21 October 2022

ECSA GA4

Quality Assurance 344

BADENHORST.M.MEJ [23539968@sun.ac.za]

INDUTRIAL ENGINEERING

Stellenbosch University

Table of Contents

List of Figures

List of Tables

Introduction

Data analysis and the manipulation of data has become an important skill to help businesses gain a competitive edge.

This report involves manipulating and analysing a file that contains customer information for an online store. The data file includes the different sales of various clients, when and why they were bought as well as the type of items bought. To assist the business in improving their online sales, it is necessary to evaluate the significance of these factors and provide conclusions and recommendations.

The data is provided as a raw data set and in some cases the data is not valid and must be cleaned or removed before the dataset can be utilized. A thorough summary Is created once the data has been cleaned in order to provide a better understanding of the data.

After using descriptive statistics to understand and analyse the data, statistical process control (SPC) is used to determine the best and worst performing classes. The SPC is also used to investigate in further detail why certain classes have the best or worst performance. The stability is also analysed

To determine how reliable the results obtained from the study are, the probability of type I and type II errors are examined. Additionally, delivery times are being decreased to assess what impact this would have on revenue, expenses and customer satisfaction.

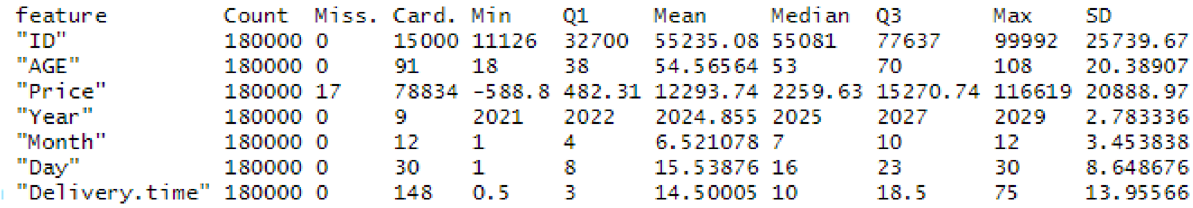
MANOVA is also used to test the hypothesis of the data included in the analysis. These MANOVA results is also discussed with appropriate graphs.

The last part of the report calculations is done on the reliability of products and services. Conclusions is also given as to why the reliability may differ in some cases.

Part 1 – Data Wrangling

Before cleaning the data, a summary was done to obtain knowledge of what mistakes the raw data has, as well as which features has these mistakes. Table 1 Is the summary of the raw data.

Table 1: Summary of Raw Data



```
feature         Count  Miss. Card. Min    Q1     Mean     Median  Q3       Max    SD
"ID"            180000 0     15000 11126  32700  55235.08 55081   77637    99992  25739.67
"AGE"           180000 0     91    18     38     54.56564 53      70       108    20.38907
"Price"         180000 17    78834 -588.8 482.31 12293.74 2259.63 15270.74 116619 20888.97
"Year"          180000 0     9     2021   2022   2024.855 2025    2027     2029   2.783336
"Month"         180000 0     12    1      4      6.521078 7       10       12     3.453838
"Day"           180000 0     30    1      8      15.53876 16      23       30     8.648676
"Delivery.time" 180000 0     148   0.5    3      14.50005 10      18.5     75     13.95566
```

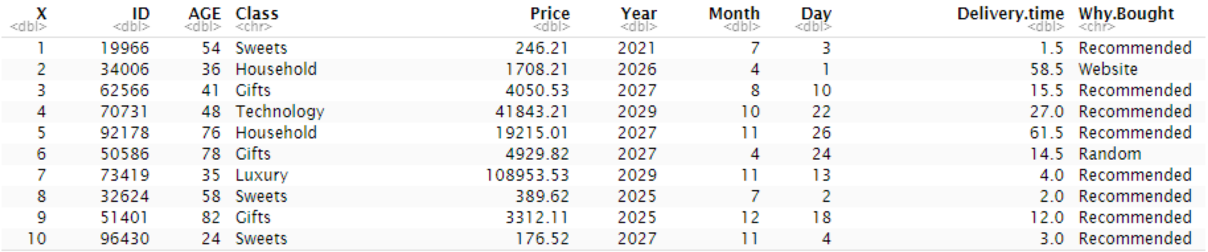
From table 1 the feature that needs to be cleaned is the price feature, because it is the only feature that contains missing values.

The raw Sales dataset was split into valid and Invalid data. The invalid data contained all the instances that had missing values as well as the instances that had negative prices.

1.1 Valid Data

The valid dataset contains 179 978 rows of data. Table 2 below shows the first rows of the Valid Dataset.

Table 2: Valid Dataset

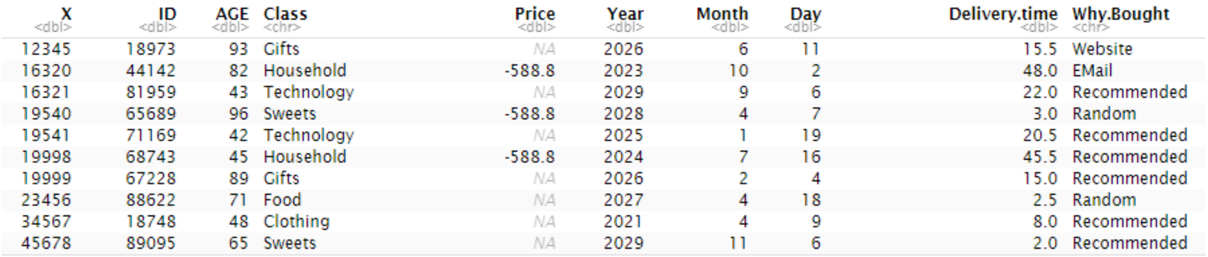


X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	19966	54	Sweets	246.21	2021	7	3	1.5	Recommended
2	34006	36	Household	1708.21	2026	4	1	58.5	Website
3	62566	41	Gifts	4050.53	2027	8	10	15.5	Recommended
4	70731	48	Technology	41843.21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

1.2 Invalid Data

The Invalid dataset contains all the rows that have negative or missing prices. There are therefore 22 rows of invalid data. Table 3 below shows the first rows of the Invalid Dataset.

Table 3: Invalid Dataset



X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
12345	18973	93	Gifts	NA	2026	6	11	15.5	Website
16320	44142	82	Household	-588.8	2023	10	2	48.0	EMail
16321	81959	43	Technology	NA	2029	9	6	22.0	Recommended
19540	65689	96	Sweets	-588.8	2028	4	7	3.0	Random
19541	71169	42	Technology	NA	2025	1	19	20.5	Recommended
19998	68743	45	Household	-588.8	2024	7	16	45.5	Recommended
19999	67228	89	Gifts	NA	2026	2	4	15.0	Recommended
23456	88622	71	Food	NA	2027	4	18	2.5	Random
34567	18748	48	Clothing	NA	2021	4	9	8.0	Recommended
45678	89095	65	Sweets	NA	2029	11	6	2.0	Recommended

Part 2 – Descriptive Statistics

Descriptive statistics will indicate what features are important as well as the correlation between features. This will help to see the effects that some features have on others.

