ECSA GA4 Report

by

C Gruber

22846042

The Department of Industrial Engineering,

Stellenbosch University

September 2022

List of figures

List of tables

Glossary

USL: Upper Specification Limit

LSL: Lower Specification Limit

CP: Process Capability

CPU: Upper one-sided Index

CPL: Lower one-sided Index
CPK: Process Capability Index

SPC: Statistical process control

Table of Contents

1 INTRODUCTION

This report documents the visual and descriptive statistical analysis of an online retailer. Client data is provided in a excel document called ‘salesTable2022’. The data set introduces various variables with respect to the company’s Sales data. The data set includes both historical, and predicted sales.

The body of the report is divided into six parts. Starting with data wrangling where the data will be analyzed and cleaned in order to be used. Secondly descriptive statistics will show various visualizations of the data. This will provide further information about the data and relationships that exists between the variables. Thirdly, statistical process control charts will be used on the delivery times to determine whether a process is in or out of control. Thereafter, the delivery process will be optimized followed by MONAVA testing. Lastly the reliability of the service and products are calculated.

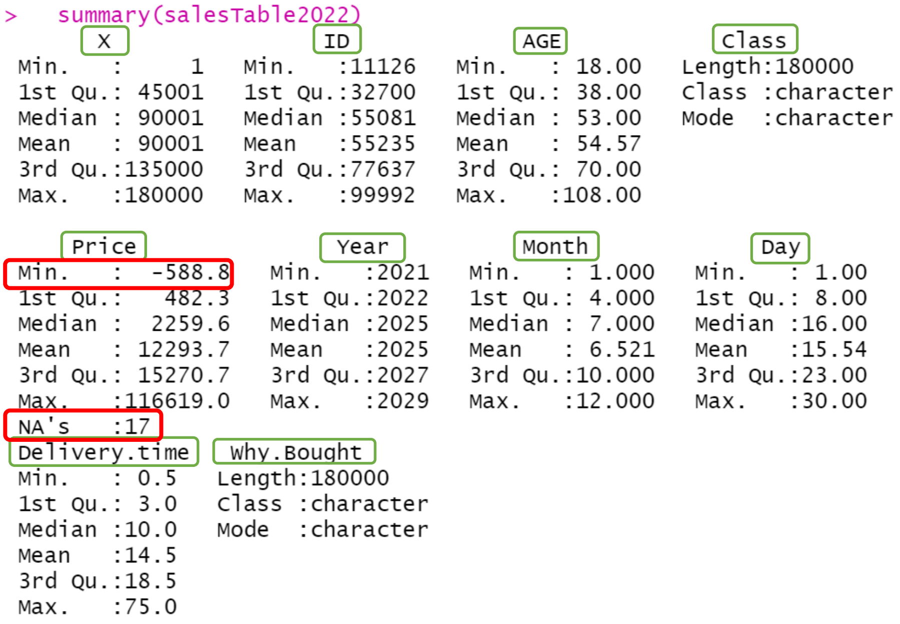
The report ends with a conclusion of the analysis, followed by references.

2 DATA WRANGLING

The ‘salesTable2022.csv’ data set was provided to the student and imported to be analysed. The data is summarized to learn more about each variable and remove any data quality issues prior to the analysis and manipulation process begins.

2.1 Data preparation

The data from ‘salesTable2022’ is summarized in Figure 1. This makes it possible to understand the data better and identify any data quality issues. It is seen that there are ten features to consider. These features are outlined in green. In the feature, *Price,* there is a data quality issue of missing values (incomplete values) and negative prices. These missing values are indicated by NA and negative prices by a minus sign.

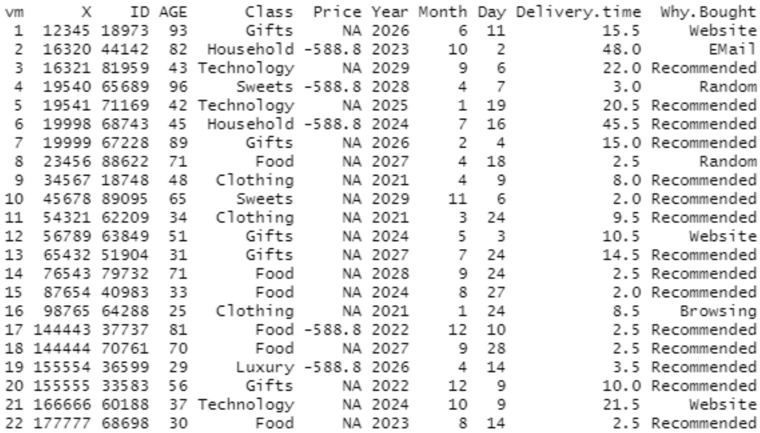


```
>  summary(salesTable2022)
       X                ID             AGE             Class          
 Min.   :     1   Min.   :11126   Min.   : 18.00   Length:180000     
 1st Qu.: 45001   1st Qu.:32700   1st Qu.: 38.00   Class :character  
 Median : 90001   Median :55081   Median : 53.00   Mode  :character  
 Mean   : 90001   Mean   :55235   Mean   : 54.57                     
 3rd Qu.:135000   3rd Qu.:77637   3rd Qu.: 70.00                     
 Max.   :180000   Max.   :99992   Max.   :108.00                     

     Price               Year          Month             Day       
 Min.   :   -588.8   Min.   :2021   Min.   : 1.000   Min.   : 1.00  
 1st Qu.:    482.3   1st Qu.:2022   1st Qu.: 4.000   1st Qu.: 8.00  
 Median :   2259.6   Median :2025   Median : 7.000   Median :16.00  
 Mean   : 12293.7   Mean   :2025   Mean   : 6.521   Mean   :15.54  
 3rd Qu.: 15270.7   3rd Qu.:2027   3rd Qu.:10.000   3rd Qu.:23.00  
 Max.   :116619.0   Max.   :2029   Max.   :12.000   Max.   :30.00  
 NA's   :17                                                        
 Delivery.time    Why.Bought       
 Min.   : 0.5   Length:180000     
 1st Qu.: 3.0   Class :character  
 Median :10.0   Mode  :character  
 Mean   :14.5                     
 3rd Qu.:18.5                     
 Max.   :75.0                     
```

Figure 1: Summary of the data from 'salesTable2022.csv'

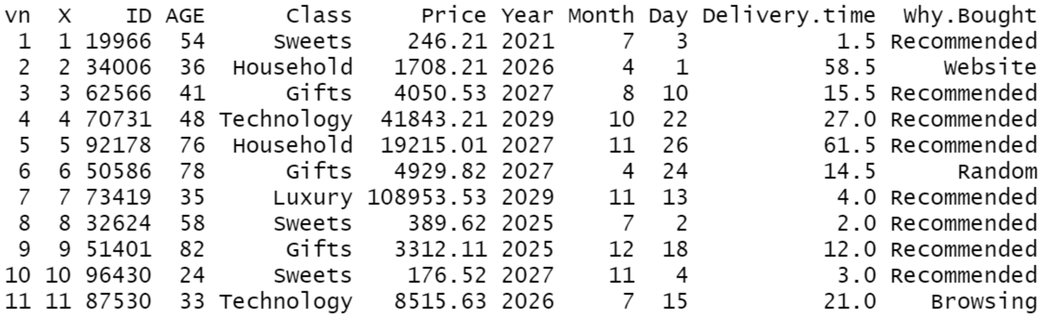
There are seventeen missing values in the variable *Price and five* negative values, which means there are 22 invalid data instances from a total of 180000 data instances. These missing and negative values are removed, and a new data set is created in a file called ‘df_invaliddata.csv’ with invalid data. Figure 2 shows all the invalid data instances.



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

Figure 2: Invalid data instances from ‘salesTable2022.csv’

The rest of the data which includes only valid data is created into a new file called ‘df_validdata.csv’. Figure 3 shows the first 11 instances in the valid data set.



```
vn  X    ID AGE      Class     Price Year Month Day Delivery.time  Why.Bought
 1  1 19966  54     Sweets    246.21 2021     7   3           1.5 Recommended
 2  2 34006  36  Household   1708.21 2026     4   1          58.5     Website
 3  3 62566  41      Gifts   4050.53 2027     8  10          15.5 Recommended
 4  4 70731  48 Technology  41843.21 2029    10  22          27.0 Recommended
 5  5 92178  76  Household  19215.01 2027    11  26          61.5 Recommended
 6  6 50586  78      Gifts   4929.82 2027     4  24          14.5      Random
 7  7 73419  35     Luxury 108953.53 2029    11  13           4.0 Recommended
 8  8 32624  58     Sweets    389.62 2025     7   2           2.0 Recommended
 9  9 51401  82      Gifts   3312.11 2025    12  18          12.0 Recommended
10 10 96430  24     Sweets    176.52 2027    11   4           3.0 Recommended
11 11 87530  33 Technology   8515.63 2026     7  15          21.0    Browsing
```

Figure 3: First 11 data instances in the valid data set orginally from 'salesTable2022.csv'

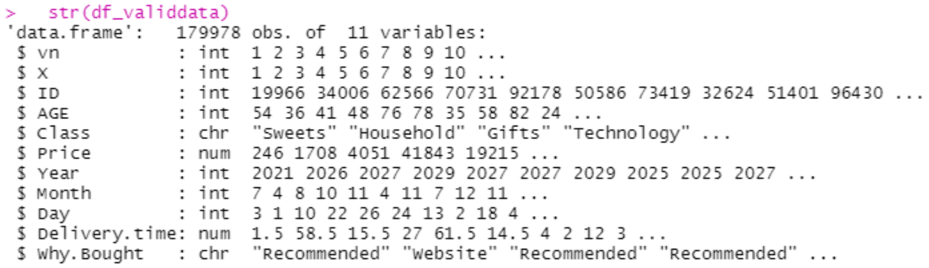
3 PART 2: DESCRIPTIVE STATISTICS

The valid data set was analysed using descriptive statistics and data visualization in order to uncover important data set properties, as well as inter-variable trends and relationships.

3.1 Data Analysis

A summary of the valid data is in shown in Figure 4.

