

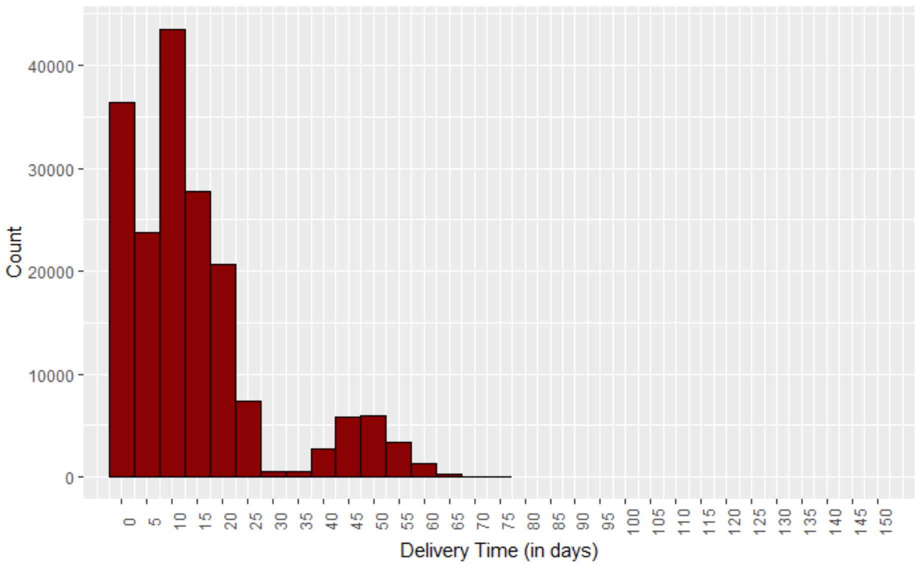
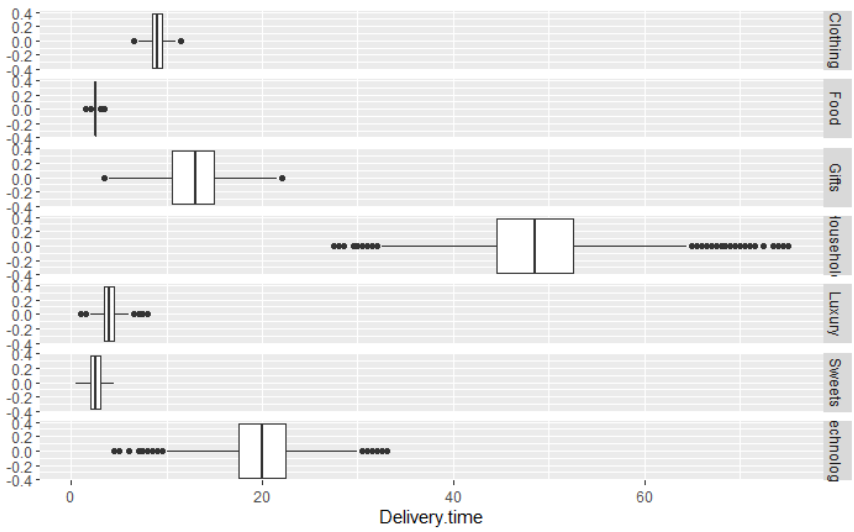


Figure 6 Histogram for Delivery Time vs Count

In Figure 7, the histogram of Delivery Time vs Count, is shown to be multimodal. There are peaks at around 10-15 and 45-50 days. This shows that customers either receive their purchase relatively soon (2 weeks) or else either an entire month later. The delivery times do not appear to be consistent. Investigation is required to determine the reasons for this.



J

Figure 7 Boxplot showing the distribution of delivery time for each Product Class

In Figure 8, the boxplot showing the distribution of delivery times for each product class type, shows that Household products have the longest delivery time as well as the largest variation in the delivery times. This is known because the median is close to 100 days (making it the longest) and the range is large (min is about 28 and the max is about 78). Product class types Food, Luxury and Sweets have the fastest delivery times, and the times are also quite consistent (little variation). This is good as food items can be perishable, meaning that fast and reliable delivery would be poignant.

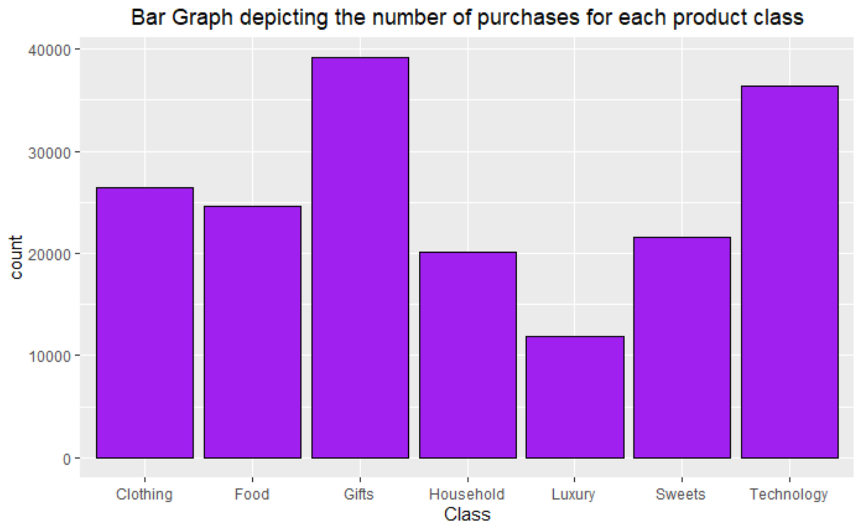


Figure 8 Bar Graph for number of purchases for each Product Class

Figure 9 depicts a simple histogram, which shows the number of items purchased from each product class. Gifts has the highest number of purchases, whilst Luxury has the lowest number of purchases.

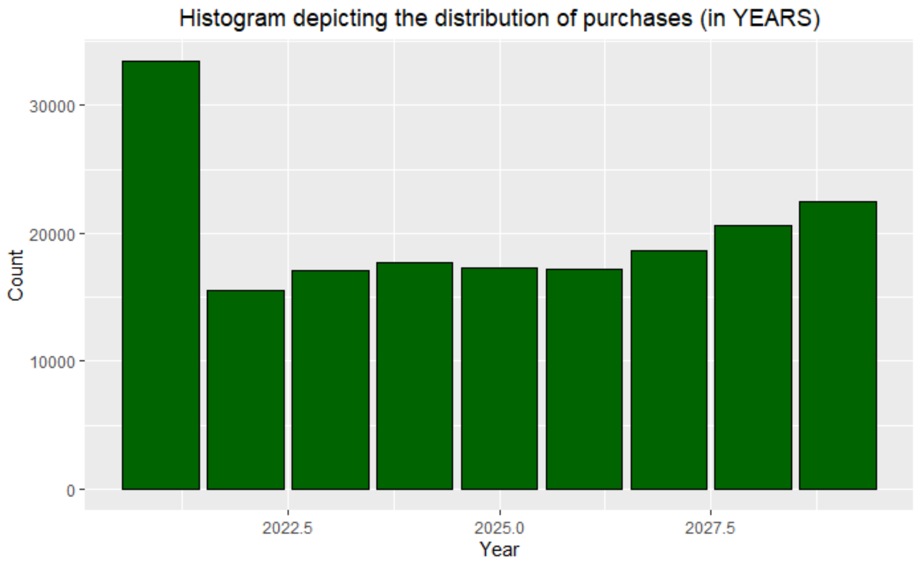


Figure 9 Histogram showing distribution of purchases in years

In Figure 9, it can be seen from the graph that the most purchases/sales were made in the Year 2021. There is a "spike" here.

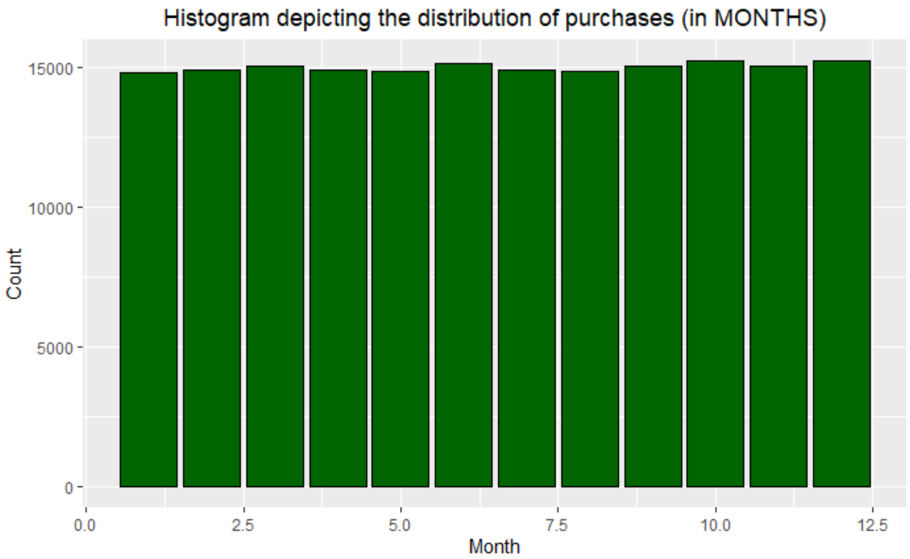


Figure 10 Histogram showing distribution of purchases in months

In Figure 10, the data is uniformly distributed, meaning that the sales/demand show no seasonality and therefore remains consistent.