The data is split into categorical and continuous features. The features were separately evaluated before evaluating the relationships between features.

2.1 Categorical Data

Table 4: Summary of Categorical Data

Feature	Count	Mode	Mode Frequency	Mode %	2nd Mode	2nd Mode Frequency	2nd Mode %
ID	179 978	41842	27	0.0150018	47570	26	0.0144462
Age	179 978	38	3 130	1.7391014	39	3115	1.7307671
Class	179 978	Gifts	39 149	21.7521030	Technology	36347	20.1952461
Year	179 978	2021	33 443	18.5817155	2029	22475	12.4876374
Month	179 978	12	15 225	8.4593673	10	15221	8.4571448
Day	179 978	17	6 126	3.4037493	25	6122	3.4015269
Why Bought	179 978	Recommended	106 985	59.4433764	Website	29447	16.3614442

Observations made from table 4 as well as figures obtained from the features are made below.

ID

From Table 4: There are 15 000 unique customer ID's which indicates that there are 15 000 customers responsible for the 179 978 sales. The customer with the ID 41842 buys the most from the company and is responsible for 27 of the total sales. The customer with the ID 47570 is responsible for the second most sales at 26 sales. The percentage of sales that these customers make is respectively 0.015% and 0.014%. We can therefore confidently conclude that there is no customer that will significantly influence the sales, since there is no customer that made more than 27 sales.

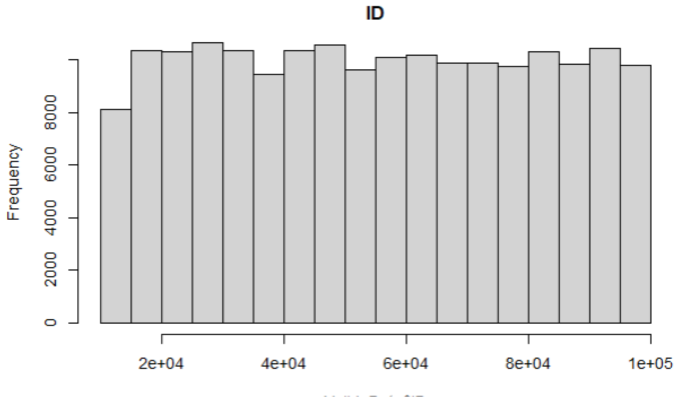


Figure 1: ID Histogram

From Figure 1: The histogram also concludes that there is not one customer that dominates the sales since the graph shows a uniform distribution of sales.

Age

From Table 4: The age Group the buys the most from the company is 38-year olds and the group that buys the second most is 39-year olds. The percentage of sales that these age groups make is respectively 1.74% and 1.73%. We can therefore confidently conclude that there is no customer that will significantly influence the sales, since the mode % and frequency remain low.

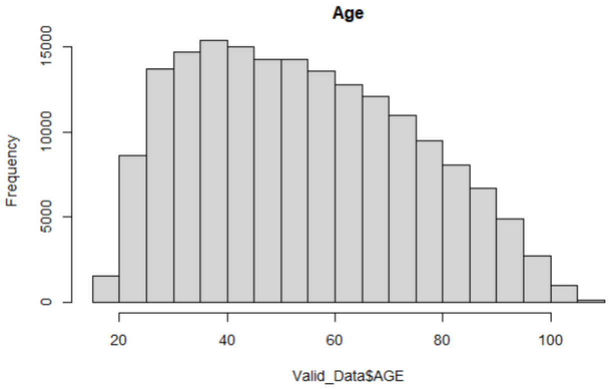


Figure 2: Age Histogram

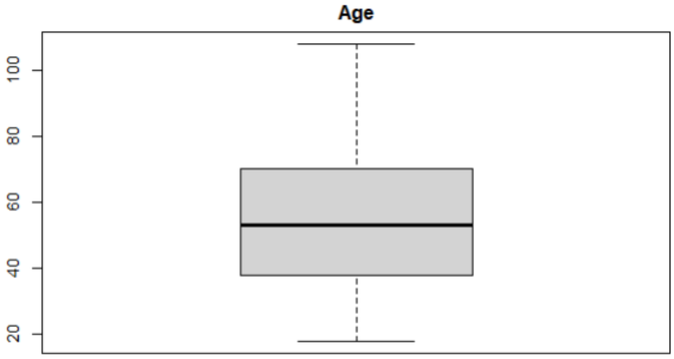


Figure 3: Age Boxplot

There are 91 different age groups that has made an online sale with this company. When referring to figure 2 and 3 the customer ages range from 18 to 108. The customers between the age of 40 and 70 contributes to the most sales and from the age of 70 the sales gradually decrease as the age increases. From the age of 18 to 25 there are only a few customers that make sales, but the sales increases after 25. The age feature has a unimodal (skewed right) distribution.

Class

From Table 4: The class that has the most sales are the Gifts class, which contributes to 21.75% of the sales. The class that has the second most sales is the Technology class, which contributes to 20.2% of the sales.

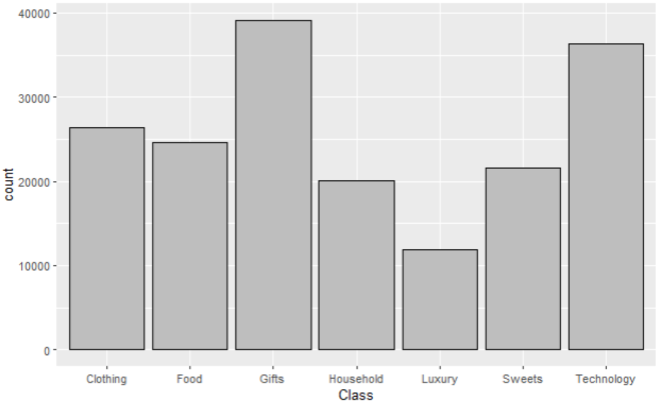


Figure 4: Class Histogram

From Figure 4: It is clear that there are 7 different classes and the ranking of the classes from most to least bought is as follows:

1. Gifts
2. Technology
3. Clothing
4. Food
5. Sweets
6. Household
7. Luxury

Year

From Table 4: The company had the most sales in 2021 (18.58% of sales) and the second most sales in 2029 (12.49% of sales).