There are 11 features (variables) and 179978 observations.



```
>  str(df_validdata)
'data.frame':   179978 obs. of  11 variables:
 $ vn          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ ID          : int  19966 34006 62566 70731 92178 50586 73419 32624 51401 96430 ...
 $ AGE         : int  54 36 41 48 76 78 35 58 82 24 ...
 $ Class       : chr  "Sweets" "Household" "Gifts" "Technology" ...
 $ Price       : num  246 1708 4051 41843 19215 ...
 $ Year        : int  2021 2026 2027 2029 2027 2027 2029 2025 2025 2027 ...
 $ Month       : int  7 4 8 10 11 4 11 7 12 11 ...
 $ Day         : int  3 1 10 22 26 24 13 2 18 4 ...
 $ Delivery.time: num  1.5 58.5 15.5 27 61.5 14.5 4 2 12 3 ...
 $ Why.Bought  : chr  "Recommended" "Website" "Recommended" "Recommended" ...
```

Figure 4: Summary of valid data instances from 'df_validdata.csv' data file

3.1.1 *Product class vs price trends*

3.1.1.1 Class of product count and price range frequency

Figure 5 shows that the ‘gifts’ class is sold most frequently and ‘Technology’ following closely behind. Luxury goods are sold the least. In figure 6 it can be seen that the graph is skewed to the right. This indicates that the most frequent bought items are in the lower price range.

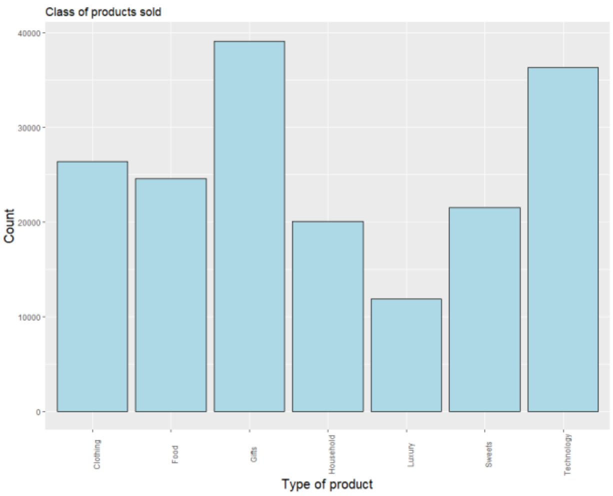


Figure 5: Histogram showing frequency of products sold per class

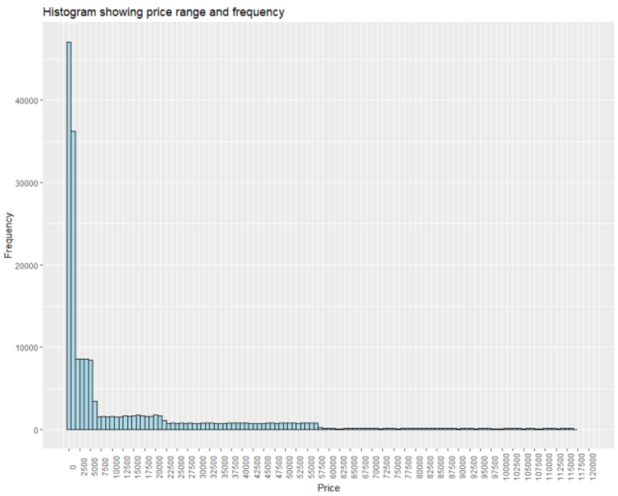


Figure 6: Histogram showing the price range and frequency

3.1.1.2 Price vs product class

In figure 7 it is seen that clothing, food, and sweets have the smallest price ranges as well as the smallest means. This is accurate as they are all product classes which contain everyday necessity type products. These product classes form part of a highly competitive market sector which are purchased frequently by customers at a priced made them accessible to many people.

Luxury and technology have the largest price ranges and means. This is expected as luxury goods are considered expensive and technology contain a large range of products of variable prices, with the top of the range prices being significantly higher than the bottom.

Management can use this information to determine the degree to which each class is capable of generating income, and the relative amount of income that each class generates.

This information also allows management to determine its target customer for each product class, which can assist in targeted marketing and production strategies. The more expensive products should be marketed and promoted strategically at the more affluent (higher income class) customers, whereas the necessity products should be marketed and promoted to target a wider range of customer groups.

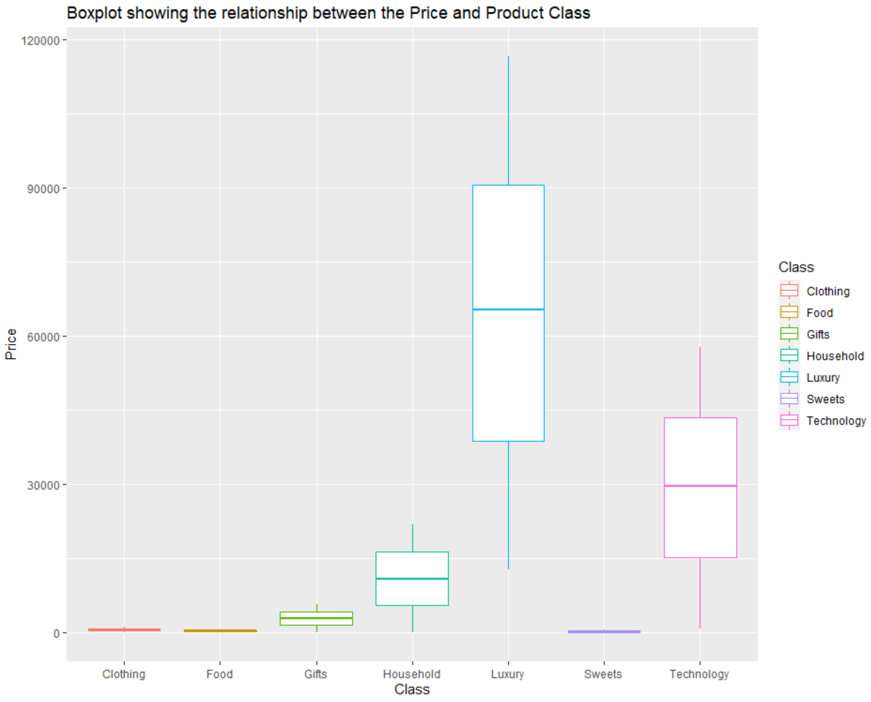


Figure 7: Box plot showing the relationship price and class

3.1.2 Product class vs purchasing reasons

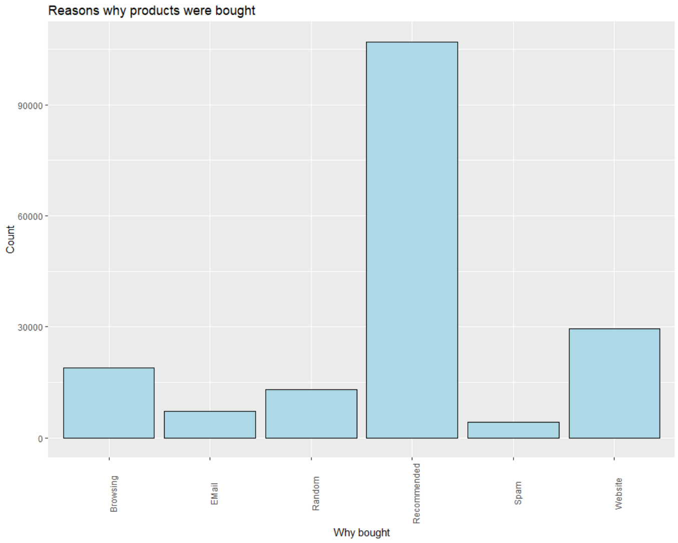


Figure 8: Histogram showing purchasing reason count

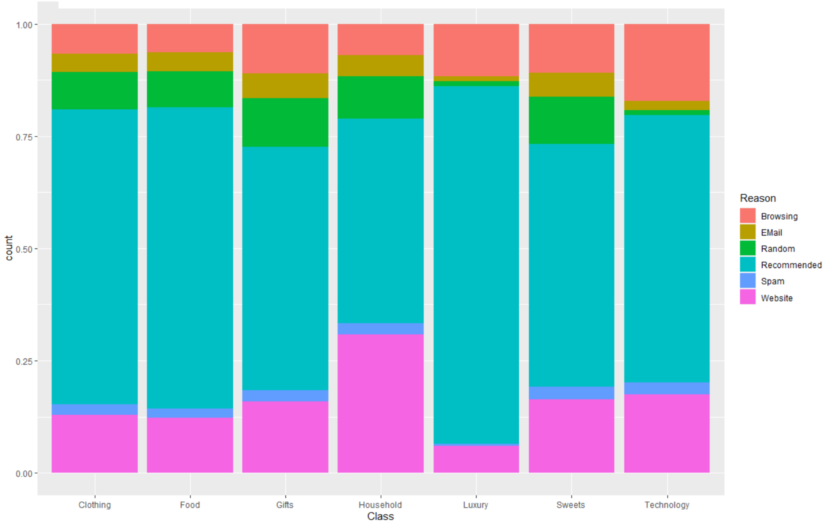


Figure 9: Stacked bar plot of purchasing reason for each product class

Figure 8 clearly shows that the main element when products are bought are when they are recommended. The second reason is when products are found on a website. Figure 9 shows the distribution of reasons why products are bought and all classes show a similar purchasing reason trend. Luxury products are mostly bought when recommended. Household products are mostly bought on websites.

The high recommended counts indicates that the company has a good market and product reputation, which management should work to maintain. Furthermore, the trends indicate the successful marketing, since website purchasing is second most frequent. Management should continue to improve to increase company and product exposure, which would lead to increased sales.