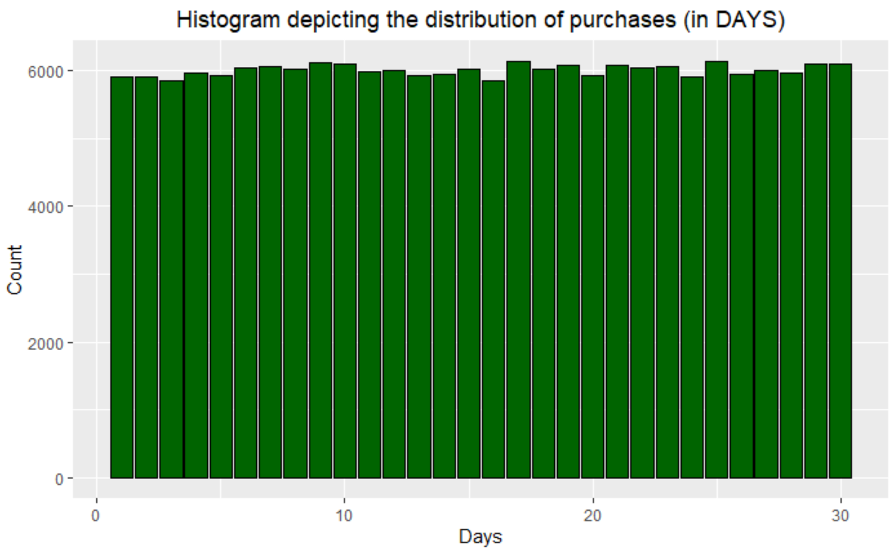


Figure 11 Histogram showing distribution of purchases in days

In Figure 11, the data is uniformly distributed, meaning that the sales/demand show no seasonality and therefore remains consistent.

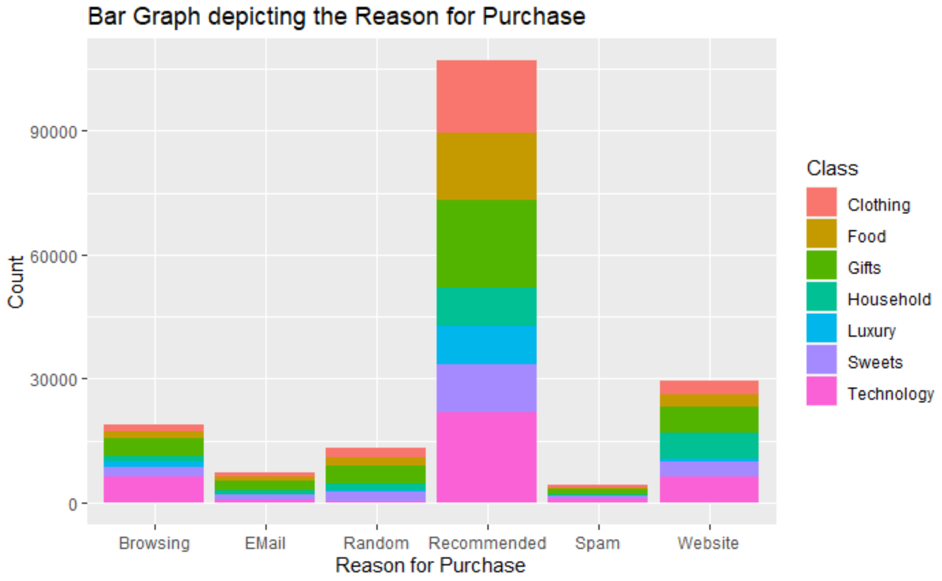


Figure 12 Histogram for Reason for Purchase, with breakdown for Product Class

Figure 12 is an interesting graph; it will provide valuable insight into how the company can increase sales. This histography shows how many products were purchased by customers as a result of exposure to different “advertising” platforms. In addition, the product class from which the items were purchased is included. In general, class ‘recommendations’ is responsible for the most purchases. It can also be seen that the most technology purchases were made due to recommendations from other customers, as well as that the most purchased recommended product class was Technology. However, although Technology was the most purchased item due to browsing, overall recommended is still the number 1 reason for technology purchases.

PROCESS CAPABILITY INDICES

Table 2 Process Capability Indices of Technology

Cp	Cpu	Cpl	Cpk
1.142207	0.3796933	1.90472	0.3796933

The Process Capabilities indices of the process delivery times for the class items ‘Technology’, shown in table 2, were calculated using a USL = 24 and LSL = 0. The USL (Upper Specification Limit) and LSL (Lower Specification Limit) are determined by the customer (VOC). An LSL of 0 makes sense because these calculations concern delivery time, and one cannot have a delivery time of less than 0 days.

Process Capability indices are a measure to determine the ability of a process to meet specifications. The Process Capability charts that are shown further on in the report, show the stability of a process, they do not however show whether the process is capable of meeting specifications as per the Voice of Customer (VOC), or whether the process is performing to its potential capability. This is why including the Process Capability Indices is important.

The Cpk value is 0.3796933. A Cpk value of 1 is considered barely stable, a value of 1.3 can be considered stable. Therefore, this extremely how Cpk value indicates that the process is not capable.

PART 3 STATISTICAL PROCESS CONTROL

3.1 INITIAL CONSTRUCTION OF S AND X-BAR CHARTS

Six-Sigma values for X-bar Chart:
In Table 3 below, the six sigma values for each respective product class for the S-chart are given.

Table 3 X-bar Chart

Class <chr>	UCL <dbl>	U2Sigma <dbl>	U1Sigma <dbl>	CL <dbl>	L1Sigma <dbl>	L2Sigma <dbl>	LCL <dbl>
Technology	22.888932	22.252544	21.313494	20.374444	19.435394	18.496345	17.859957
Clothing	9.390601	9.261041	9.115521	8.970000	8.824479	8.678959	8.549399
Household	47.464268	48.926693	47.744458	46.562222	45.379987	44.197751	45.660177
Luxury	5.468973	5.307688	5.021622	4.735556	4.449489	4.163423	4.002138
Food	228.912240	489.602253	357.551126	225.500000	93.448874	-38.602253	222.087760
Gifts	9.451412	9.124865	8.742988	8.361111	7.979234	7.597357	7.270811
Sweets	2.889680	3.545668	3.015363	2.485057	1.954752	1.424447	2.080434

Six-Sigma values for the S Chart
In Table 4 below, the six sigma values for each respective product class for the X-Bar Chart are given.

Table 4 S-chart

Class <chr>	UCL <dbl>	U2Sigma <dbl>	U1Sigma <dbl>	CL <dbl>	L1Sigma <dbl>	L2Sigma <dbl>	LCL <dbl>
Technology	5.1146591	4.7120218	4.0037748	0.70824699	2.5872808	1.8790338	1.4763965
Clothing	0.8555346	4.7120218	0.6373038	0.08605728	0.4651893	0.3791320	0.2469585
Household	1.8348295	3.1641500	2.1731927	0.99095727	0.1912782	-0.7996791	0.5296415
Luxury	1.4918272	3.1641500	1.1426962	0.18146738	0.7797615	0.5982941	0.4306305
Food	6.9407550	4.4721360	4.4721360	0.00000000	4.4721360	4.4721360	2.0035169
Gifts	2.2177540	1.9636522	1.6963087	0.26734351	1.1616217	0.8942782	0.6401764
Sweets	0.8230339	0.7521517	0.6412285	0.11092318	0.4193822	0.3084590	0.2375768

As briefly touched upon previously, the purpose of Statistical Process Control (SPC) charts is to determine whether a particular process is in control or out of control.

For the first 30 samples, we will construct S and X-bar charts to determine whether the particular process is in or out of control. For variable data, such as time (what is considered here-Delivery Time), two charts need to be constructed. First, the S-chart is constructed. The S-chart measures the standard deviation of each sample (of 15 instances). If the samples are in an acceptable range of +-3 sigma from the Centre Line (CL) then the data is considered accurate to construct the X-bar Chart.

The X-Bar Chart measures the mean of the sample. If the sample does not confine to the control limits (UCL, CL, LCL), then these samples must be removed before constructing the X-bar chart. The X-bar chart does not mean anything, unless the standard deviation is known. It is the X-bar chart that we use to determine if a sample is in or out of control.

Both charts are constructed for 30 samples (of 15 instances) of each product class type (Gifts, Clothing etc).