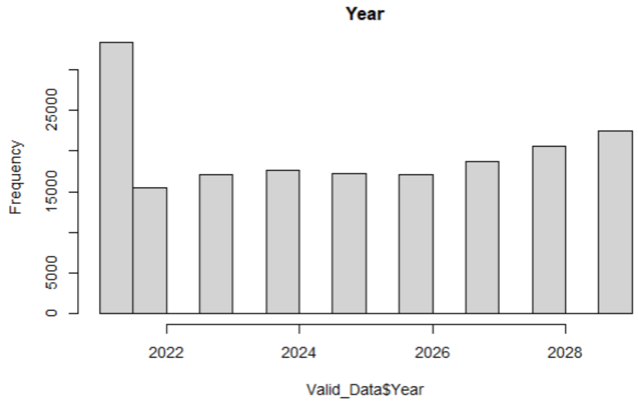


Figure 5: Year Histogram

From Figure 5: The data given ranges from the year 2021 to the year 2029. The histogram also shows a that there was a large spike in sales in 2021. This could have been a result of the coved pandemic since people were buying out of the fear of being in lockdown for a very long time. Excluding 2021 the company shows an upwards trend in sales from 2022 to 2029 which indicates that they are still growing as a business.

Month

From Table 4: The company had the most sales on the 17th day of the month (8.46% of sales) and the second most sales in October (8.46% of sales).

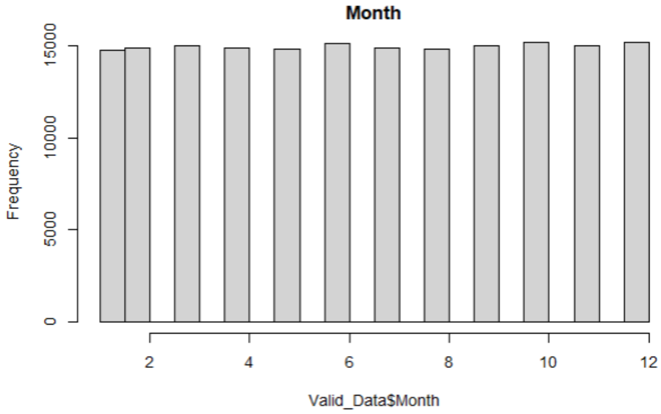


Figure 6: Month Histogram

From Figure 6: The histogram concludes that there is not one month in which there is a spike in sales, because the graph shows a uniform distribution of sales.

Day

From Table 4: The company had the most sales in December (3.4% of sales) and the second most sales on the 25th day of the month (3.4% of sales).

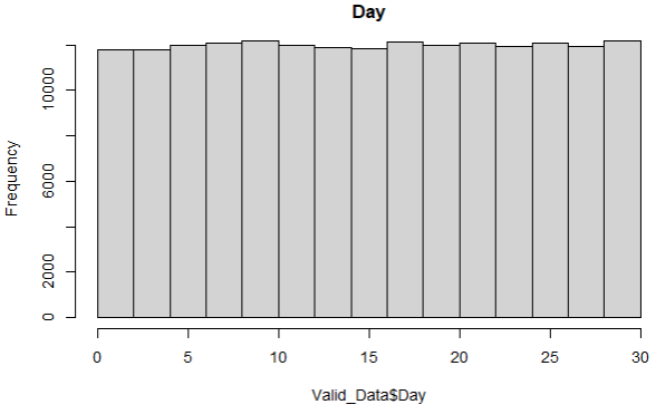


Figure 7: Day Histogram

From Figure 7: The histogram concludes that there is not one day of the month in which there is a spike in sales, because the graph shows a uniform distribution of sales. There is therefore no seasonality at a specific time of the month, which means that there aren't more sales at the beginning or end of the month.

When looking at the histograms for month and days it is clear that the company does not have peak times when they should carry more stock or have more delivery trucks since there is an even spread of sales throughout the year.

Why Bought

From Table 4: Most of the customers bought from this company because it was recommended to them. 59.4% of the sales had this reason. The second reason that customers bought from this company is because they found them on a website.

The reason why people buy products is therefore very important information for a company to have. They can use this information to improve their marketing tactics.

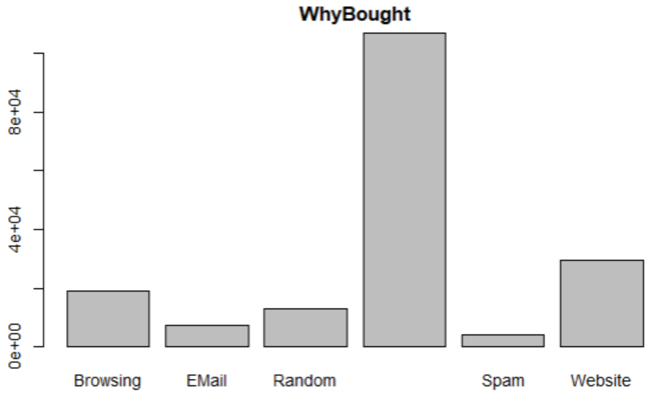


Figure 8: Why Bought Bar graph

From Figure 8: There are 6 different reasons why people bought something from this company. The ranking of the different reason ranked from most to least used is as follows:

1. Recommended
2. Website
3. Browsing
4. Random
5. Email
6. Spam

The fact the most people buy from the company because of a recommendation, indicates that the company has great customer satisfaction and loyalty. The histogram in figure 8 also indicates that the company should stop using Emails and Spam as a form of marketing, or they should improve this marketing tactic.

2.2 Continuous Data

Table 5: Summary of Continuous Data

Feature	Count	Minimum	1st quartile	Mean	Median	3rd quartile	Maximum	Std. deviation
Price	179978	35.65	482.31	2259.63	12294.098	15270.97	116618.97	20889.15
Delivery time	179978	0.5	3	10	14.5	18.5	75	13.956

Observations made from table 5 as well as figures obtained from the features are made below.

Price

From Table 5: The cheapest product was sold at a price of R35.65 and the most expensive product was sold at a price of R116 618.97. The greatest number of products sold fell between the price range of R482.31 to R 15 270.97. The most expensive product sold was a clear outlier with a standard deviation of 20 889.15.

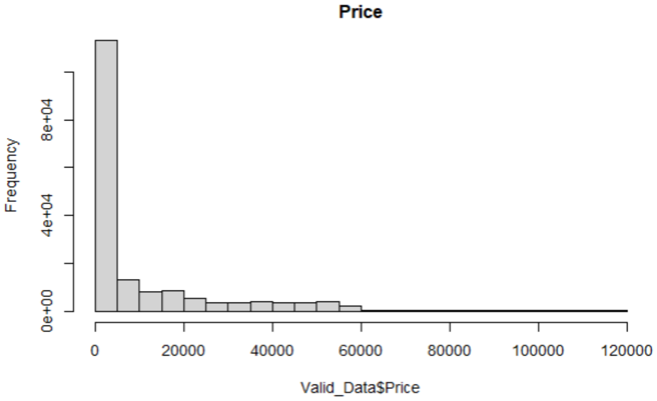


Figure 9: Price Histogram

From Figure 9: The price feature has an exponential distribution. Most of the products sold is in the lower price range, from R0 to R5000, with very little sales being made that has a price above R5000.

Delivery time

From Table 5: The shortest delivery time for a sale is 0.5 days and the longest delivery time for a sale is 75 days. Most of the sales arrive at the customer between 3 and 18.5 days and the standard deviation is 13.96. Both the maximum and minimum delivery times are therefore outliers.