3.1.3 Product class vs date trends

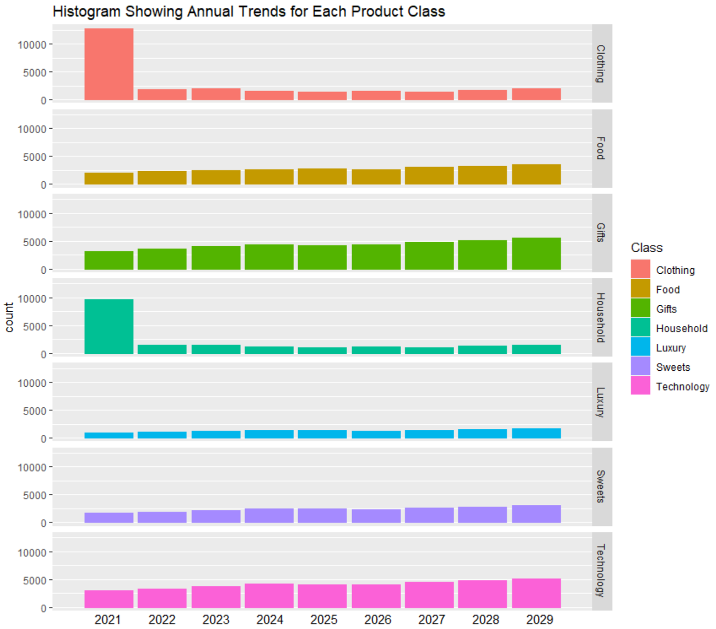


Figure 11: Histogram showing annual trends for each product class

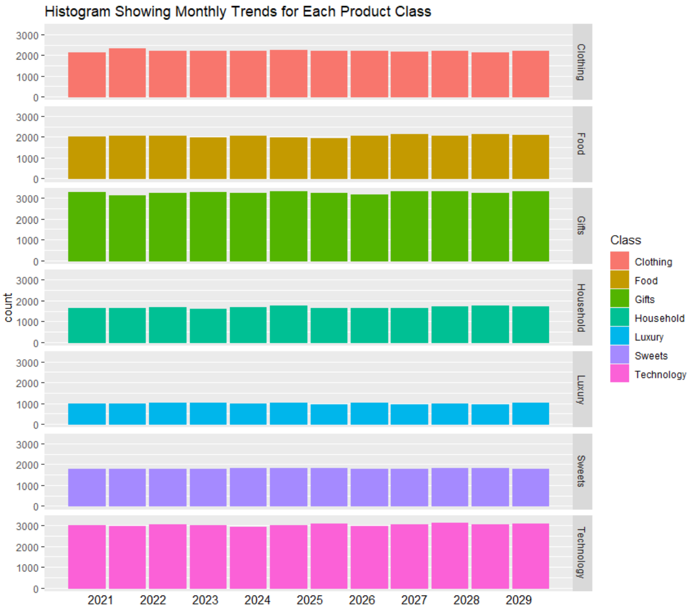


Figure 10: histogram showing monthly trends for each product class

Figure 10 indicates that there are no significant monthly trends per product class. Gifts and Technology are most frequently bought all year round. This may be as a result of various factors. Gifts have a very wide product range and thus a large variety of items would be classified under gifts when compared to more specific product classes (such as clothing). Furthermore, gifts are purchased all year round for multiple reasons and occasions, which further increases their count. Technology also has a wide product range, as well as being an emerging, very high demand market, which all increase its count.

As seen in Figure 11, in 2021 there was a significant spike in purchasing household and clothing items. This could be due to the covid – 19 pandemic causing people to have tight finances and inly purchasing necessities.

These trends can be used by management for product manufacturing and purchasing planning. The information would help to ensure that management plan for the production of the required amount of each product class throughout the year, as well as focusing on developing and keeping enough stock of the higher demand classes, being gifts and technology.

3.1.4 Product class vs age trends

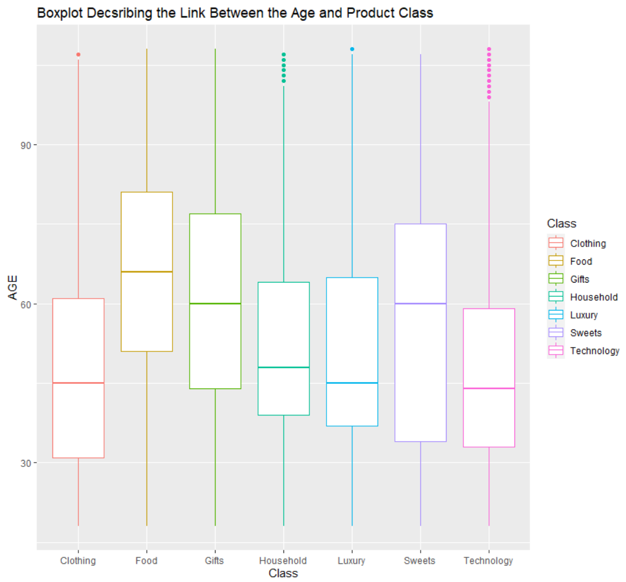


Figure 13: Box plot showing the link between age and product class

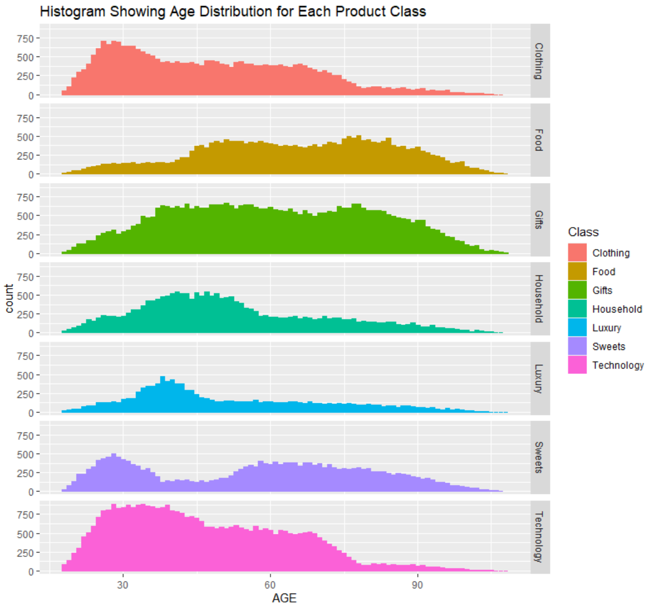


Figure 12: Histogram showing the age distribution for each product class

As seen in figure 13, the mean purchasing age of clothing, household, and technology are lower than that of the other product classes. This shows that gifts, sweets and food are mainly purchased by older customers. Figure 12 indicates that all of the distributions are skewed, except for the gift's distribution. This is expected, as gifts are bought by all age groups for various reasons. Management can use these age group data trends to strategically develop marketing and promotional strategies.

3.1.5 Product class vs delivery times

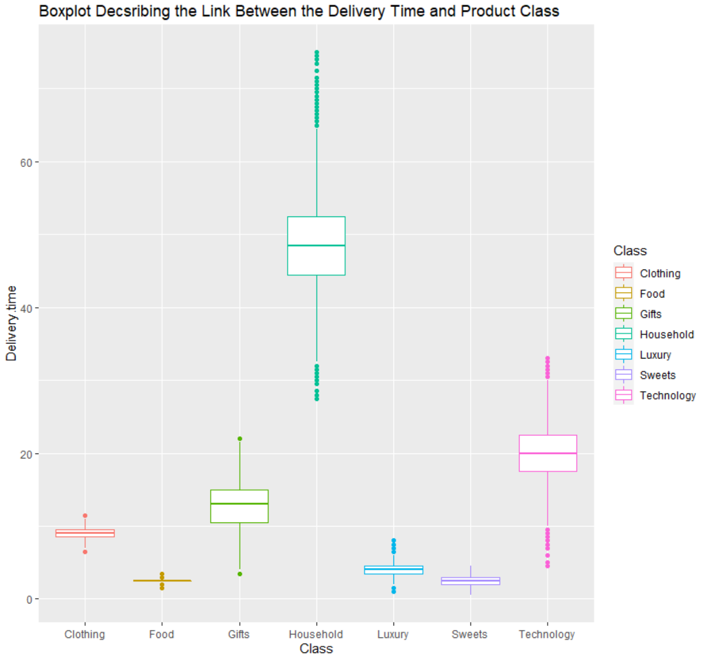


Figure 14: Boxplot describing the link between the delivery time (in hours) and product class

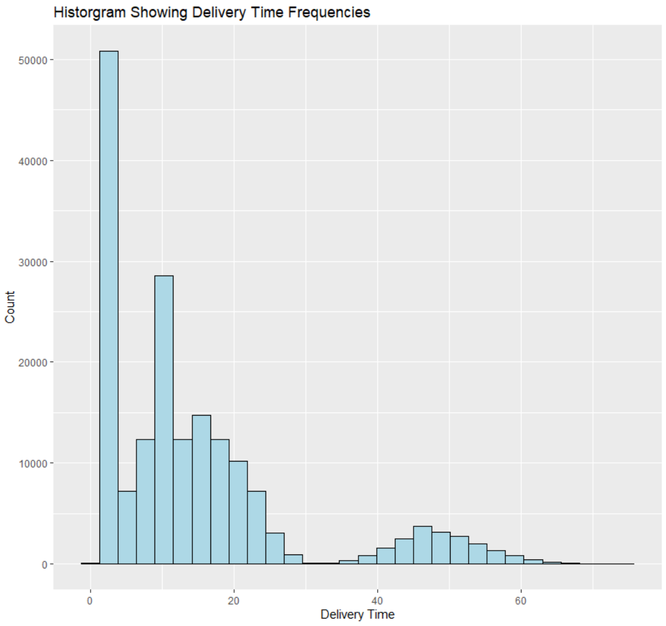


Figure 15: Histogram showing delivery time frequencies in hours

Figure 15 shows the overall delivery time in hours achieved by the company follows a right skewed, multi-modal distribution. The most occurring delivery times are significantly lower than the remaining. This shows that, despite the variation, the overall delivery process time achieved by the company is less than 30 hours.

Figure 14 indicates that food and sweet have the quickest delivery times. This is expected as they are perishable products and ready-made products which need to delivered quickly once an order has been placed. Furthermore, these products are usually ordered locally, and thus have short delivery distances, associated with shorter delivery times.

3.2 Process Capability indices

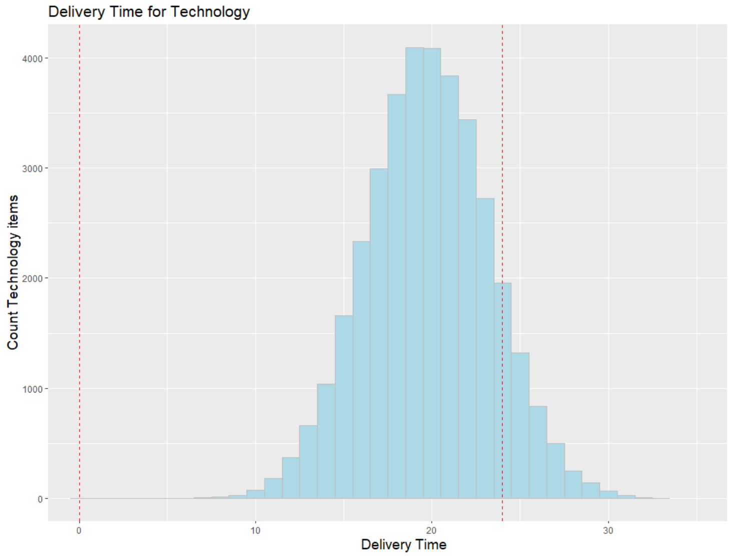


Figure 16: Histogram showing the relationship between delivery times in hours and technology items

Process Capability is "the ability of a process to produce output that conforms to specifications." Process capability is measured by computing numerical indices. Figure 16 shows the distribution of delivery times for technology items. The red dotted lines indicate the lower and upper specification limit

Using USL = 24 and LSL = 0.