GIFTS

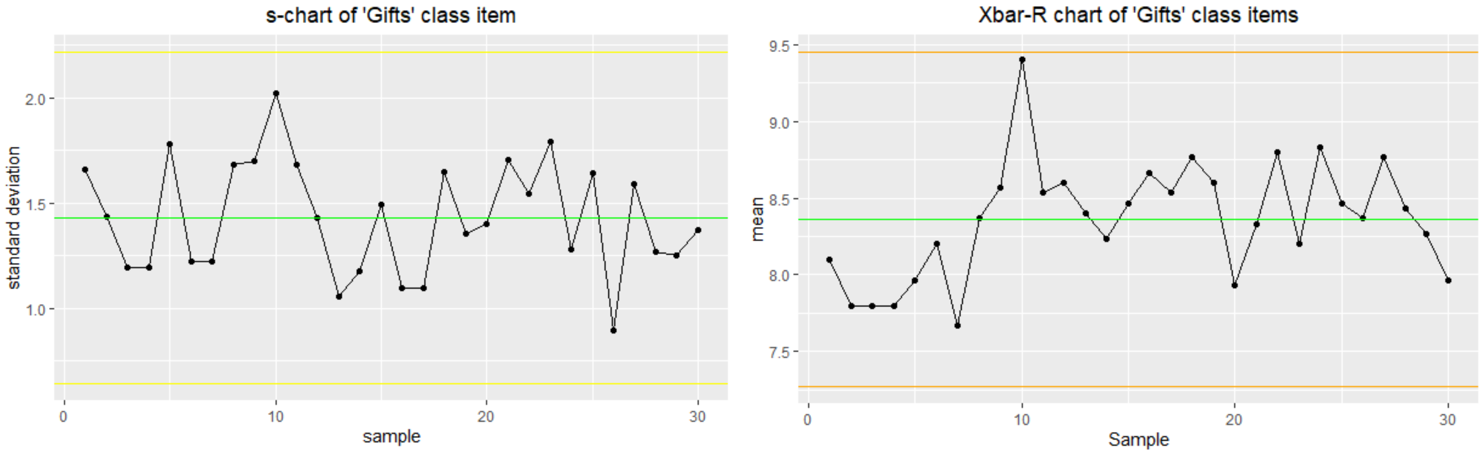


Figure 13 s-chart and x-bar chart for Gifts

In figure 13, in the S-chart it can be seen that none of the sample points are beyond the 3-sigma limit. Therefore, the X-bar chart can be constructed because it is now known that the X-bar chart will have accurate control limits. In the X-bar chart, it can be seen that no sample instances are beyond the 3-sigma limits, therefore the process is considered to be in statistical control.

CLOTHING

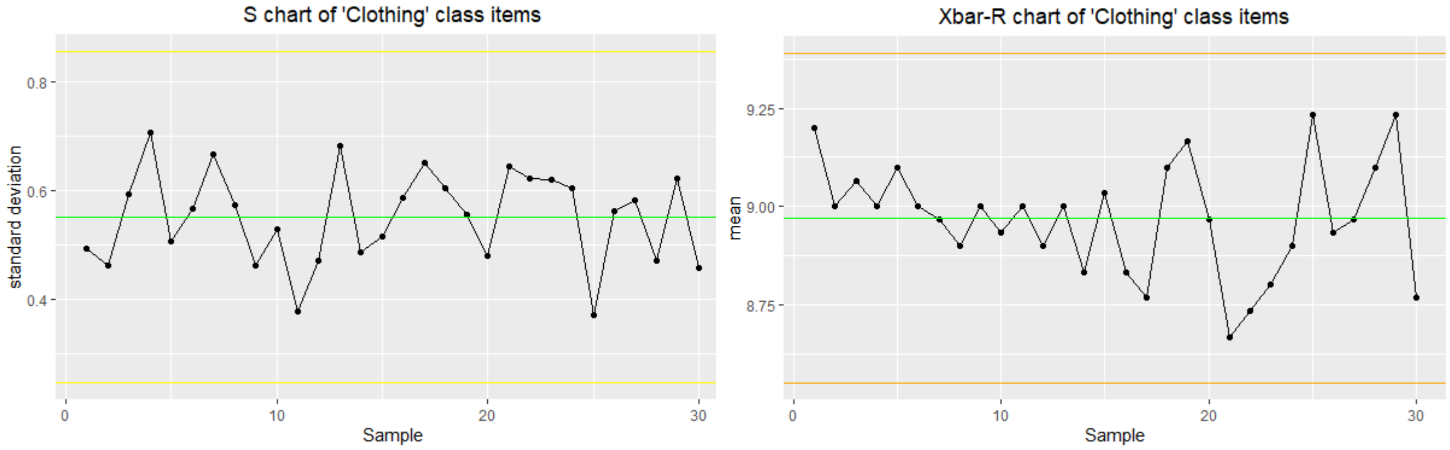


Figure 14 s-chart and x-bar chart for Clothing

In figure 14, in the S-chart it can be seen that none of the sample points are beyond the 3-sigma limits. Therefore, the X-bar chart can be constructed because it is now known that the X-bar chart will have accurate control limits. In the X-bar chart, it can be seen that no sample instances are beyond the 3-sigma limits, therefore the process is considered to be in statistical control.

HOUSEHOLD

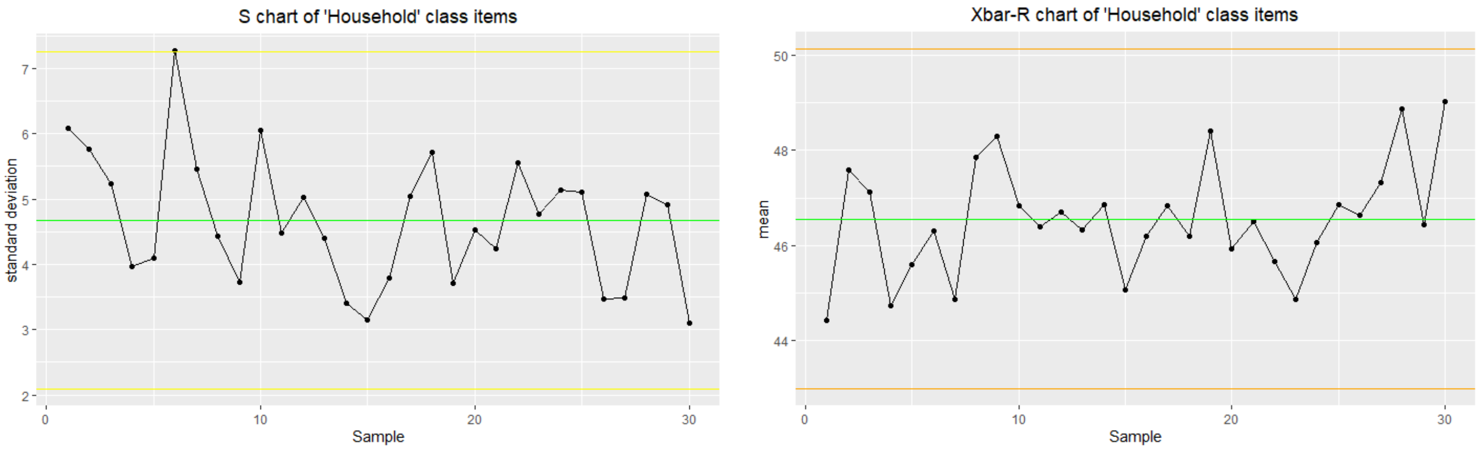


Figure 15 s-chart and x-bar chart for Household

In figure 15, in the S-chart it can be seen that there is one sample instance on the 3-sigma upper control limit. This is still acceptable, and therefore the data can be used to construct an X-bar chart with accurate control limits. It can be seen on the X-bar chart that all sample instances are within the specified control limits, and thus once again this process can be considered in control.