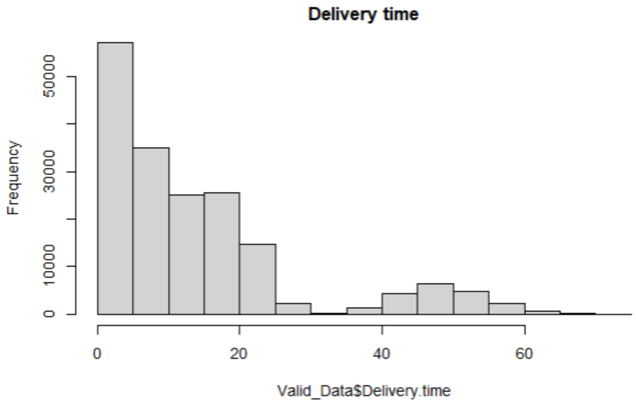


Figure 10: Delivery time Histogram

From Figure 10: The delivery time feature has a multimodal distribution. Most of the sales made will have a delivery time between 0 and 5 days following a downward trend to 25 days. a small number of sales have a delivery time between 40 and 60 days.

2.3 Feature Relationship

The graphs below compare features with one another to establish the kind of relationships between these features.

Price VS Delivery times

A Scatterplot is the best way to establish whether more expensive items have longer or shorter delivery times.

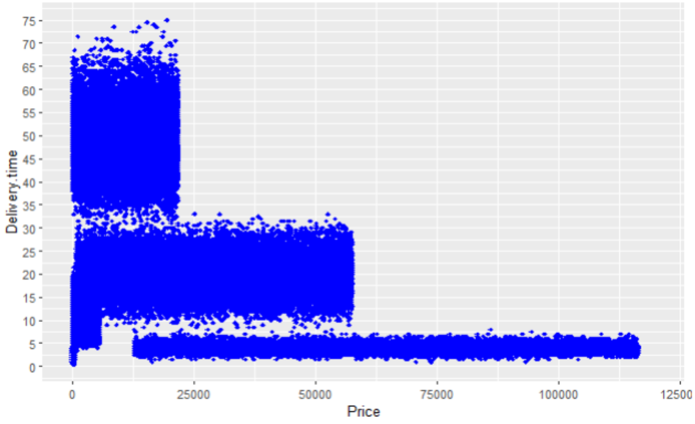


Figure 11: Price VS Delivery time - Scatterplot

From Figure 11: The scatterplot indicates that if the price is under R20 000 the delivery time range from the minimum (0.5 days) to the maximum (75 days). If the price range between R20 000 and R60 000 the delivery times are between 0 and 35 days. And as soon as the product has a price of above R60 000 the delivery times range between 0 and 10 days.

Thus, the higher priced items tend to have shorter delivery times. This can be due to the fact that when customers pay a large amount for an item, they expect it to arrive as fast as possible. This scatterplot also indicates that the lower the price, the more time it can take for the sale to arrive. Depending on if it is perishables.

Price VS Year

A Heatmap is the best way to see in which year the company sold the most of what price range.

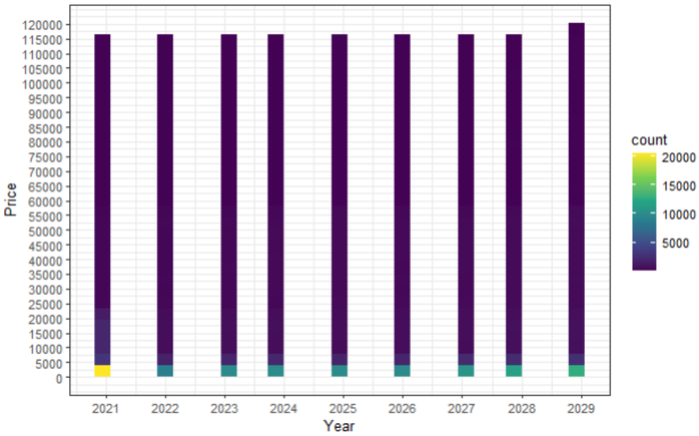


Figure 12: Price VS Year - Heatmap

From Figure 12: The heatmap clearly indicates that across all the years the greatest number of items sold were in the lower price range. The year 2021 however had a very large number of cheaper items being sold. This could have been a cause of the covid pandemic, where people were buying more cheaper items. Over all the years the company still made sales in all the price ranges.

Price VS Class

The boxplot is the best way to visualize the distribution and skewness of the prices for the different classes, by displaying the data in percentiles and averages.

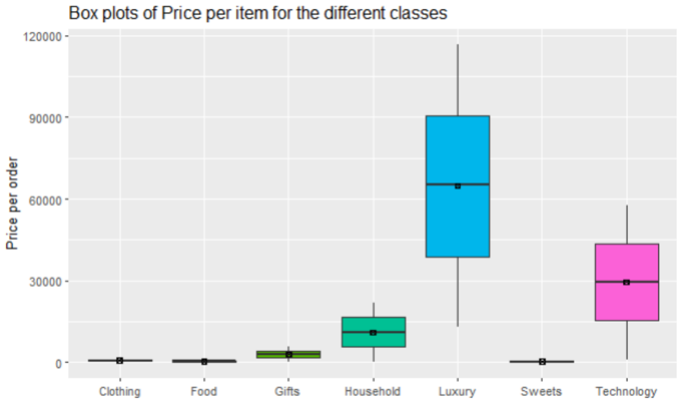


Figure 13: Price VS Class - Boxplot

From Figure 13: Luxury items sold have the widest price range with most of these items being sold at a price of between R40 000 and R90 000. They are also the most expensive items with the mean above R60 000.

Technology also has a wide range of prices and is also expensive but not as expensive as luxury items with the mean at about R30 000.

The items that belong to the classes, clothing, food and sweets, price range is very small and on the cheap side and these items won't cost more than R1000.

Table 6 below shows the exact values of what is witnessed in the boxplot in figure 13.

Table 6: Summary of Price VS Class

Class	Max Price	Min Price	Median Price	Mean Price
Clothing	1154.02	127.76	642.04	640.5253
Food	691.96	127.76	408.37	407.8153
Gifts	5774.49	172.61	2961.59	2931.8414
Household	21935.33	127.76	10960.88	11009.2738
Luxury	116618.97	12825.37	65342.14	64862.6386
Sweets	576.38	35.65	303.25	304.0704
Technology	57735.40	935.18	29653.90	29508.0626

Price VS Age

A heatmap is the best way to visualize which age group buys in which price range.