The following capability indices are obtained:

Index	**Value**	**Interpretation**
CP	1.142	Process is capable of meeting requirements
CPU	0.379	The process mean is not centred between the specification limits.
CPL	1.904	The process mean is closer to the upper specification limit.
CPK	0.379	The process is not actually very capable of meeting requirements

Table 1: Process capability indices

A LSL of zero is logical, as it is impossible to achieve a delivery time of less than 0 hours.

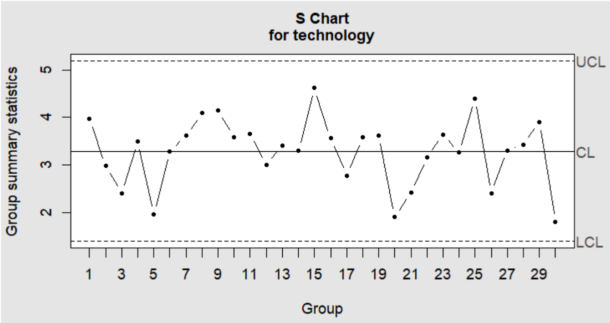
CP is more than one, which means process is capable. CPK is less than CP which means process is not centred between the specific limits. This can also be seen in Figure 16. The process needs to be improved by moving the mean to the left by improving the delivery time for technology.

4 PART 3: STATISTICAL PROCESS CONTROL

Statistical Process Control is a method used to help monitor and analyse process conditions to determine the processes performance and prescribe corrective actions that are required. (Molaksingh, 2022)

The X&s charts are constructed using 30 samples of 15 Sales each. Before the charts can be calculated and constructed, the data first needs to be ordered according to the date. The data is ordered from oldest to newest according to the year, then month, then day.

4.1 Analysis of first 30 samples



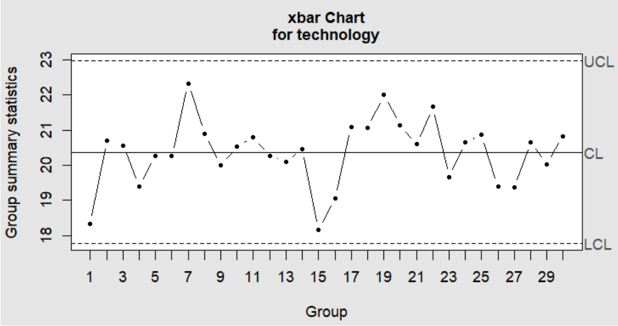
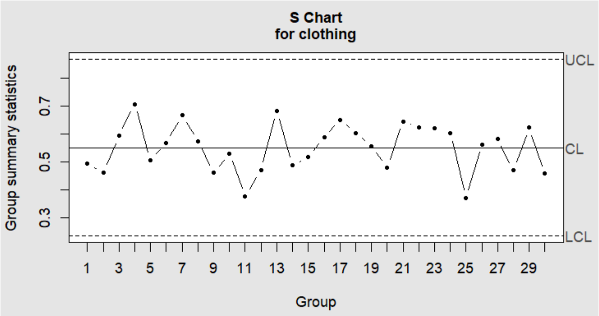


Figure 18: S Chart for technology class

Figure 17: X bar chart for technology class

In figure 18, the first 30 samples in the s chart indicates that the technology class is in control and is no cause of variation in the process of technology orders. The X-bar chart, in figure 17 can be evaluated due to the satisfactory S bar chart.



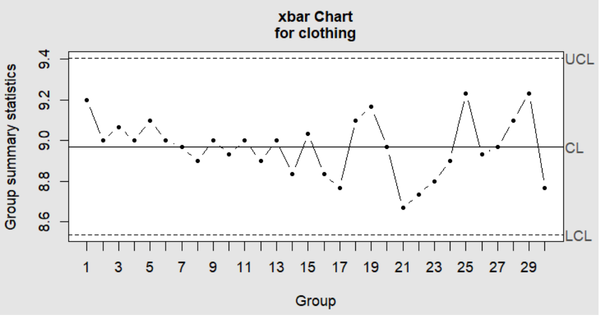


Figure 20: S chart for clothing class

Figure 19: X bar chart for clothing class

In figure 20, the first 30 samples indicates that the clothing class is in control and is no cause of variation in the process of clothing orders. The X-bar chart, in figure 19 can be evaluated due to the satisfactory S bar chart.

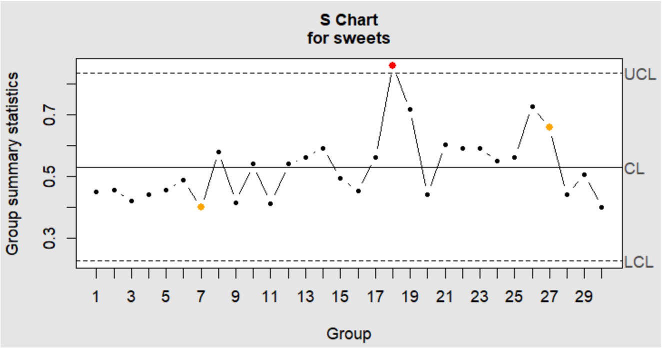


Figure 22: S chart for sweets class

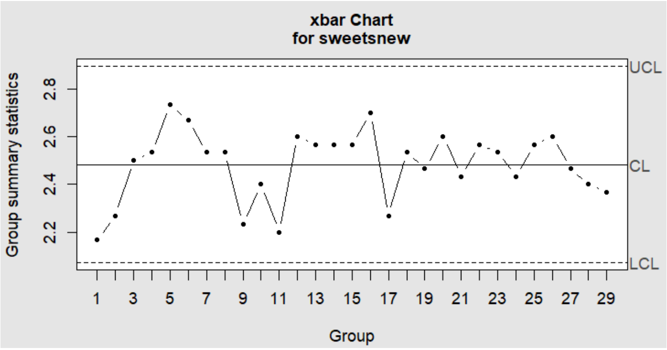


Figure 21: X bar chart for sweets class

In figure 22, the first 30 samples indicates that sweets class is in control, except in the S chart, sample 18's standard deviation is out of the control limits. This indicates that sample 18 needs to be removed before creating the x bar chart. This was done and therefore the x bar chart in figure 21 can be evaluated.

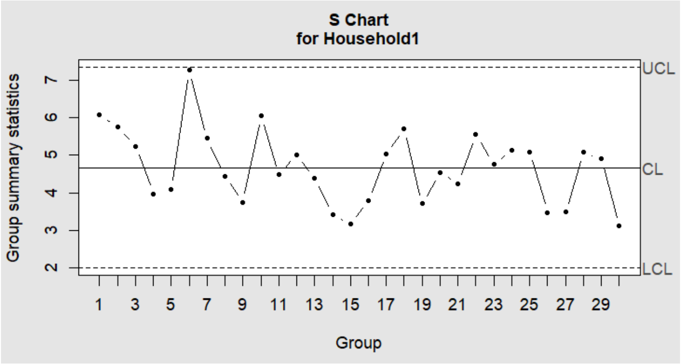


Figure 24: S chart for household class

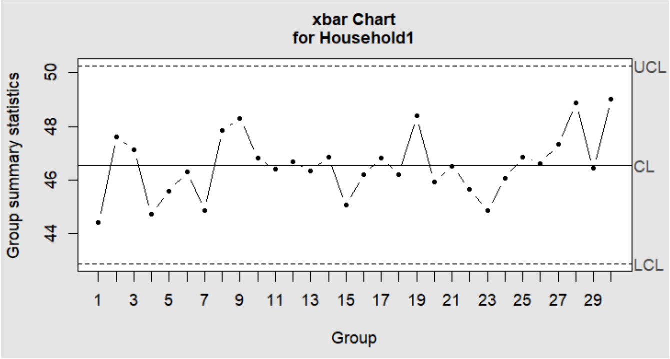


Figure 23: X bar chart for household class

In figure 24, the first 30 samples indicates that the household class is in control and is no cause of variation in the process of household orders. The X-bar chart, in figure 23 can be evaluated due to the satisfactory S bar chart.

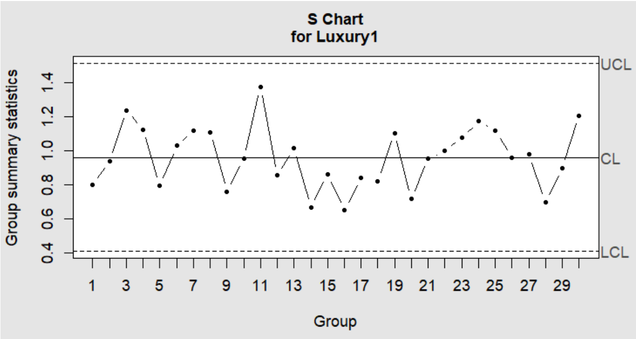


Figure 25: S chart for luxury class

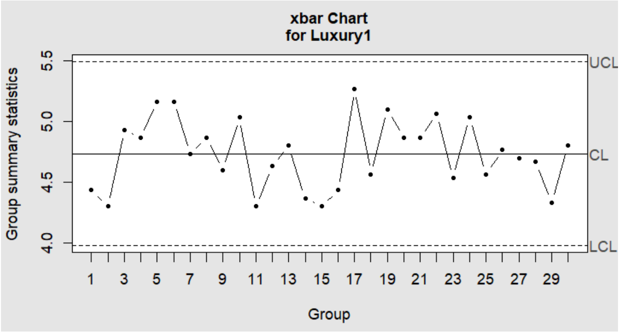


Figure 26: X bar chart for luxury class

In figure 26, the first 30 samples indicates that the luxury class is in control and is no cause of variation in the process of luxury orders. The X-bar chart, in figure 25 can be evaluated due to the satisfactory S bar chart.