SWEETS

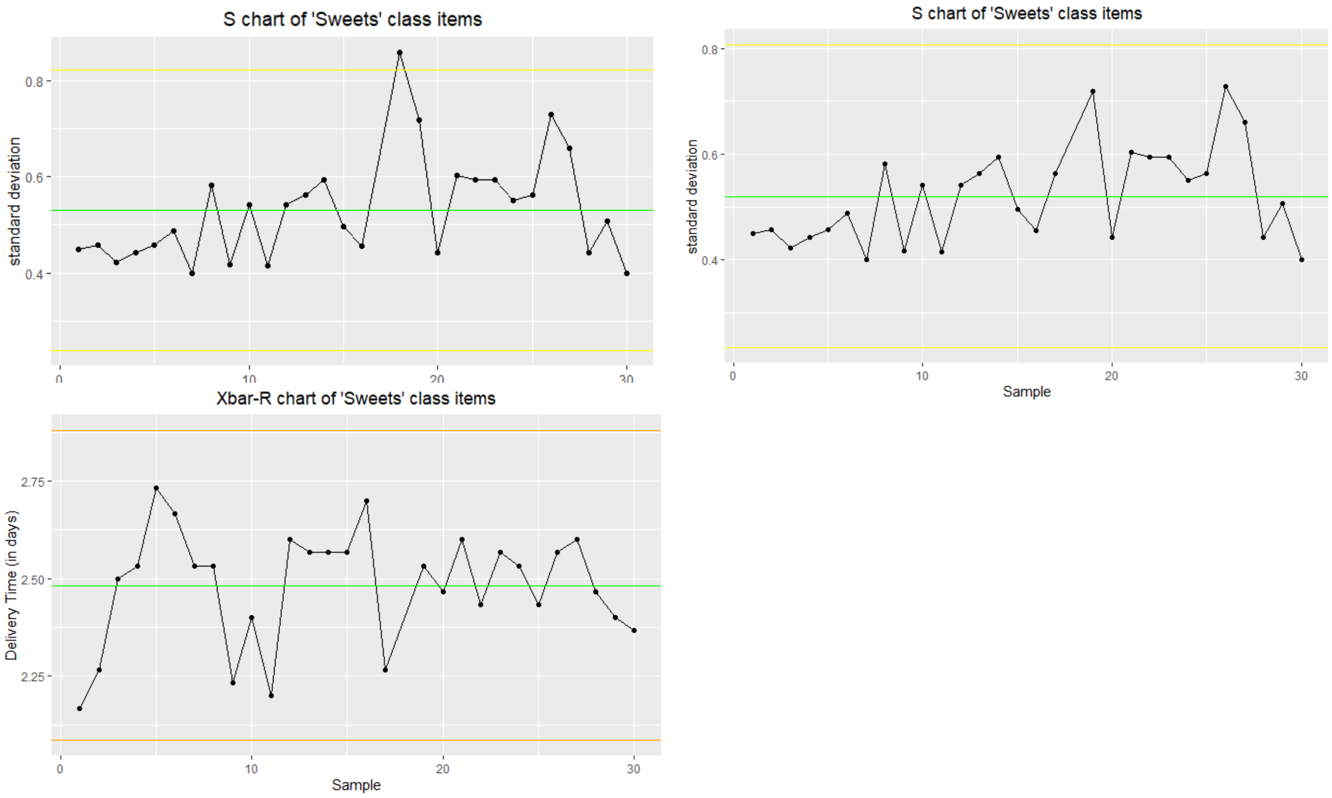


Figure 16 s-charts and x-bar chart for Sweets

In the Figure 16, in the S-chart it can be seen that there is one sample instance that is above the +3-sigma upper control limit. Therefore, this sample instance 18 must be removed before the X-bar chart can be constructed. Once the instance is removed, the Cl, UCL and LCL for ‘Sweets’ are re-calculated and the X-bar chart is constructed using these new values. From the X-bar chart it is seen that all instances are within the specified control limits. Therefore, the process is in control.

TECHNOLOGY

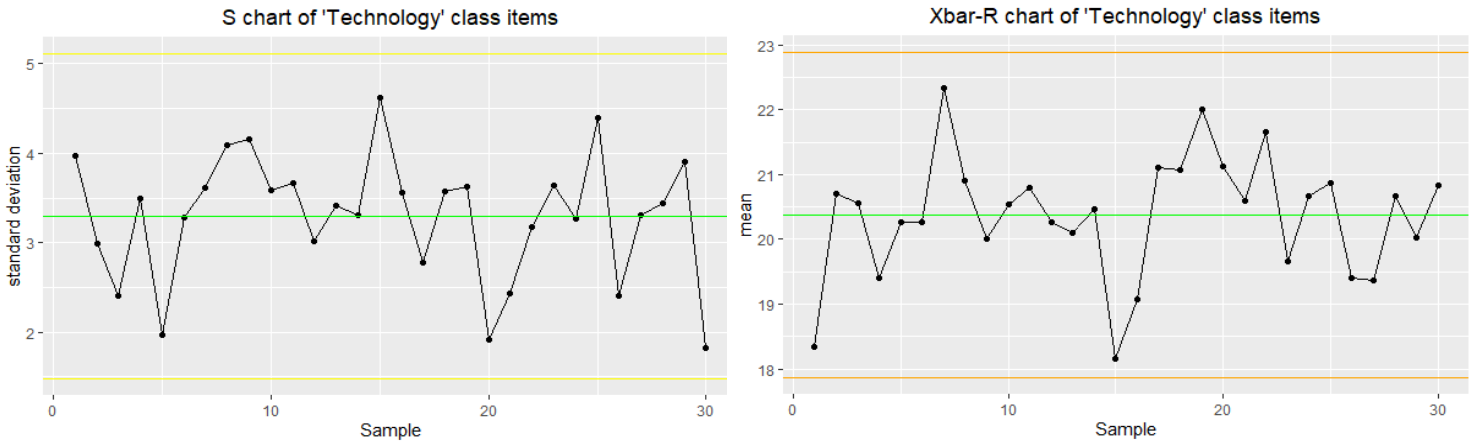


Figure 17 s-chart and x-bar chart for Technology

In Figure 17, in the S-chart it can be seen that none of the sample points are beyond the 3-sigma limit. Therefore, the X-bar chart can be constructed because it is now known that the X-bar chart will have accurate control limits. In the X-bar chart, it can be seen that no sample instances are beyond the 3-sigma limits, therefore the process is considered to be in statistical control.

LUXURY

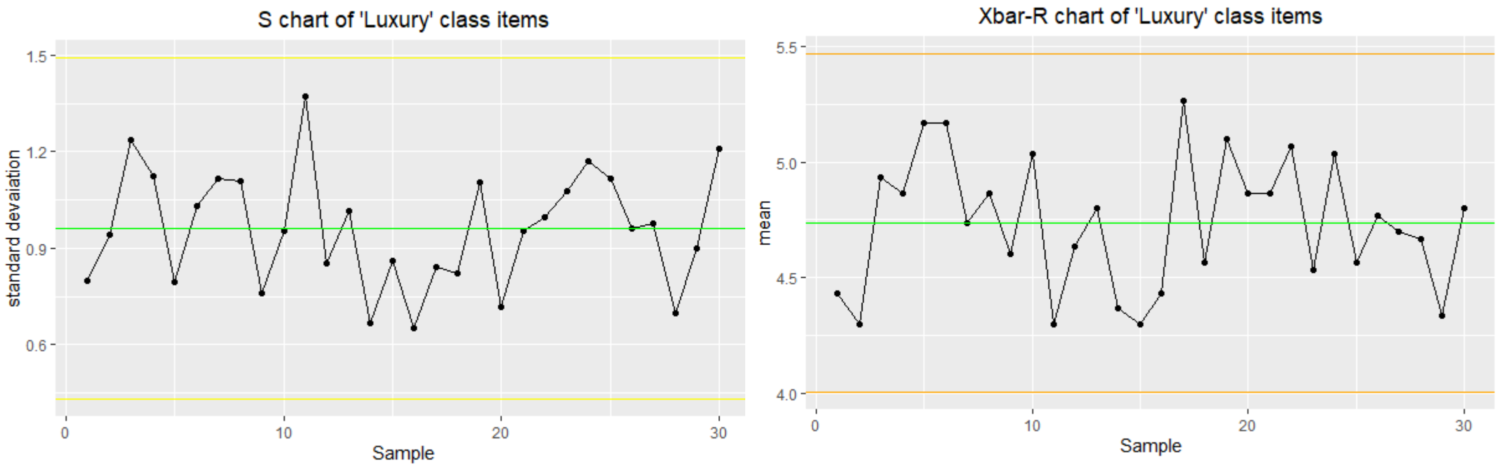


Figure 18 s-chart and x-bar chart for Gifts

In the Figure 18, in the s-chart it can be seen that none of the sample points are beyond the 3-sigma limit. Therefore, the X-bar chart can be constructed because it is now known that the X-bar chart will have accurate control limits. In the X-bar chart, it can be seen that no sample instances are beyond the 3-sigma limits, therefore the process is considered to be in statistical control.