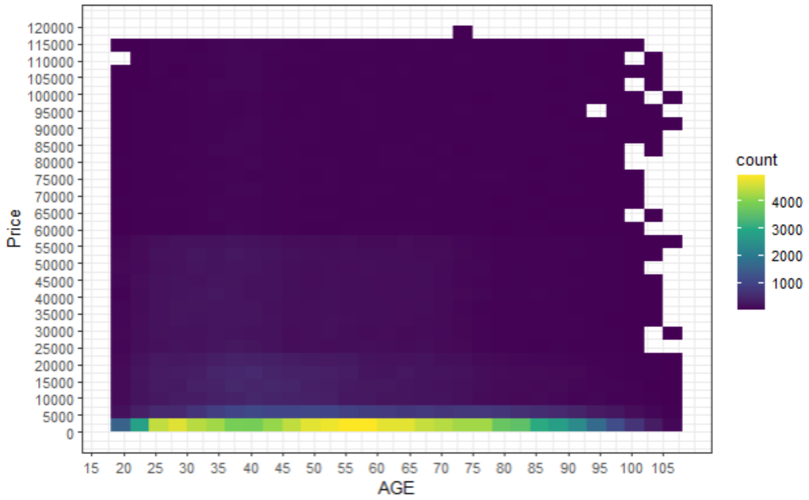


Figure 14: Price VS Age - Heatmap

From Figure 14: The heatmap illustrates that people of all ages buy products of all prices, but it seems that between the age range of 25 and 80 most people tend to buy the cheaper options, but there are still people that buy the expensive items. At a very young and very old age the same amount of people that buy cheap items buy expensive items. This can be because young people don’t know how to work with money yet and older people have enough money to buy the more expensive items.

This information can be used to sell more products to the age group, 25 to 80, that contribute to the most sales.

Age VS Class

The boxplot is the best way to visualize the distribution and skewness of the ages that purchase items in different classes, by displaying the data in percentiles and averages.

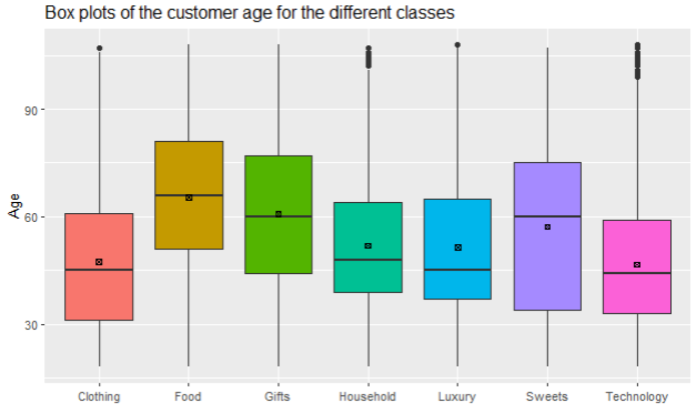


Figure 15: Age VS Class - Boxplot

From Figure 15: People between the ages of 50 and 80 tend to buy more food, sweets and gifts. People between 40 and 60 tend to buy more clothing, household and luxury items. The class bought by the largest variety of ages is sweets.

The heatmap in figure 16 is used to give more information on the relationship between the age and the class.

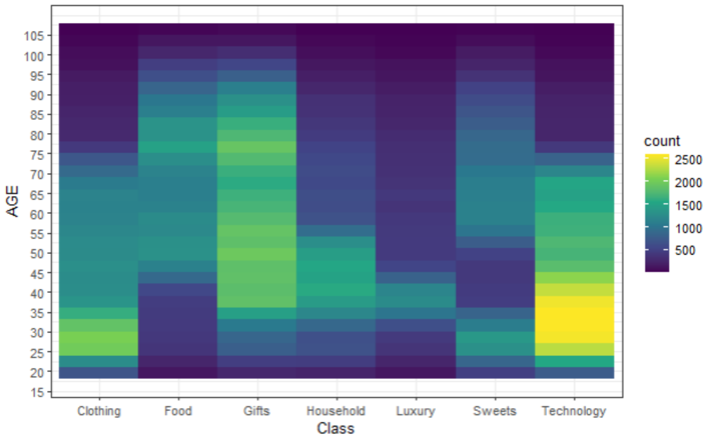


Figure 16: Age VS Class - Heatmap

From Figure 16: Between the ages of 25 and 45 people tend to buy mostly form the technology class.

The information gathered from figure 15 and figure 16 can be used to know which products to market to which age group.

Age VS Why Bought

The boxplot is the used to visualize the distribution and skewness of the ages and the reasons why items were bought, by displaying the data in percentiles and averages.

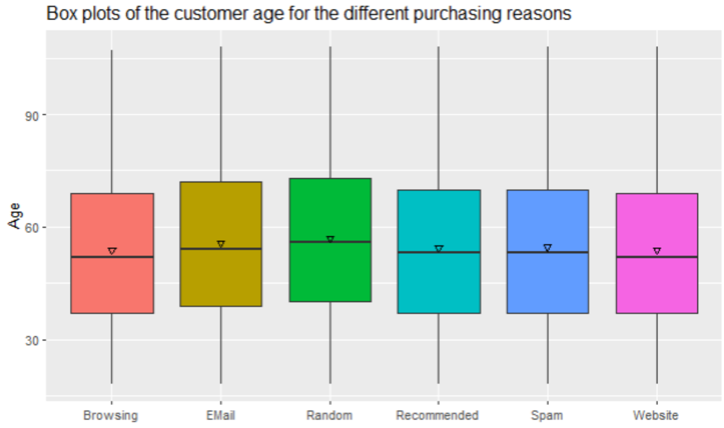


Figure 17: Age VS Why Bought - Boxplot

From Figure 17: There isn't a specific marketing tactic that can be established for different age groups as the reason certain products were bought is evenly spread for all ages.

Delivery time VS Class

The boxplot is the used to visualize the distribution and skewness of the different classes and the delivery times, by displaying the data in percentiles and averages.

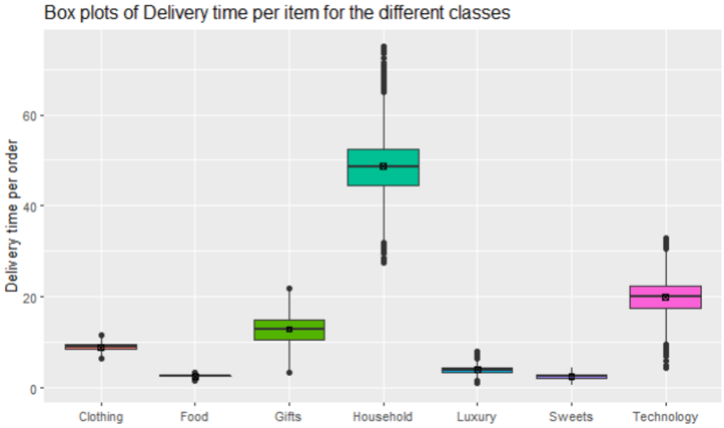


Figure 18: Delivery time VS Class

From Figure 18: Food and sweets have the lowest delivery times. This is because the items in these classes are usually perishable and therefore it needs to arrive at the customer before the expiry date. Luxury items also have short delivery times. Since people pay so much for these items, they don't want to wait long before their items arrive.

The items in the household class have the longest delivery times. This can be because these items are usually very large and bulky, and it takes more time to move them around.