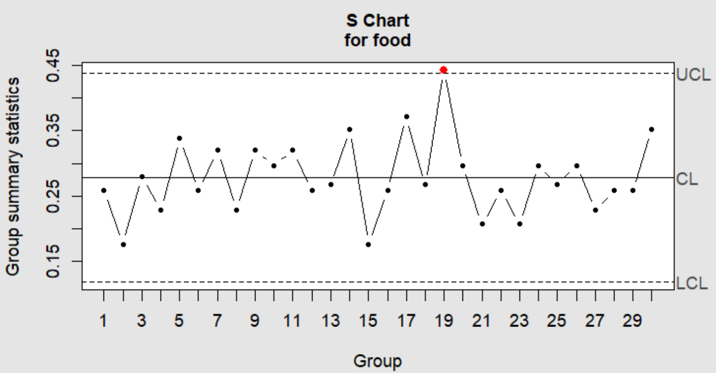


Figure 28: S chart for food class

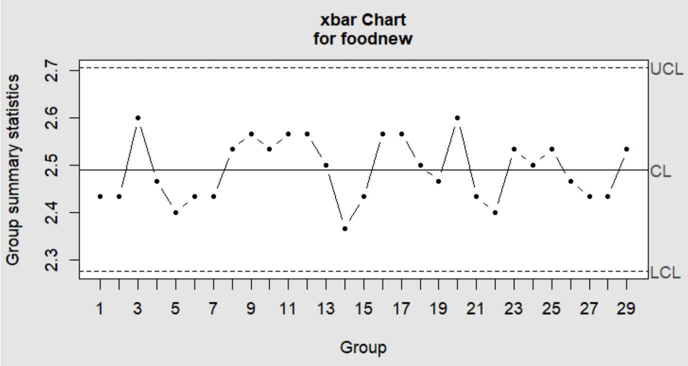


Figure 27: X bar chart for food class

In figure 28, the first 30 samples indicates that the food class is in control, except sample 19's standard deviation is out of the control limits. This indicates that sample 19 needs to be removed before creating the x bar chart. This is done which can be seen in figure 27 as it contains 29 samples. The x bar chart can now be evaluated.

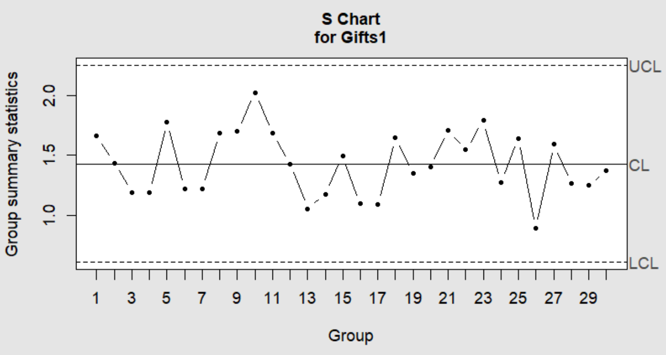


Figure 30: S chart for gifts class

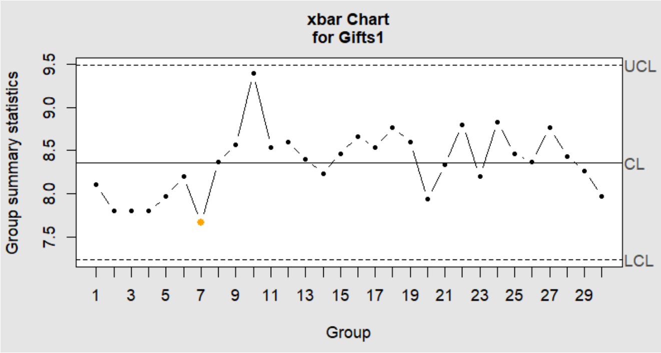


Figure 29: X bar chart for gifts class

In figure 30, the first 30 samples indicates that the gifts class is in control and is no cause of variation in the process of gifts orders. The X-bar chart, in figure 29 can be evaluated due to the satisfactory S bar chart.

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22.9731	22.10688	21.24066	20.37444	21.24066	22.10687	17.77579
Clothing	9.404681	9.259787	9.114894	8.97	9.114894	9.259787	8.535319
Sweets	2.892867	2.757125	2.617451	2.482759	2.617451	2.757125	2.07265
Household	50.24618	49.018193	47.790207	46.56222	47.790207	49.018193	42.87826
Luxury	5.493524	5.240868	4.988212	4.735556	4.988212	5.240869	3.977587
Food	2.705683	2.636220	2.563110	2.490805	2.563110	2.636220	2.275926
Gifts	9.487909	9.112310	8.736710	8.361111	8.736710	9.112310	7.234313

Table 2: X bar chart table showing control limits of each class

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.179912	4.551784	3.923656	3.295528	3.923656	4.551785	1.411143
Clothing	0.8664496	0.7613819	0.6563142	0.5512465	0.6563142	0.7613818	0.2360435
Sweets	0.8352331	0.7339508	0.6326685	0.5313862	0.6326685	0.7339508	0.2275393
Household	7.343248	6.4527886	5.5623293	4.67187	5.562329	6.4527880	2.000493
Luxury	1.51086	0.594808	0.778019	0.9612289	0.778019	0.594808	0.4115978
Food	0.4371911	0.172117	0.225132	0.2781467	0.225132	0.172117	0.1191023
Gifts	2.246048	1.973687	1.701326	1.428965	1.701326	1.973687	0.6118823

Table 3: s chart table showing control limits of each class

4.2 Analysis of all samples

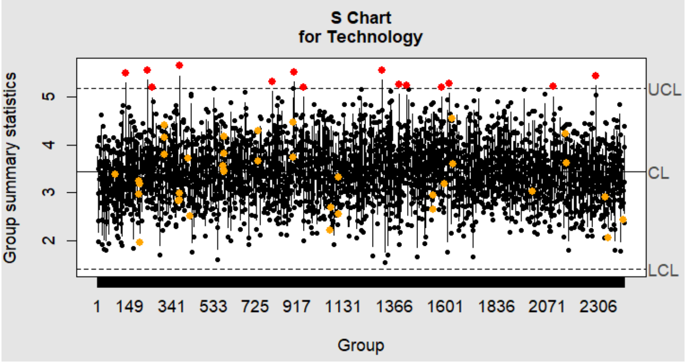


Figure 32: S chart for technology class on all samples

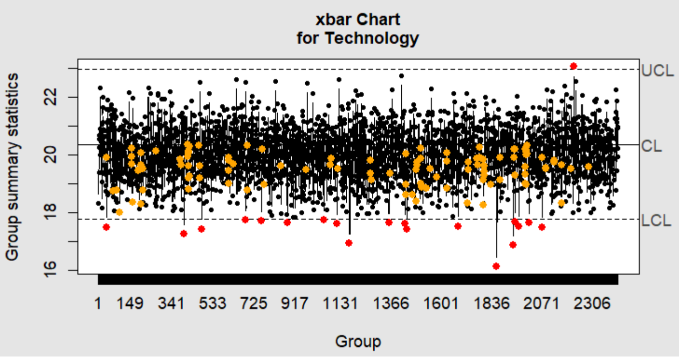


Figure 31: X bar chart for technology class on all samples

In Figure 32, it can be seen that most of the samples in the technology class are between the control limits. In the S chart there are only 14 samples that are out of the control, and most samples are even distributed around the center line. Therefore, the X-bar chart in figure 31 is appropriate to use as a guidance for managers to analyze samples that are above the UCL in the x bar chart. Samples that are above the UCL in the x bar chart are flagged since they deviating too much from the mean which means their delivery times is too long.

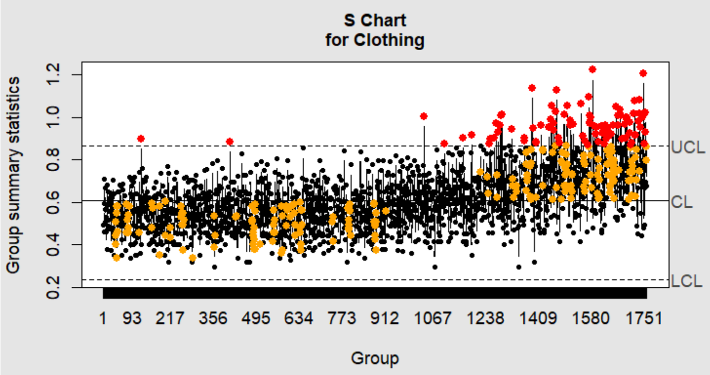


Figure 33: S chart for clothing class on all samples

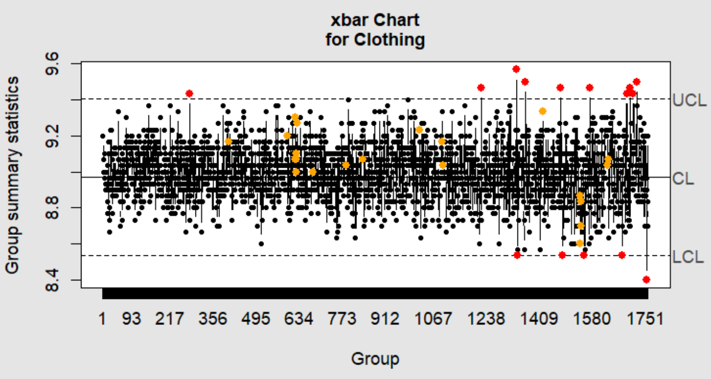


Figure 34: X bar chart for clothing class on all samples

In Figure 33, it can be seen that many samples are out of control limits. From the total number of 1760 groups, 90 are beyond the control limits in the S chart, which means those samples increasingly deviate from the center line and will make the x bar chart, in figure 34 less accurate to analyze.

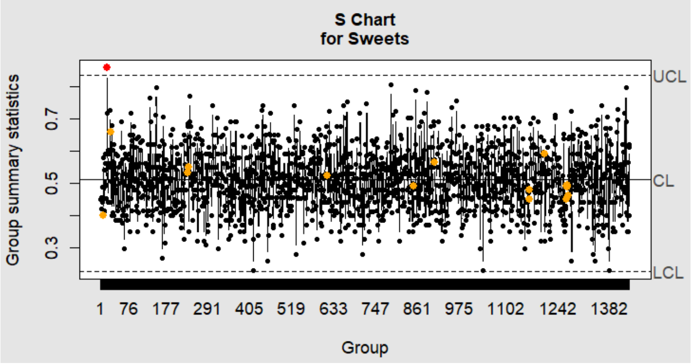


Figure 36: S chart on sweets class on all samples

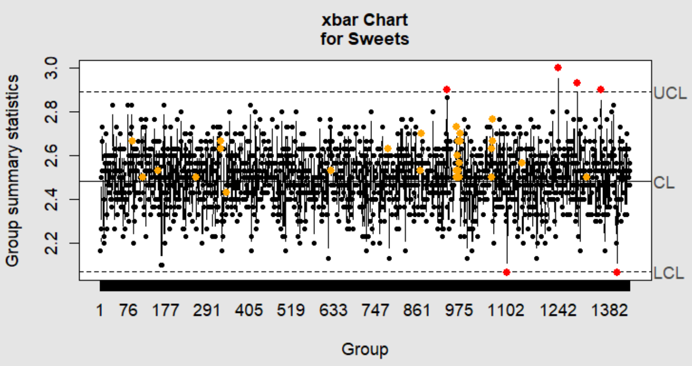


Figure 35: X bar chart for sweets class on all samples

In Figure 35, there is only one sample in the sweets class that is out of the control limits, while 1436 samples in the S chart are between the control limits. Therefore, the X-bar chart in figure 36 is appropriate to use. In the x bar chart, the outliers above the UCL should be flagged for the manager to analyze why the delivery times are taking too long.