FOOD

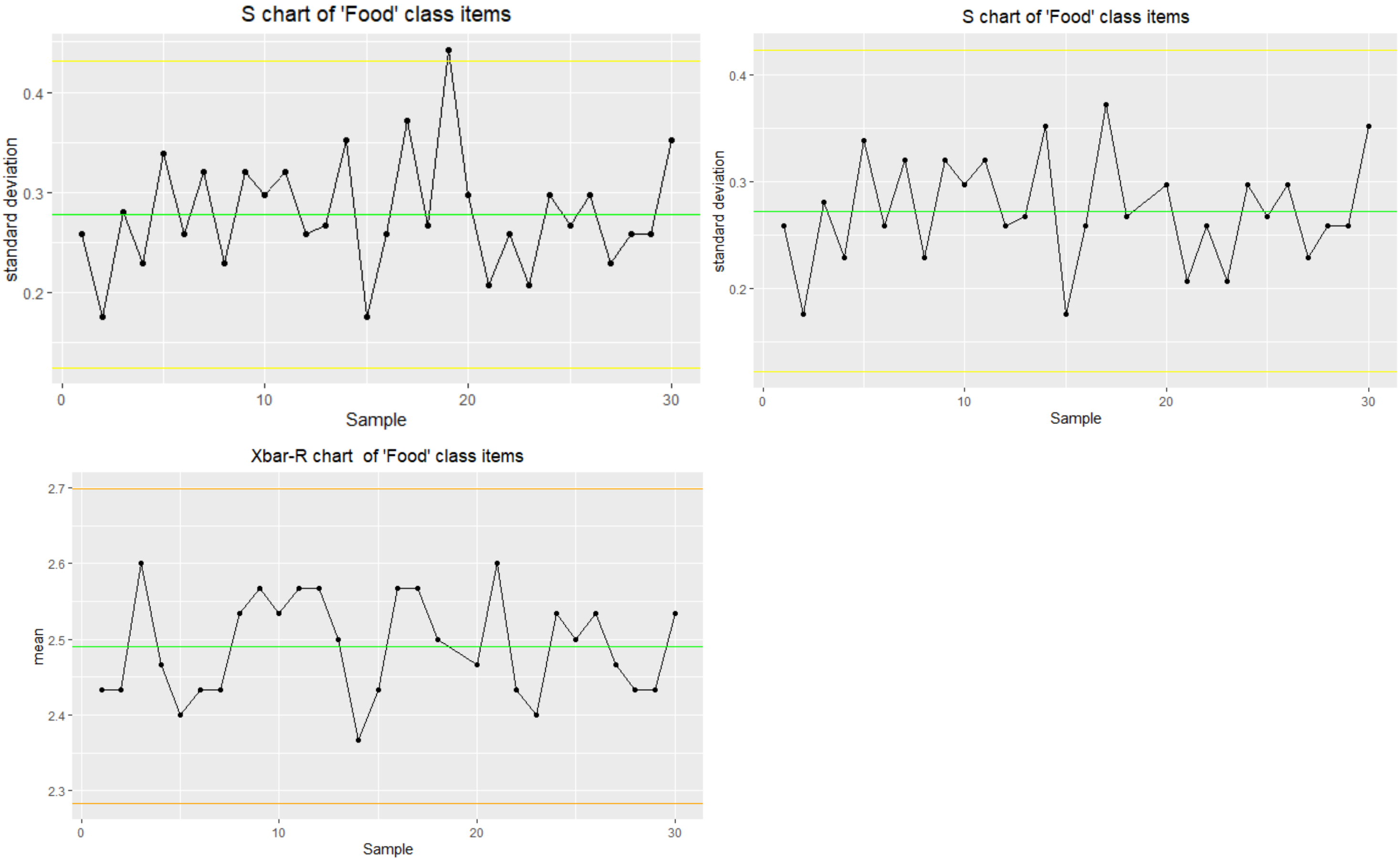


Figure 19 s-charts and x-bar chart for Food

In the figure 19, in the S-chart it can be seen that there is one sample instance that is above the +3-sigma upper control limit. Therefore, this sample instance 19 must be removed before the X-bar chart can be constructed. Once the instance is removed, the Cl, UCL and LCL for 'Sweets' are re-calculated and the X-bar chart is constructed using these new values. From the X-bar chart it is seen that all instances are within the control limits. Therefore, the process is in control and stable.

3.2 CONTINUED CONSTRUCTION OF X-BAR CHARTS

For this section, x-bar charts for number of samples greater than 30 will be constructed. Samples in multiples of 15 data instances will be created, until the data has been exhausted (highest multiple of 15 in that specific class). Data instances that do not form part of a sample of 15, will be disregarded. Again, the x-bar charts are used to determine whether the process is in or out of control, by evaluating whether the graphs confine to the control limits.

GIFTS

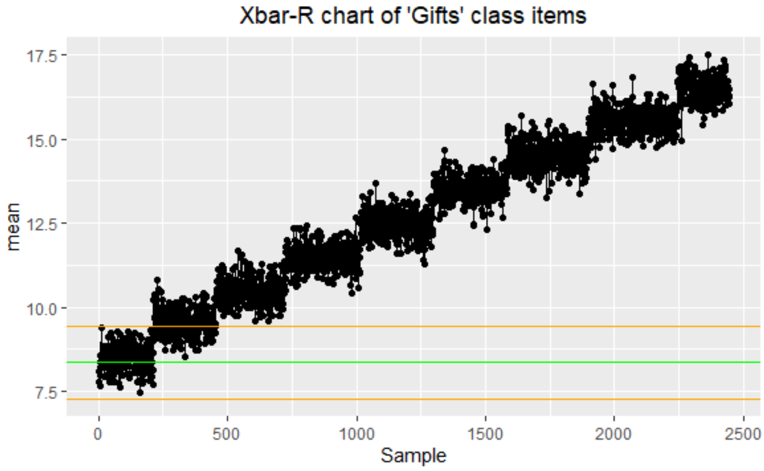


Figure 20 X-bar chart for Gifts

In figure 20, it can be seen that there that most of the graph is above the UCL (upper control limit). This indicates that this process is drastically out of statistical control. Upward trends are not a good thing, they show that the process is “worsening”. The delivery times for this product class much slower than before, as shown by the fact that the mean of the delivery times is periodically increasing drastically. Action needs to be taken to determine the reason for this, and then subsequently rectify as soon as possible.

CLOTHING

Figure 21 X-bar chart for Clothing

In figure 21, it can be seen that the process is out of statistical control. There are 18 outliers. There are 15 samples above the UCL and 3 samples below the LCL limit. However, in comparison to the number of sample considered, this is relatively few outliers. The majority of data points are within the specified control limit, and are evenly distributed above and below the centre line, indicating the presence of normal distribution. Overall, technically this process is out of control, but is far less of a concern than the previous class of ‘gifts’, for instance. The process is relatively stable.

HOUSEHOLD

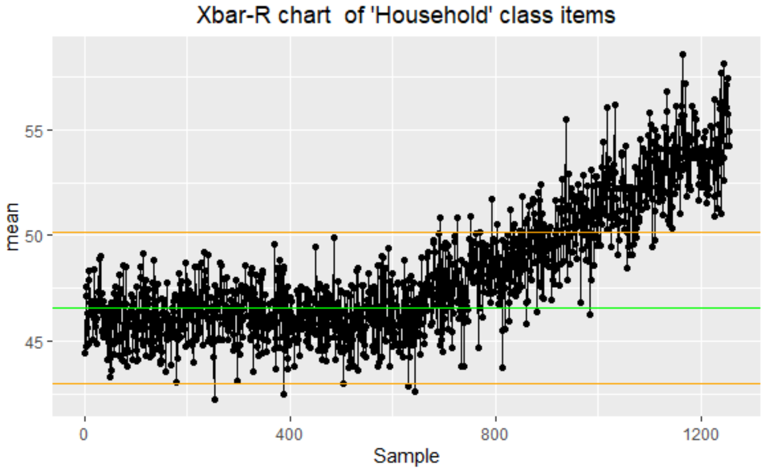


Figure 22 X-bar chart for Clothing

In figure 22, it can be seen that the process is out of statistical control. The process is unstable. There is an increasing trend in the graph, showing that the graph begins to shift upward and deviate from the mean (green centre line) at around the half way mark in the sample instances. This indicates that the mean of the samples has shifted, and so the mean should be adjusted upwards to account for this. There are many potential reasons for this shift-one could be that the delivery method for this product class was changed, resulting in sample mean that deviates largely from the original. Upward trends are not a good thing, they show that the process is “worsening”. The delivery times for this product class are slower than before, as shown by the fact that the mean of the delivery times has increased.

LUXURY

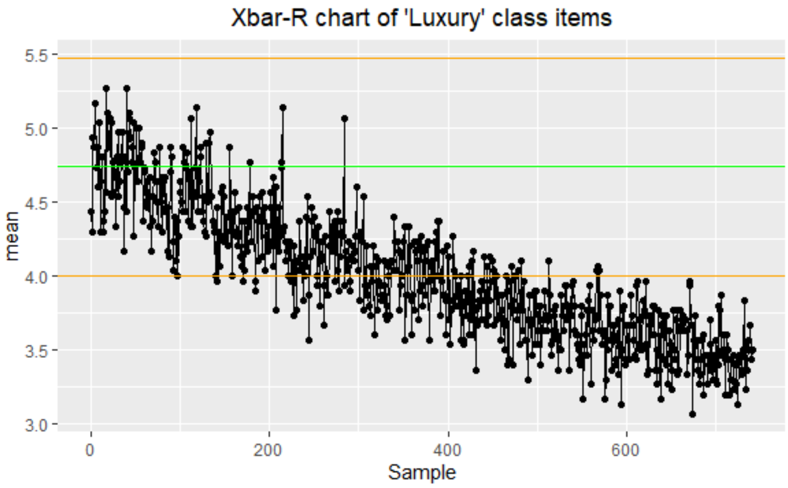


Figure 23 X-bar chart for Luxury

In figure 23, it can be seen that the graph is out of statistical control and therefore the process is not stable. The graph exhibits a downward trend, showing that as the samples increase the data strays further and further away from the centre line, with roughly half the samples being below the LCL of negative 3 sigma. This shows that the mean of the samples has shifted over time, and so the mean should be adjusted downward accordingly to compensate for this. Again, there are a variety of potential reasons for this. Perhaps, because luxuries are an expensive commodity, customers expect a quicker delivery time and so the delivery method changed to accommodate this. A downward trend shows that the process is "improving" in a sense, the delivery time for this product class is faster than it was previously (the new mean is lower than the original).