Delivery time VS Year

A Heatmap is the best way to see in which year the delivery times increased and which delivery times are most frequent.

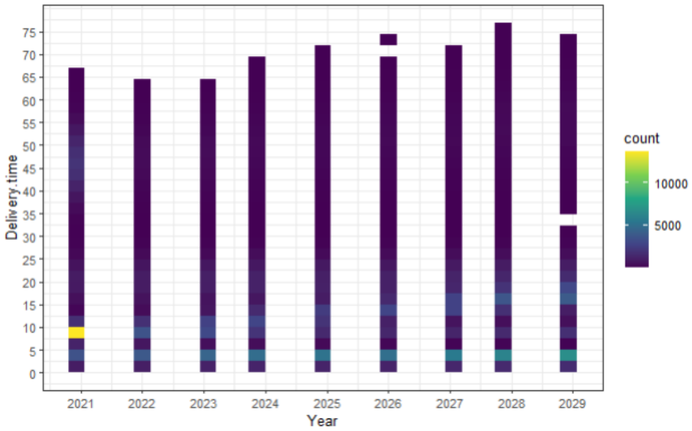
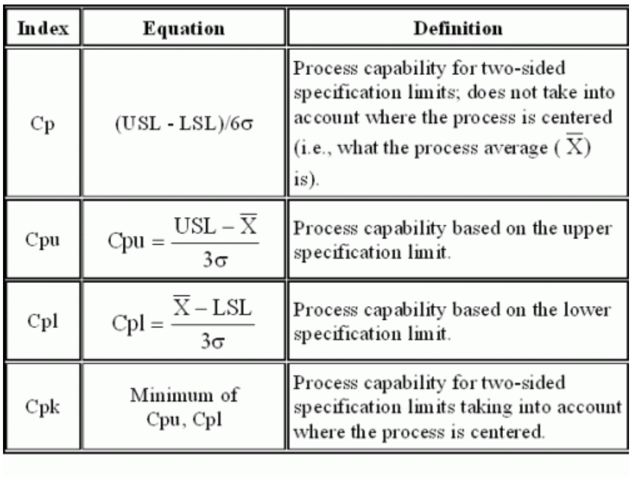


Figure 19: Delivery time VS Year - Heatmap

From Figure 19: The heatmap indicates a slight upward trend for the delivery times over the years. The delivery times increased slightly over the years. This can be caused by a decrease in workforce and that there are not enough employees to deliver the items. It can also be caused by the company receiving more orders than previous years but not increasing the workforce. The company should therefore look into increasing their workforce, because the delivery times seem to get worse as they sell more products. This should be addressed immediately, since the problem will only increase if left unattended.

2.4 Process Capabilities

Table 7: Process Capability Indices



Index	Equation	Definition
Cp	(USL - LSL)/6σ	Process capability for two-sided specification limits; does not take into account where the process is centered (i.e., what the process average ($\overline{X}$) is).
Cpu	$Cpu = \frac{USL - \overline{X}}{3\sigma}$	Process capability based on the upper specification limit.
Cpl	$Cpl = \frac{\overline{X} - LSL}{3\sigma}$	Process capability based on the lower specification limit.
Cpk	Minimum of Cpu, Cpl	Process capability for two-sided specification limits taking into account where the process is centered.

The process capabilities for the technology and delivery time features are calculated using table 7 above. The Upper Specification limit (USL) is chosen to be 24 hours and the Lower Specification Limit (LSL) is chosen to be zero. This is a logistical choice since it is ideal for products to be delivered immediately. It also can't be less than zero.

Table 8: Summary of Process Capabilities

Sigma	**Mean**	**cp**	**cpl**	**cpu**	**cpk**
3.5019927426298	20.0109500096294	1.142207	1.90472	0.3796933	0.3796933

The Cp is larger than one. We can therefore conclude that this process is capable of meeting the specifications.

The cpl indicates whether the process can meet the Lower Specification Limit (LSL). The cpl of 1.90472 is larger than one. It therefore meets the Lower Specification Limit.

The cpu indicates whether the process can meet the Upper Specification Limit (USL). The cpu of 0.3797 is smaller than one. It therefore does not meet the Upper Specification Limit.

The mean is smaller than the UCL and this indicates that the process is centred with a cpk of 0.3797. The cpk measures the process capability more accurately than the Cp alone, because it also takes whether the process is centred into account. A cpk is therefore desired to indicate whether the process is centred

Part 3 – Statistical Process Control

Statistical Process Control (SPC) is the use of statistical techniques to control a process or production method. SPC tools can help to monitor a process' behaviour, discover issues and find solutions. It is also used to measure the quality of a process during manufacturing. It is thus very useful, as it will help to identify problematic processes.

The first step was to order the sales data chronologically. This is necessary since the data is real time data.

The first 30 samples of size 15 was used to develop control charts as seen in 3.1. These SPC values are critical values and will be used in 3.2 and 4.1 to control the process and it will be used to identify out of control processes.

3.1 First 30 samples

The Xbar-chart is used to examine the process mean over the time. It can be used to determine whether the process is stable and predictable.

Table 9: X-Chart

X-Chart							
Class	**UCL**	**U2Sigma**	**U1Sigma**	**CL**	**L1Sigma**	**L2Sigma**	**LCL**
Technology	22.974616	22.107892	21.241168	20.37444	19.507721	18.640997	17.774273
Clothing	9.404934	9.259956	9.114978	8.97	8.825022	8.680044	8.535066
Household	50.248328	49.019626	47.790924	46.56222	45.333520	44.104818	42.876117
Luxury	5.493965	5.241162	4.988359	4.735556	4.482752	4.229949	3.977146
Food	2.709458	2.636305	2.563153	2.49	2.416847	2.343695	2.270542
Gifts	9.488565	9.112747	8.736929	8.361111	7.985293	7.609475	7.233658
Sweets	2.897042	2.757287	2.617532	2.477778	2.338023	2.198269	2.058514

The S-chart is used to examine the standard deviation over the time.

Table 10: S-Chart

S-Chart							
Class	**UCL**	**U2Sigma**	**U1Sigma**	**CL**	**L1Sigma**	**L2Sigma**	**LCL**
Technology	5.1805697	4.5522224	3.9238751	3.2955278	2.6671805	2.0388332	1.4104859
Clothing	0.8665596	0.7614552	0.6563509	0.5512465	0.4461422	0.3410379	0.2359335
Household	7.3441801	6.4534101	5.5626402	4.6718703	3.7811003	2.8903304	1.9995605
Luxury	1.5110518	1.3277775	1.1445032	0.9612289	0.7779546	0.5946803	0.4114060
Food	0.4372466	0.3842133	0.3311800	0.2781467	0.2251134	0.1720801	0.1190468
Gifts	2.2463333	1.9738773	1.7014213	1.4289652	1.1565092	0.8840532	0.6115971
Sweets	0.8353391	0.7340215	0.6327039	0.5313862	0.4300686	0.3287509	0.2274333

The values that were out of control within the first 30 samples was not used to plot the SPC values, because those values can create the wrong SPC values to control the process.