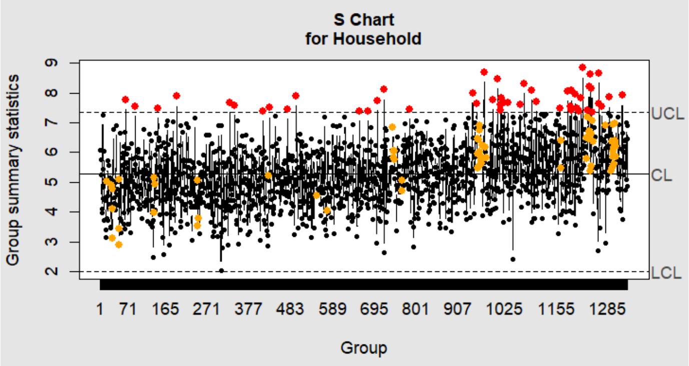


Figure 38: S chart for household class on all samples

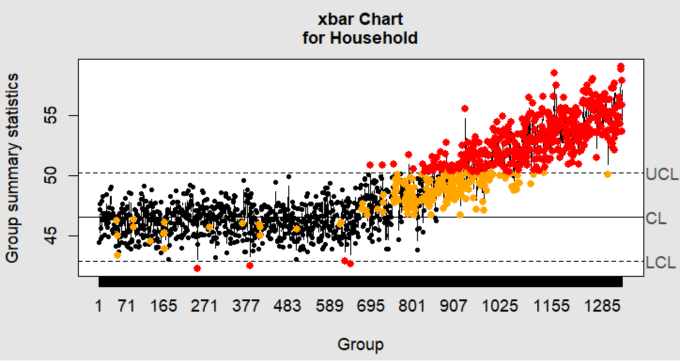


Figure 37: X bar chart for household class on all samples

In figure 37, the S chart there are 90 samples out of 1337 that are out of the control limits indicating the S chart is out in control. All the outliers are above the UCL indicating a far deviation from the center line. Therefore, the X-bar chart in figure 38 is in line with the s chart. Figure 38 shows that that the delivery time for household products increases significantly over time with a significant number of samples out of control limits indicating a unstable variation in data.

There is a significant upward trend in the samples from samples 907 onwards. There is a need to investigate why the delivery times increase continuously. One reason which could cause this trend is the raising use of online shopping, which can cause the delivery times to be put under pressure.

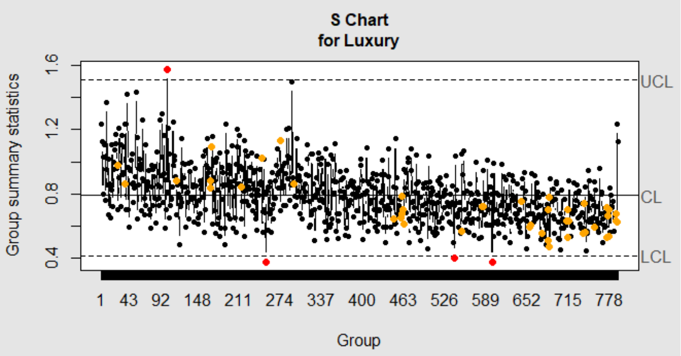


Figure 40: S chart for luxury class on all samples

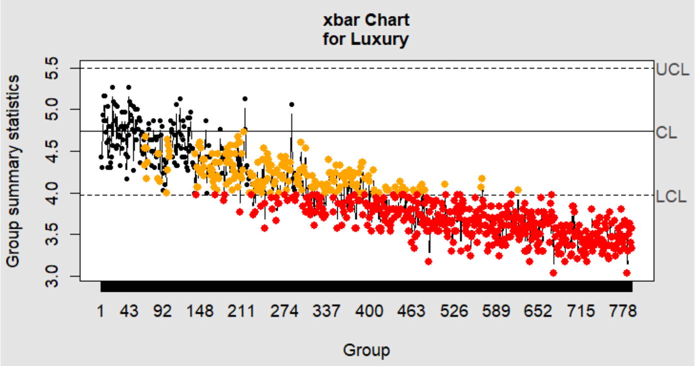


Figure 39: X bar chart for luxury class on all samples

In Figure 40, there is only one sample in the Luxury class that is out of the control limits above the UCL indicating the deviation from the center line is large for that sample. The rest of the samples are hugged around the centerline with a slight downward trend. This is reflected in the x bar chart with the majority of the samples out of control. Figure 39 exhibits a downwards trend with the delivery times of luxury items decreasing. This can be seen as a good aspect for the company as they are improving.

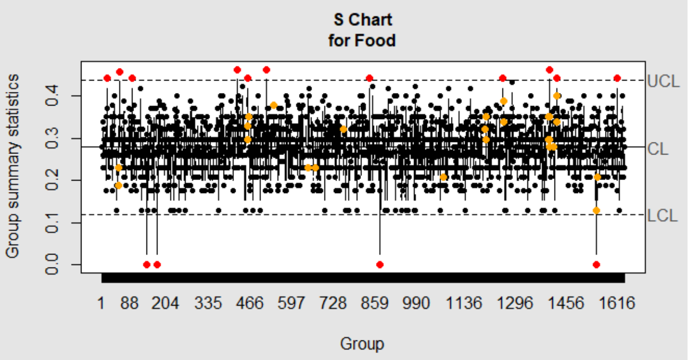


Figure 42: S chart for food class on all samples

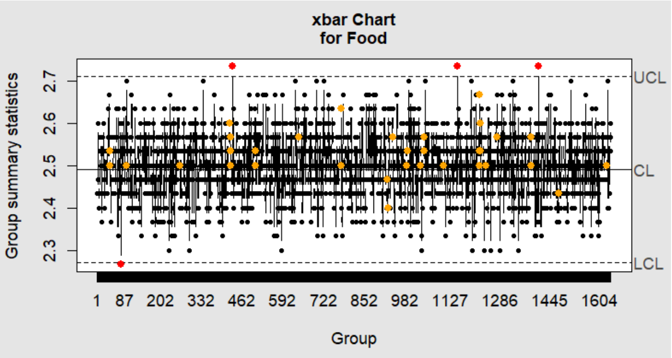


Figure 41: X bar chart for food class on all samples

In Figure 41, it can be seen that most of the samples in the food class are between the control limits. In the S chart there are only 15 samples that are out of the control, and most samples are even distributed around the center line. Therefore, the X-bar chart in figure 42 is appropriate to use as a guidance for managers to analyze samples that are above the UCL in the x bar chart. Samples that are above the UCL in the x bar chart are flagged since they are deviating too much from the mean which means their delivery times is too long.

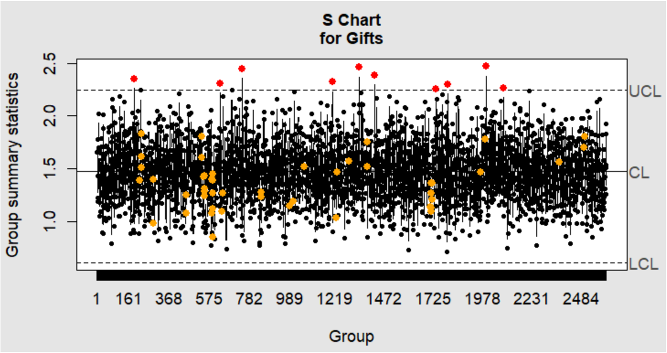


Figure 44: S chart for gifts class on all samples

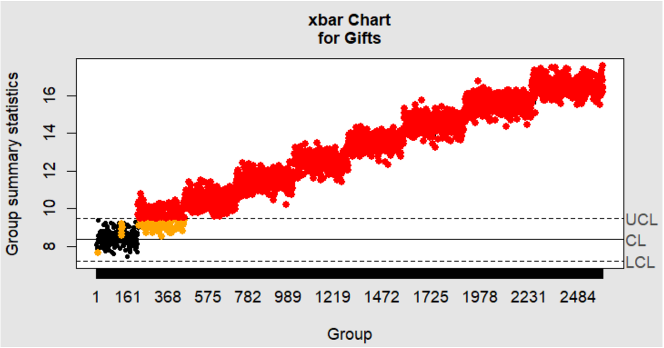


Figure 43: X bar chart for gifts class for all samples

In Figure 43, it can be seen that most of the samples in the Gifts class are between the control limits. In the S chart there are 10 samples that are out of the control. The 10 samples out of control are all above the UCL indicating a large deviation from the center line upwards. This indicates an increase in delivery time. Figure 44 shows a sharp increase in mean of samples. All the samples from sample 161 are above the UCL. This indicated a huge increase in delivery time for food which needs to be investigated. Upward trends are not a good thing since they show that the process is “worsening”. The delivery times for this product class is much slower than before, as shown by the fact that the mean of the delivery times is periodically increasing drastically.

5 PART 4: OPTIMIZING THE DELIVERY PROCESS

5.1 Part 4.1 Out of control processes

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

Class	Total	1st sample	2nd sample	3rd sample	3rd last sample	2nd last sample	Last sample
Technology	20	37	398	483	2009	2071	2218
Clothing	15	282	1222	1337	1713	1723	1756
Sweets	6	942	1104	1243	1294	1358	1403
Household	396	252	387	629	1335	1336	1337
Luxury	434	140	169	182	787	788	789
Food	4	73	430	1147	-	-	1406
Gifts	2291	213	216	218	2608	2609	2610

Table 4: Samples out of control for x bar means

As seen in table 4, the product classes with the highest number of outliers in descending order are:

1. Gifts
2. Luxury
3. Household

This corresponds to the graphs in part 3 since these graphs showed a large deviation from the mean. This problem for Household, Luxury and Gifts items should be further investigated to determine exactly what is causing the delivery times to be outside the control limits.