SWEETS

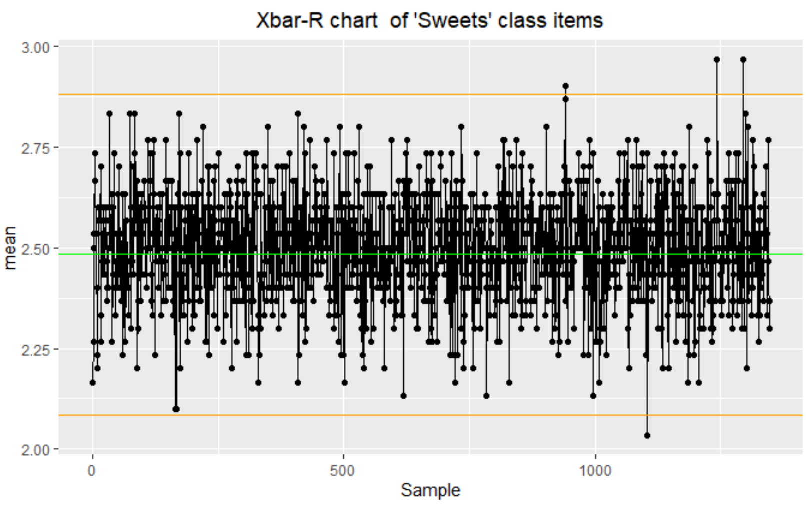


Figure 24 X-bar chart for Sweets

As seen in Figure 24, this process is in statistical control. This can be seen by the fact there are 4 outliers, 3 above the UCL and 1 below the LCL. The outliers are few, practically negligible. The majority of points are within the control limits and evenly distributed above and below the centre line, indicating a normal distribution. Although technically out of control, this process can be considered to be relatively stable.

TECHNOLOGY

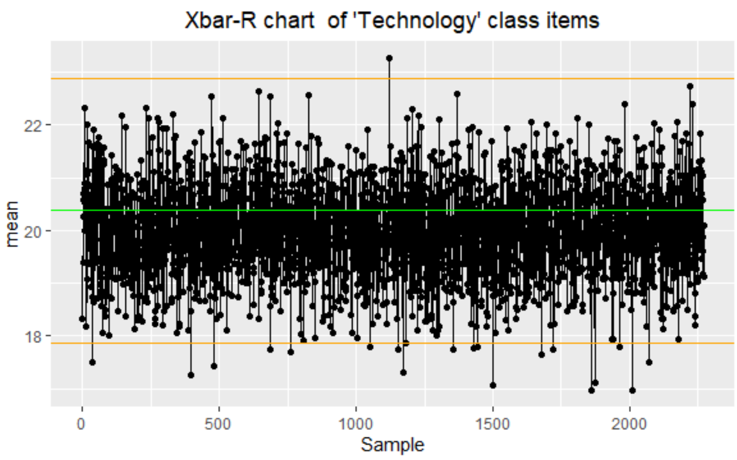


Figure 25 X-bar chart for Technology

As seen in Figure 25, this process is out of statistical control. This can be seen by the fact there are 20 outliers, 1 above the UCL and 19 below the LCL. However, the majority of points are still within the control limits and evenly distributed above and below the centre line, indicating a normal distribution. The process can be considered to be relatively stable.

FOOD

Figure 26 X-bar chart for Food

As seen in Figure 26, this process is in statistical control. This can be seen by the fact there are 5 outliers, 1 above the UCL and 4 below the LCL. The majority of points are within the control limits and evenly distributed above and below the centre line, indicating a normal distribution. This process can be considered to be relatively stable.

PART 4 OPTIMISING THE DELIVERY PROCESSES

4.1 OUT OF CONTROL PROCESSES

RULE A: X-bar/sample means outside of the outer control limits

Only product class types with outliers>6 will be considered.

Table 5 Rule A

		Sample Index					
Class	**Nr. Outliers**	**1st**	**2nd**	**3rd**	**3rd Last**	**2nd Last**	**Last**
Gifts	2133	213	216	218	2444	2445	2446
Clothing	19	148	217	455	1626	1640	1644
Household	328	252	387	629	1252	1253	1254
Luxury	405	98	140	142	739	740	741
Sweets	4						
Technology	20	37	398	483	1961	2009	2071

Food	5						

As seen in the table 5 above, the 2 product classes with the highest number of outlier, in descending order, are:

1. Gifts
2. Luxury
3. Household

This is in line with expectations, as in Part 3 the graphs of these product classes showed large deviation from the mean/centre line. ‘Gifts’ and ‘Household’ showed an upward trend. ‘Luxury’ showed a downward trend.

RULE B: Find most consecutive samples between -0.3 and +0.4 sigma

Here the new control limits are added onto the original s-charts for the different product class types. In essence, the new UCL = 0.4 and the new LCL = -0.3. Now, the longest consecutive string of samples with the same value found within these control limits is determined.

Table 6 Rule B

Class	**Longest Pattern**
Gifts	7
Clothing	4
Household	3
Luxury	4
Sweets	5
Technology	6
Food	4

Therefore, it can be seen in table 6, the product class type with the largest consecutive sample string is ‘Gifts’.

4.2 TYPE I MANUFACTURERS ERROR

A type I error is the probability that the results indicate something is wrong, when in reality it is are correct. The error incurs a cost to the manufacturer, thus it is referred to as the Manufacturers error.

H0: the mean of the process is within the upper and lower control limits

HA: the mean of the process is not within the upper and lower control limits.

Table 7 Type I Error

	Rule A	**Rule B**
Type 1 Error Probability	0.002699796	0.7266668

Therefore from table 7, it can be seen that the probability of a Type I error for Rules A is 0,27%, and the probability of a Type I error for Rules B is 72,67%

4.3 DELIVERY PROCESS BEST PROFIT

Currently:

Mean hour for delivering technology products: 20.01 hours

Total sales beyond 26 hours: 1356 sales

If every lost sale costs the company R329, the total loss will be R446 124.

It costs R2.5 per hour over 26 hours to move the mean of the distribution considered to the left.

Therefore total cost is R636 125.

Therefore, calculating the minimum delivery cost is vital, because this will decrease the loss of revenue due to lost sales. This minimum delivery cost is calculated by looping through the potential delivery times and by storing the costs related to these hours.

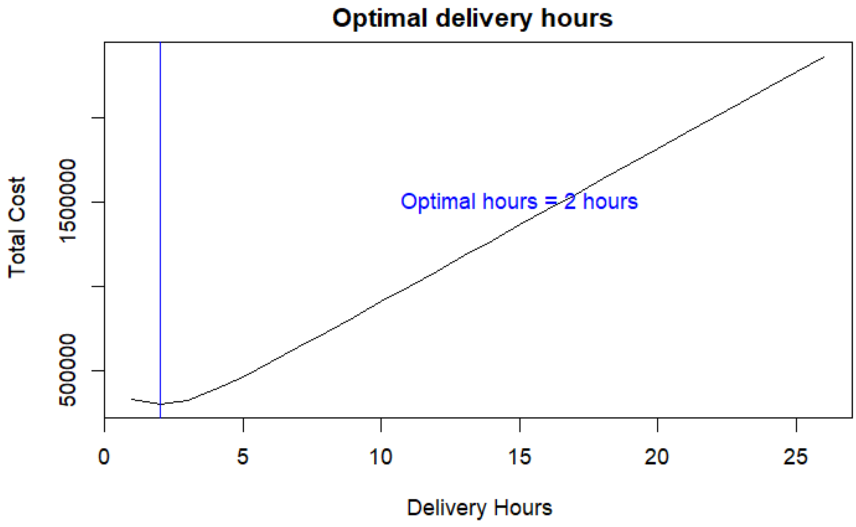


Figure 27 Graph for Optimal Delivery Time

In figure 27, it can be seen from the graph that the average of the distribution should be shifted 2 hours to the left, in order to achieve the minimum delivery time.

4.4 TYPE II CONSUMERS ERROR

A type I error is the probability that the results indicate something is correct, when in reality it is are wrong. The error incurs a cost to the consumer, thus it is referred to as the Consumers error.