However, removing these values can create the illusion that the process is in control.

Before interpreting the Xbar-chart, the S-Chart must first be examined to determine if the process variations are in control. If the S-Chart is not in control, then the control limits on the Xbar-Chart are not accurate. (support.minitab.com, n.d.)

Technology

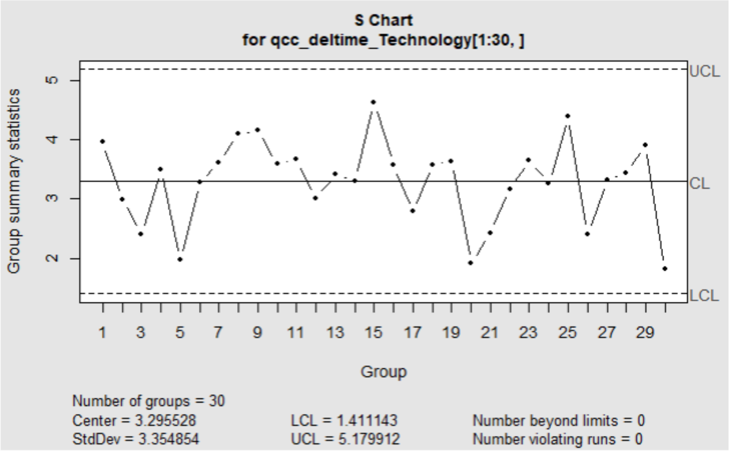


Figure 20: Technology - S-Chart

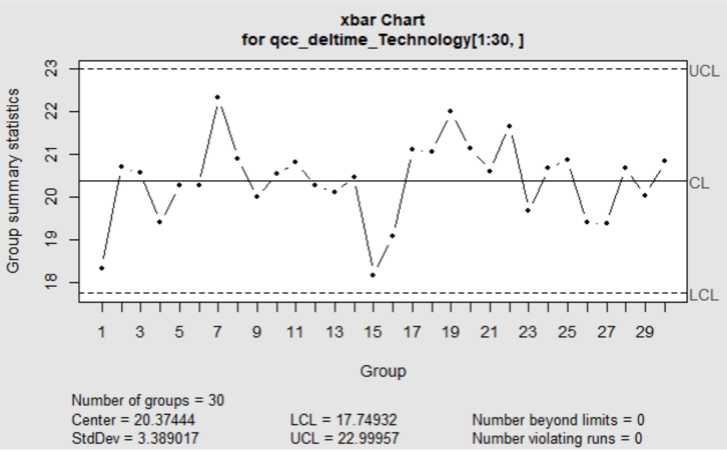


Figure 21: Technology - Xbar-Chart

From figure 20 the S-chart is in control, since all the variations fall between the limits. Therefore, the Xbar-chart's control limits are accurate.

The Xbar-chart in figure 21 has a lot of variation in the runs. This is because of the high standard deviation seen the S-chart. The Xbar-chart is still in control event though it has a lot of variation, because all the runs for both charts fall between the limits

Clothing

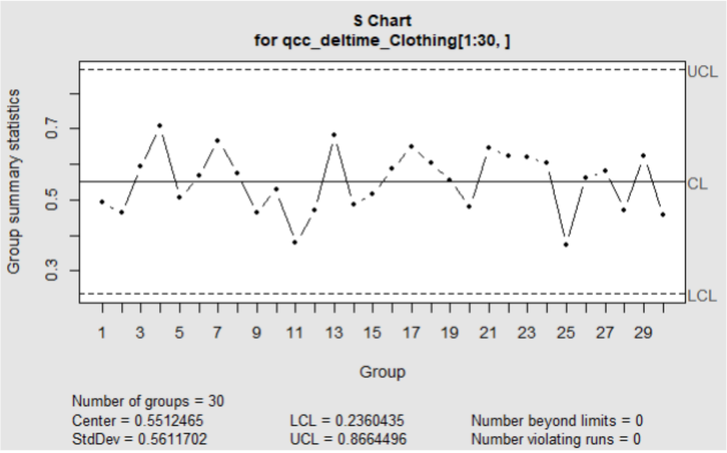


Figure 22: Clothing - S-Chart

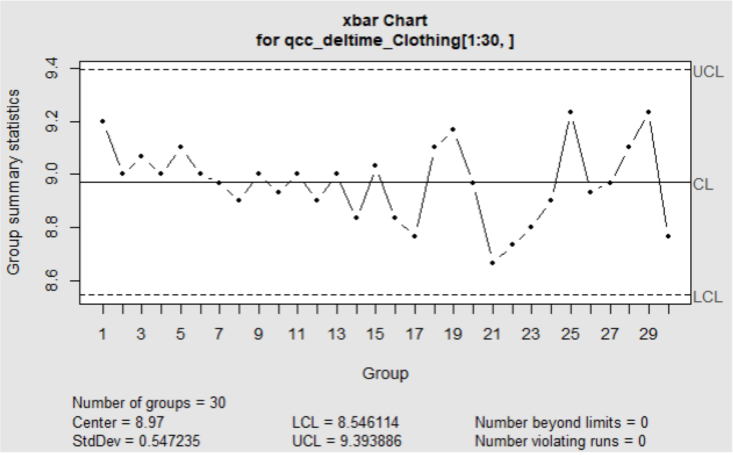


Figure 23: Clothing - Xbar-Chart

From figure 22 the S-chart is in control, since all the variations fall between the limits. Therefore, the Xbar-chart's control limits are accurate.

The Xbar-chart in figure 23 does not have a lot of variation. All the samples are situated close to the centreline, with just a couple of outliers. The Xbar-chart is therefore relatively stable. This can be explained by the small standard deviation in the S-chart, and all the points on the S-chart is also close to the centreline.