The following graphs show the first 3 and last 3 outliers of the Gifts class since it most out of control class. This means the deliveries for the gifts class will almost always be not as expected. Figure 45 shows the last 10 samples of the gifts class, and they are all out of control which tells us the predicted delivery times in 2029 are very inaccurate. Figure 46 shows the first 3 outliers.

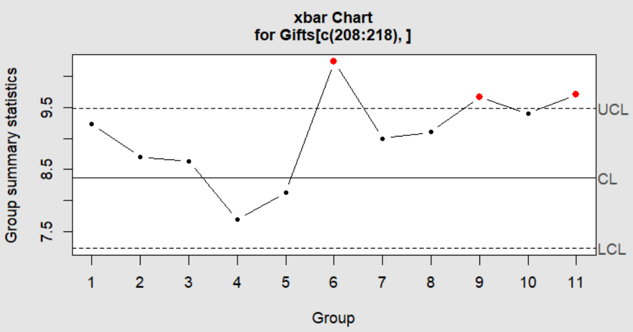


Figure 46: x bar chart of first 3 outliers of gifts class

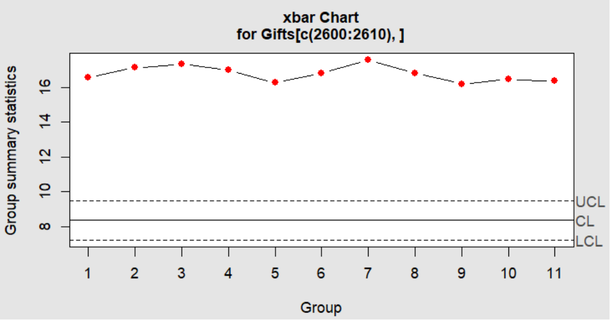


Figure 45: X bar chart showing last 3 outliers of gifts class

Rule B: most consecutive samples of "s-bar" between -0.3 and +0.4 sigma control limits

Class	Longest Pattern	Ending index
Technology	6	372
Clothing	4	223
Sweets	4	179
Household	4	46
Luxury	4	61
Food	5	754
Gifts	9	513

Table 5: Most consecutive samples of s bar chart between sigma limits

Therefore, the product class type with the largest consecutive sample string is 'Gifts'

5.2 Part 4.2 Type 1 manufacturers error

A type I error is the probability that the results indicate something is wrong, when in reality it is 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.

Question	Probability of performing type 1 error
A	0.002699796, thus 0.27% This indicates the probability of making a mistake that the product in reality is delivered on time, but company thinks the products is not delivered on time
B	0.7266668, thus 72.67%

Table 6: Type I error probabilities

5.3 Part 4.3 Delivery process best profit

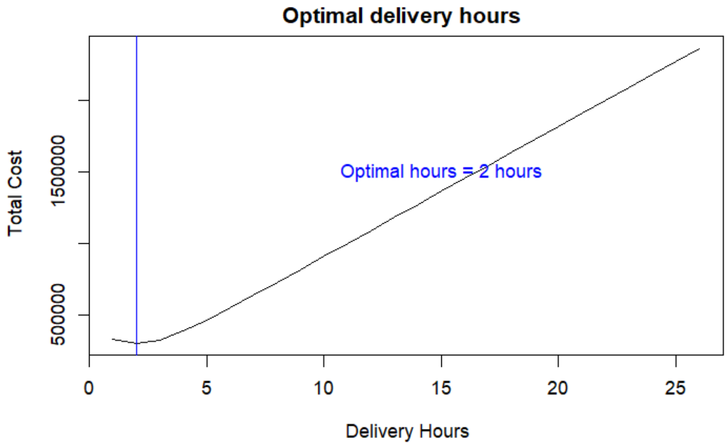


Figure 47: Optimal delivery hours

The cost for the relevant hours is compared to one another by looping through each hour. Currently the mean hour for delivering technology products are 20.01 hours. The total amount of sales beyond 26 hours are 1356 sales. If each lost sale costs R329, this will yield a total loss of R446 124. If it costs R2.5 per hour over 26 hours to move the mean of the distribution to the left and taking the whole distribution into account, will yield R636 125.

Therefore, the minimal delivery cost for technology needs to be calculated to decrease the loss in sales but also taking the effect of moving the distribution into account. The minimal delivery cost is determined by looping through the possible delivering times and by storing the costs related to these hours.

Figure 47 indicates that the average of the distribution should be shifted 2 hours. This means the cost will be minimized if the average hour is shifted 2 hours to the left. The new max delivery time in hours is 24 hours.

5.4 Part 4.4 Type II consumer's error

A type II error is the probability that the results indicate something is correct, even though in reality it is wrong. The error incurs a cost to the consumer and therefore is referred to as a consumer's error. In this case, a type II error occurs when the delivery time for a product from the Technology class is delivered late, but the company thinks the technology product is delivered on time.

The likelihood of making a type II error is 0.48761, thus 48, 76%

This probability is relatively high, and the company must be aware of ensuring that the product is delivered on time and not just assume the product is delivered on time.

6 PART 5: MANOVA TEST

6.1 Hypothesis test 1:

A MANOVA test was conducted on the Class Categorical Variable, against the continuous Price and Age variables.

Assuming a significance value of 0.001.

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

Table 7: Hypothesis test 1

Manova Test: test whether there is a feature that has influence

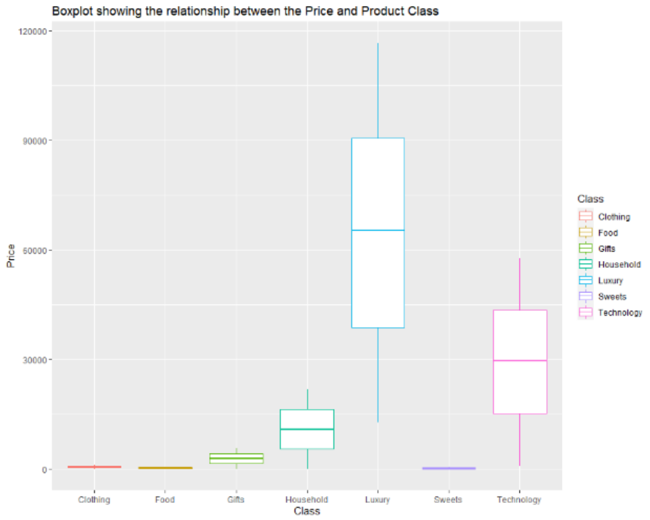
P value < 2.2e-16

Therefore, reject null hypotheses. At least one dependent variable average differs. From this output, Price or Age have a significant effect on the class of product chosen to buy.

Dependent variables	**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.

Table 8: Hypothesis test 1 MANOVA test results

Visualizing the analysis:



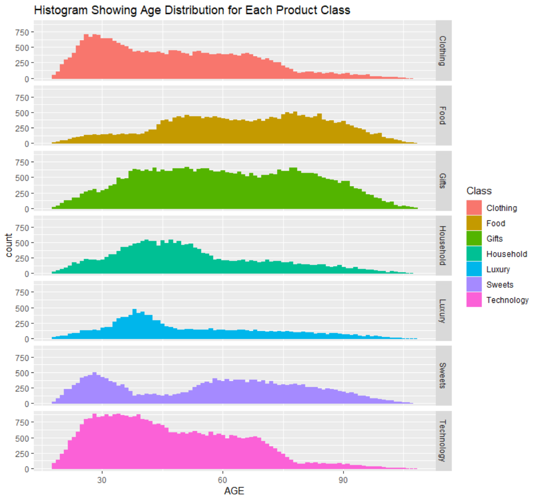
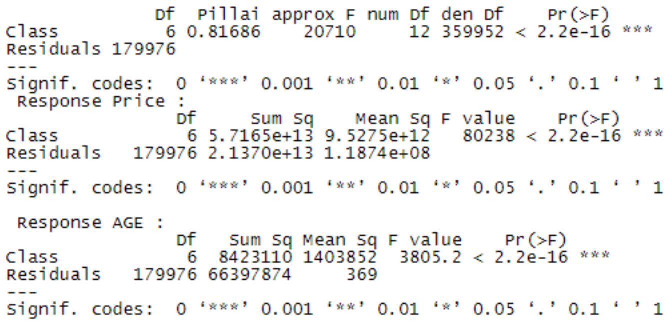


Figure 48: Box plot showing the link between age and product class

Figure 49: Histogram showing the age distribution for each product class



```
              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 50: Results of MANOVA test 1

Conclusion for MANOVA test 1:

As seen in figure 12, luxury items are most popular at the age around 40 years and are the most expensive as seen in figure 13. This tells us that age and price influence the type of product bought. The reliability and service for luxury items must be high to increase revenue and ensure customer satisfaction. Technology items are most popular below the age of 75 and most popular around the age of 35. Management can use this information to target the correct age group for marketing purposes.

6.2 Hypothesis test 2:

A MANOVA test was conducted on the Class categorical variable, against the continuous delivery time and Year variables.

Assuming a significance value of 0.001.

Dependent variables	Delivery time and Year
Independent variable	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.

Table 9: Hypothesis test 2

Manova Test: test whether there is a feature that has influence

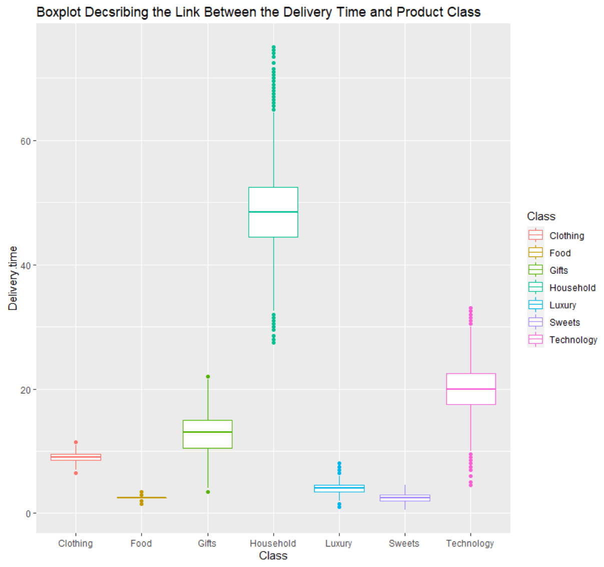
P value < 2.2e-16

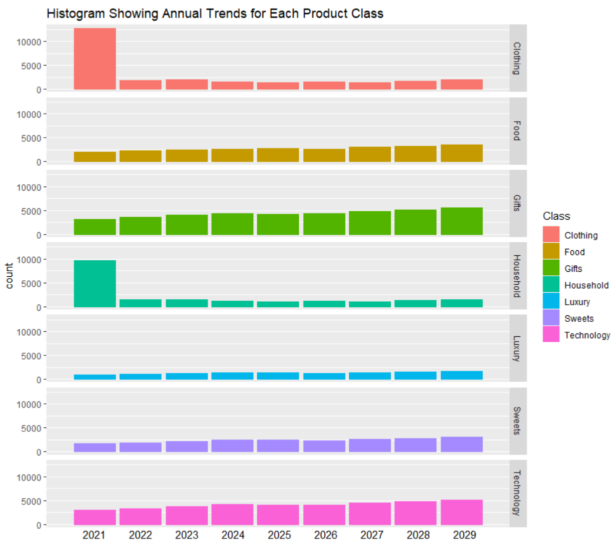
Therefore, reject null hypotheses. At least one dependent variable average differs. From this output, delivery time or year bought have a significant effect on the class of product chosen to buy.

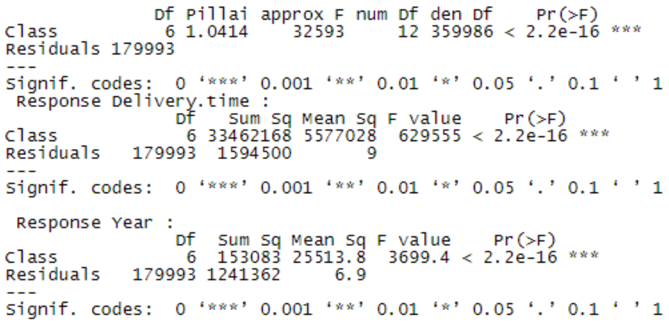
Dependent variables	**P value**	**Analyses**
Delivery time	2.2e-16	P value is smaller than 0.001. This means that the delivery time differs depending on what class the product is.
Year	2.2e-16	P value is smaller than 0.001. This means that the Year of a product being bought differs depending on what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

Table 10: hypothesis test 2 MANOVA test results

Visualizing the analysis:







```
              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 51: Results of MANOVA test 2

Conclusion for MANOVA test 2:

Household products have a long delivery time and since 2021 they a large sharp decline in the number of products being bought. Since household products have the longest delivery time by far it is clear that the class of products affects the delivery time. The reliability and service for household products needs to be improved to retain customers.

7 PART 6: RELIABILITY OF THE SERVIE AND PRODUCTS

Problem 6: A blueprint specification for the thickness of a refrigerator part at Lafrigeradora is 0.04+-0.035cm. It costs $30 to scrap a part that is outside of specifications. Determine the Taguchi loss function for this situation.

1. Calculate constants

$$L(x) = k(x - T)^2$$
$$30 = k(0.035)^2$$
$$k = \frac{30}{0.035^2} = 24489.796$$

2. Calculate loss function

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

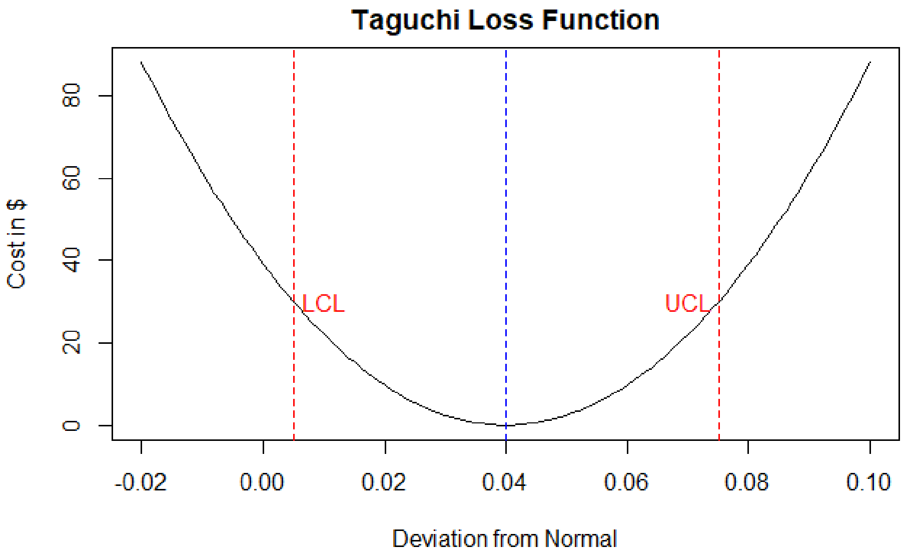


Figure 52: Taguchi loss function problem 6

As seen in figure 50, the more a product deviates from the target value of 0.04 the more expensive the product becomes. As the product becomes more expensive, it will cost the company more money eventually causing the company to make a loss. Moreover, as the product deviates from the target value, it will become more unreliable, and the service efficiency will decrease.

Problem 7: A team was formed to study the refrigerator part at Lafrigeradora, Inc. described in problem 6. While continuing to work to find the root cause of scrap, they found a way to reduce the scrap cost to $25 per part.

a. Determine the Taguchi loss function for this situation

1. Calculate constants

$$L(x) = k(x-T)^2$$
$$25 = k(0.035)^2$$
$$k = \frac{30}{0.035^2} = 20408.1633$$

2. Calculate loss function

$$L(x) = k(x-T)^2$$
$$L(x) = 20408.1633(x-0.04)^2$$

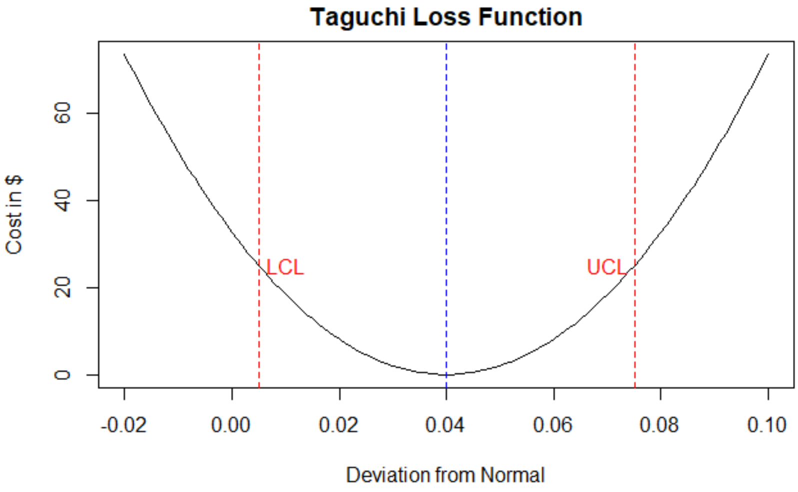


Figure 53: Taguchi loss function problem 7

As seen in figure 51, the more a product deviates from the target value of 0.04 the more expensive the product becomes. As the product becomes more expensive, it will cost the company more money eventually causing the company to make a loss. Moreover, as the product deviates from the target value, it will become more unreliable, and the service efficiency will decrease.

b. If the process deviation from the target can be reduced to 0.0027cm, what is the Taguchi loss?

$$L(0.027) = 20408.1633(0.027)^2 = \$14.878$$

Therefore, the company would make a loss of $14.878 per product if the process deviation from the target is reduced to 0.027cm.

Part 6.2

Problem 27:

a) Analyse the system reliability, assuming only one machine at each stage (all the back up machines are out of operation)

$$Reliability = reliability(Machine\ A) \times reliability(Machine\ B) \times reliability(Machine\ C)$$

$$Reliability = 0.85 \times 0.92 \times 0.90 = 0.7038$$

b) How much is the reliability improved by having two machines at each stage?

$$Reliability = reliability(Machine\ A1\ and\ A2) \times reliability(Machine\ B1\ and\ B2) \times reliability(Machine\ C1\ and\ C2)$$

$$Reliability = (1-(1-0.85)^2) \times (1-(1-0.92)^2) \times (1-(1-0.90)^2) = 0.9615$$

$$Improved\ reliability = 0.9615 - 0.7038 = 0.2577$$

Having two machines in parallel at each stage will improve the reliability by 25.77%. This is because if one machine breaks down, the identical machine in parallel can still operate. The company would greatly improve reliability by running the two machines simultaneously.

Part 6.3

1. Reliability of vehicle if the number of vehicles is 20 and is independent of the number of drivers.
   Vehicle reliability probability = 0.956732
   Number of reliable days = 349.2 days

   Reliability of drivers
   Driver reliability probability = 0.9963746
   Number of reliable days = 363.67 days   Total reliable days = **347.94 days**

   If number of vehicles increase to 21

2. Reliability of vehicle if the number of vehicles is 21 and is independent of the number of drivers.
   Vehicle reliability probability = 0.995082
   Number of reliable days = 363.2 days

   Reliability of drivers
   Driver reliability probability = 0.99988
   Number of reliable days = 364.95 days   Total reliable days = **363.16 days**
   Therefore it is recommended to increase the number of vehicles to 21 as it increase the total number of reliable days from 347 days to 363 days.

8 CONCLUSIONS

Data analytics and manipulation help businesses gain a competitive edge.

This report documents various data analytics techniques on a company’s online store “salesTable2022’ data set. The data is cleaned and sorted to ensure only valid data is used in the analysis which includes both visual and process analysis techniques.

After gaining a thorough understanding of all the data through descriptive statistics, followed by visual analysis, the sales are statistically analyzed. Thereafter control charts were constructed.

Visually it was seen that the classes Household, then technology, then gifts have the longest delivery times which is seen in figure 14. The control charts are in agreement that the classes household, technology and gifts are out of control. They all have many outliers above the upper specification limit with x bar charts that show an increasing delivery time. This is a clear indication that the company need to investigate further into an issue that is causing delivery times to increase. Technology is also the second most bought product which means it’s a vital part of the companies income. The probability of making a type I error is significantly smaller than making a type II error. The company should therefor put emphasis on ensuring products are delivered on time rather than assuming the product are delivered on time.

In conclusion, the company can take vital information from explorative analysis conducted to improve and benefit the company.

9 REFERENCES

344, Q. (2022). *QA344 Statistics.* Cape Town: THEUNIS GYSBERT Dirkse van Schalkwyk.

Lindsay, a. E. (2022). *Managing for Quality and Performance Excellence.* Cape Town: Cengage Learning.

Molaksingh, P. (2022, 09 16). *Geeksforgeeks*. Retrieved from https://www.geeksforgeeks.org/process-capability-index-cpk-formula/

Starcape. (2022, August). *Starcape*. Retrieved from http://www.starcape.com/

Zach. (2022, 10 1). *How to conduct a MANOVA in r*. Retrieved from statology: https://www.statology.org/manova-in-r/