H0: the mean of the process is within the upper and lower control limits

HA: the mean of the process is not within the upper and lower control limits.

Table 8 Type II Error

	Technology
Probability of Type II error	0.48761

Therefore, as seen is table 8, the probability of Type II error for Technology is 48.76%

PART 5 MANOVA

HYPOTHESIS TEST 1

For a MANOVA test to be conducted, a dependent and independent variables must be chosen.

Independent – The different classes of each product

Dependent – Price and Age

Assume a significant value of 0.001

Table 9 Hypothesis Test 1

Dependent	Price and Age
Independent	Class of each product
H0	Price and Age made no significant change to the class of product chosen to buy.
H1	At least one feature has an influence on the buying pattern.

In table 9, the summary of the variables for Hypothesis 1 is shown.

MANOVA test is used to test whether the dependent has influence on the independent. The p-value is less than 2.2e-16, therefore the null hypothesis is rejected.

Table 10 Results of Hypothesis 1

Dependent	**P-value**	**Analyses**
Price	2.2e-16	P value is smaller than 0.001. This means that the price differs depending on what class the product is.

Age	2.2e-16	P value is smaller than 0.001. This means that the age of a person differs depending on what class the product is.

In table 10, the results of the Hypothesis Test 1 are summarized

VISUALIZATION

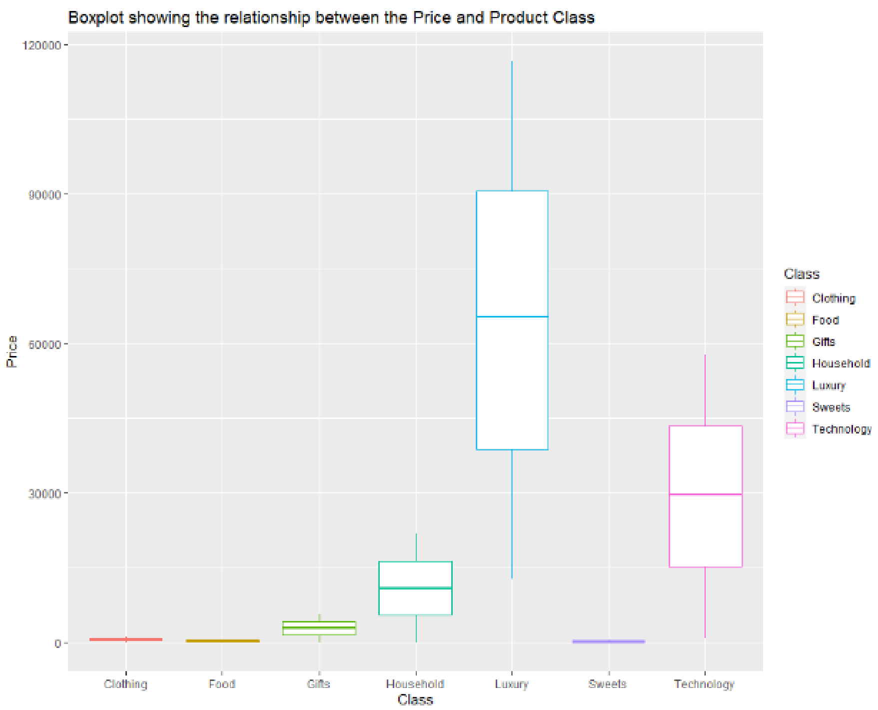


Figure 28 Boxplot showing Distribution of Price per Product Class

In Figure 28, it is seen that the MANOVA was able to identify a link between the distribution of Price and Product Class.

Items such as clothing, food and gifts have a much lower average mean than classes such as technology and Luxury. Class Luxury has the highest mean.

This means that on average, customers spend the most money per purchase for Luxury goods, and the least money per purchase on product classes such as clothing, food, sweets and gifts.

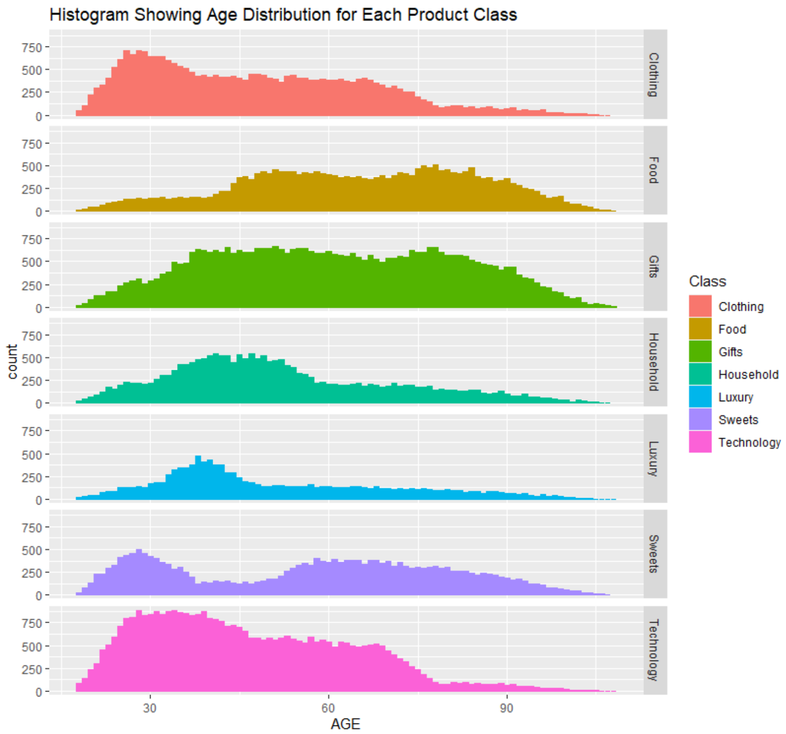


Figure 29 Histogram showing Distribution of Age for each Product Class

In Figure 29, it is seen that the MANOVA was able to identify a link between the distribution of Age and Product class

Here it can be seen the older customers (60+ years) purchased mainly from class types Food and Gifts, whilst the younger customers (30-50years) purchased mainly from classes Clothing and Technology.

This is known from examination of where each respective graph for the different product classes peaked.

```
              Df  Pillai approx F num Df den Df    Pr(>F)
Class          6 0.81686    20710     12 359952 < 2.2e-16 ***
Residuals 179976
---
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 5.7165e+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 AGE :
                Df   Sum Sq Mean Sq F value    Pr(>F)
Class            6  8423110 1403852  3805.2 < 2.2e-16 ***
Residuals   179976 66397874     369
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 30 Summary of Hypothesis Test 1

Figure 30, simply shows a summary of the results obtained from the Hypothesis Test 1.

HYPOTHESIS TEST 2

A MANOVA was conducted, a dependent and independent variable must be chosen.

Independent - Class of each product type

Dependent - Delivery Time and Year

Assume a significance value of 0.001

Table 11 Hypothesis Test 1

Dependent	Delivery time and Year
Independent	Class of each product
H0	Delivery time and year made no significant change on the class of product chosen to buy.
H1	At least one feature has an influence on the buying pattern.

In table 11, the summary of variables needed for the calculations of Hypothesis Test 2.

MANOVA test is used to test whether the dependent has influence independent. The p-value is less than 2.2e-16, therefore the null hypothesis is rejected.

Table 12 Results of Hypothesis Test 2

Dependent	**p-value**	**Analyses**
Delivery time	2.2e-16	P value is smaller than 0.001. This means that the delivery time differs depending on what class the product is.
Year	2.2e-16	P value is smaller than 0.001. This means that the Year of a product being bought differs depending on what class the product is.

In table 12, the results of the Hypothesis Test 2 were summarized.

VISUALIZATION

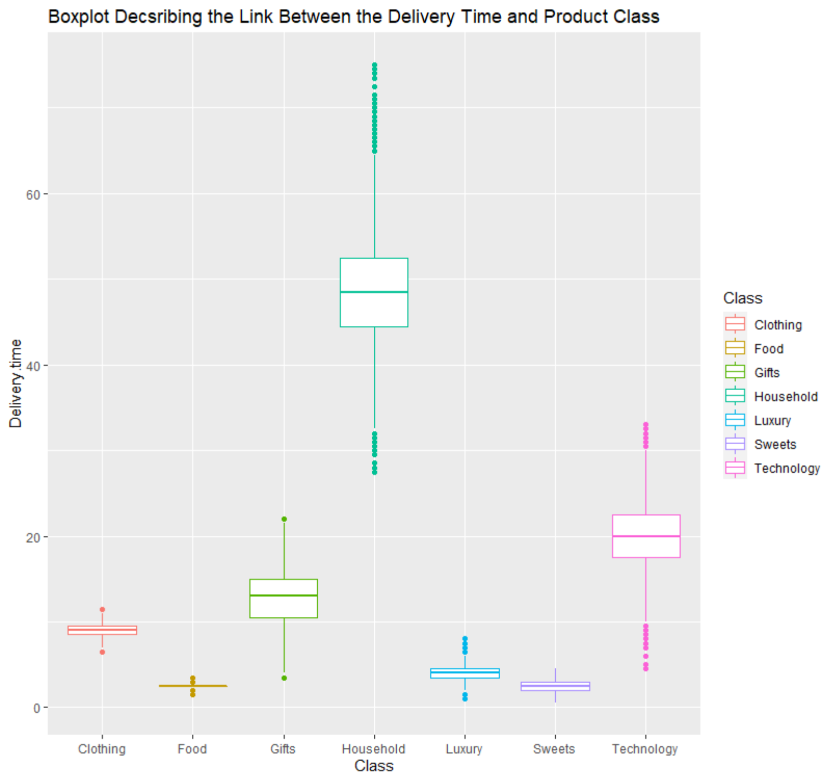


Figure 31 Boxplot for Product class vs Delivery time

In Figure 31, it is seen that the MANOVA was able to identify a link between Product Class and Delivery time

Product class Household has the highest mean of delivery time, followed by Technology and Gifts. This means purchases of product class Household, on average, took the longest to be delivered to customers.

Conversely, purchases of product classes Sweets, Luxury had the shortest delivery times.

Figure 32 Histogram of Annual Trends

In Figure 32, the MANOVA identified a link between Product Class and Year.

AS can be seen, Product classes clothing and Household experienced “a spike” in purchases for the year 2021.

Otherwise, the buying trends are consistent from 2022-2029

```
              Df Pillai approx F num Df den Df    Pr(>F)
Class          6 1.0414    32593     12 359986 < 2.2e-16 ***
Residuals 179993
---
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 33462168 5577028  629555 < 2.2e-16 ***
Residuals   179993  1594500       9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Year :
                Df  Sum Sq Mean Sq F value    Pr(>F)
Class            6  153083 25513.8  3699.4 < 2.2e-16 ***
Residuals   179993 1241362     6.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 33 Summary of Hypothesis 2

Figure 33, shows the results obtained from Hypothesis Test 2.

PART 6 RELIABILITY OF SERVICES AND PRODUCTS

6.1. PROBLEMS FROM TEXTBOOK

PROBLEM 6

1. Calculate constants

$$L(x)=k(x-T)^2$$
$$30=k(0.035)^2$$
$$k=30/0.035^2 = 24489.796$$

2. Calculate loss function

$$L(x) = k(x-T)^2$$
$$L(x) = 24489.796(x - 0.04)^2$$

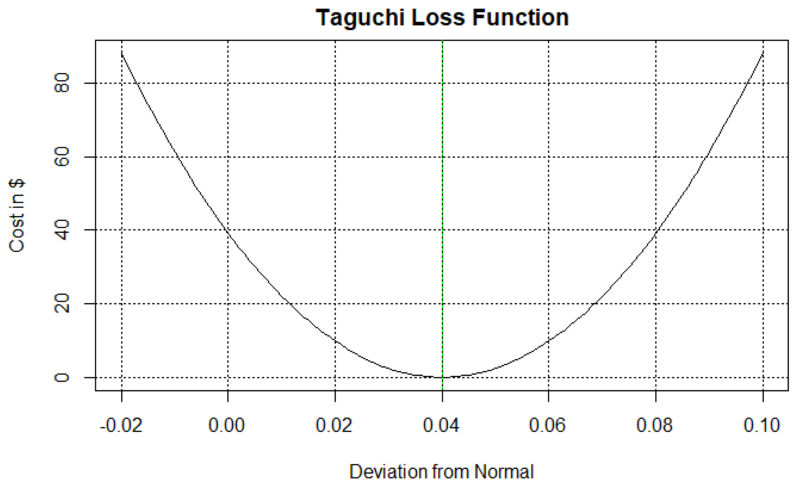


Figure 34 Taguchi loss function of Problem 6

In figure 34, it can be seen that as the thickness of the product part deviates from standard specification, the loss value increases.

The loss value increases due to customer dissatisfaction

PROBLEM 7

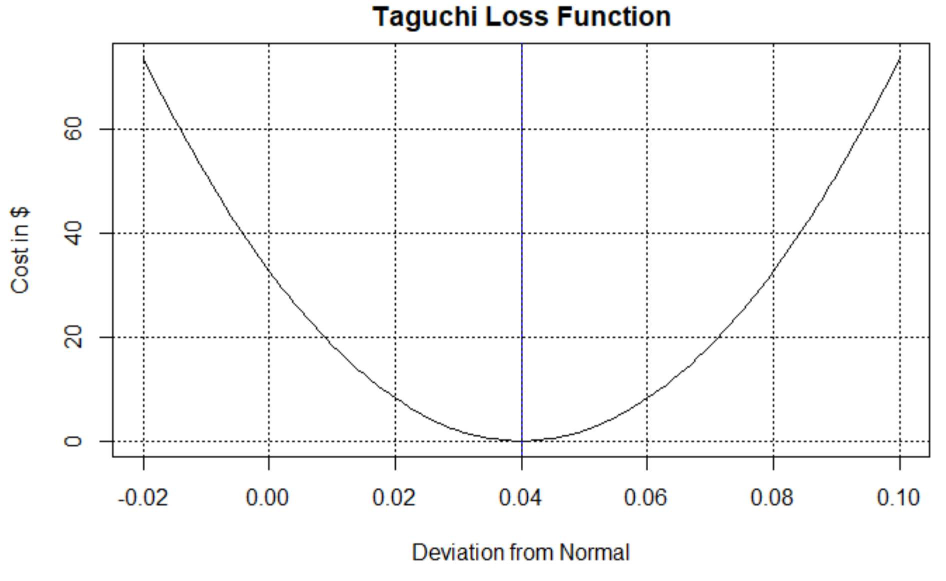
a)

1. Calculate constants

$$L(x) = k(x - T)^2$$
$$25 = k(0.035)^2$$
$$k = 30/0.035^2 = 20408.1633$$

2. Calculate loss function

$$L(x) = k(x - T)^2$$
$$L(x) = 20408.16339(x - 0,04)^2$$



In Figure 35, the increase in cost, as the deviation from specification increases, is reduced in this figure (in comparison to figure 35).

This is primarily due to the fact that scrap rate price was decreased.

Figure 35 Taguchi loss of Problem 7

b)

$$L(0,027) = 20408.1633(0,0270^2$$
$$= 14,878$$

Therefore, this company would make a loss of $14.878 per product, given that the process deviation from the target is reduced to 0,027cm.

6.2 PROBLEMS FOR TEXTBOOK

PROBLEM 27

a) Reliability = reliability(Machine A) x reliability(Machine B) x reliability(Machine C)

Reliability = 0.85 x 0.92 x 0.90 = 0.7038

b) Reliability = reliability(Machine A1 and A2) x reliability(Machine B1 and B2) x reliability(Machine C1 and C2)

Reliability = (1-(1-0.85)^2) x (1-(1-0.92)^2 x (1-(1-0.90)^2) = 0.9615

Improved reliability = 0,9615 – 0,7038 = 0,2577

This therefore shows that having two machines in parallel at each stage will improve the reliability by 25,77%. If one machine breaks down, the identical machine in parallel can still operate. The reliability of the company would greatly improve by making use of both these machines simultaneously.

6.3 BINOMIAL DISTRIBUTION

These questions were completed using the pbinom function in R.

If the number of vehicles is 20, independent of the number of drivers then,
Delivery reliability = 349.2072

If the number of vehicles is increase to 21, independent of the number of drivers then,
Delivery reliability = 363.2049

If we take the number of available drivers into account, then;
Delivery reliability = 364.9587

Therefore, is recommended to increase the number of vehicles to 22, as it increases the delivery reliability.

CONCLUSION

The analysis conducted for the valid data set revealed relationships between the different features of the data set. As is detailed in this report, it is incredibly important and useful for management to be aware and understand the affects that different features can have on each other.

Statistical process control charts were constructed and analysed, to determine whether the process was in or out of control. If the process was shown to be out of control, then this provides management with the opportunity to rectify these issues.

Through the creation of this report, it is hoped that management was able to gain further insight into the operational decisions of the business, as well as the distribution of revenue over the product line. It is hoped that management will be able to use this information to improve the business.

REFERENCES:

Hernandez, F., 2021. *Data Analysis with R - Exercises.* [online] Fch808.github.io. Available at: http://fch808.github.io/Data-Analysis-with-R-Exercises.html. Accessed on : 5 October 2022

Sthda.com. 2021. *MANOVA Test in R: Multivariate Analysis of Variance - Easy Guides - Wiki - STHDA.* [online] Available at: http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance. Accessed on: 8 October 2022.

University of Minnesota, 2020. Probability Distributions in R. [online] Available at: https://www.stat.umn.edu/geyer/old/5101/rlook.html#:~:text=pbinom%20is%20the%20R%20function,do%20exactly%20the%20same%20thing. Accessed on:10 October 2022.