Household

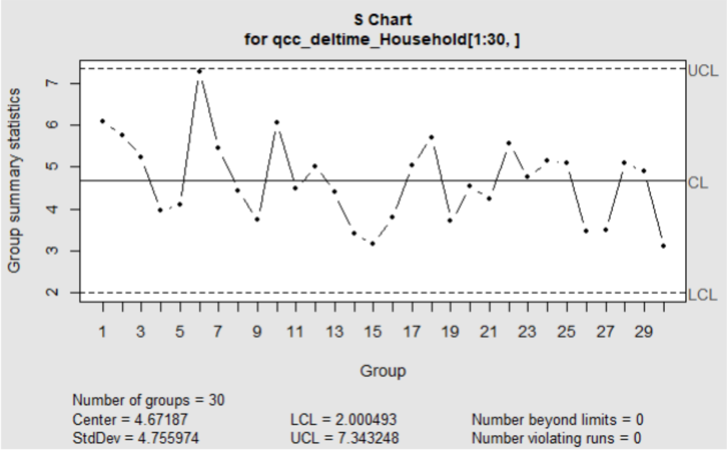


Figure 24: Household - S-Chart

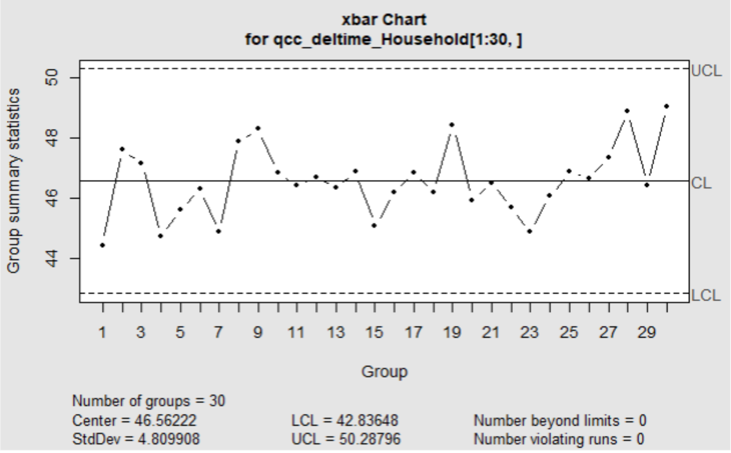


Figure 25: Household - Xbar-Chart

From figure 24 the S-chart is in control, since all the variations fall between the limits. Therefore, the Xbar-chart's control limits are accurate.

The Xbar-chart in figure 25 Has a lot of variation, but most of the points are still situated near the centreline. The large variation can be explained by the high standard deviation seen in the S-chart. The S-chart also has a lot of variation, but all the samples still fall between the limits. Both charts are therefore in control.

Luxury

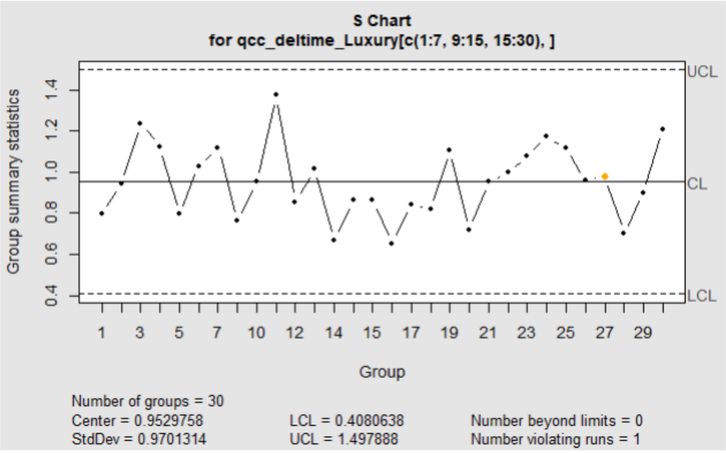


Figure 26: Luxury - S-Chart

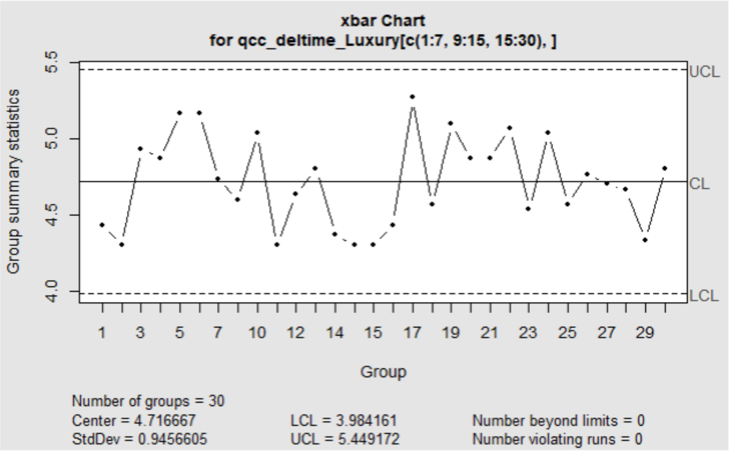


Figure 27: Luxury - Xbar-Chart

From figure 26 the S-chart is in control, since all the variations fall between the limits. Therefore, the X-chart's control limits are accurate.

The Xbar-chart in figure 27 does not have a lot of variation. This is explained by the small standard deviation in the S-chart. Some of the samples are near the centreline and some are near the limits. The S-chart has one violating run. Both charts are in control.

Food

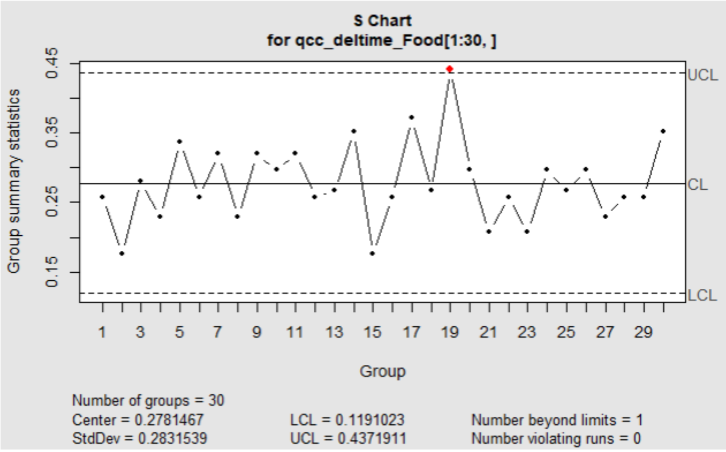


Figure 28: Food - S-Chart

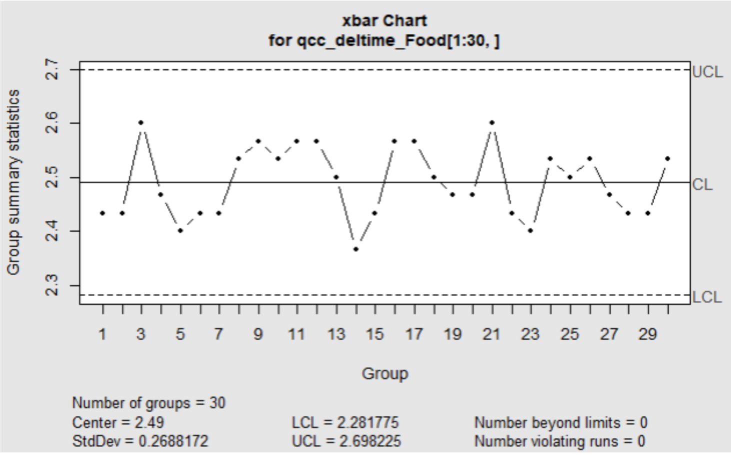


Figure 29: Food - Xbar-Chart

From figure 28 the S-chart is not in control, since there is one variation that fall beyond the limits. Therefore, the X-chart's control limits are not accurate.

The Xbar-chart in figure 29 Does not show a lot of variation. This can be explained by the small standard deviation in the S-chart. The S-chart is not in control due to a sample being beyond the limits. The X-chart on the other hand is in control and stable, since all the samples are located near the centreline.

Gifts

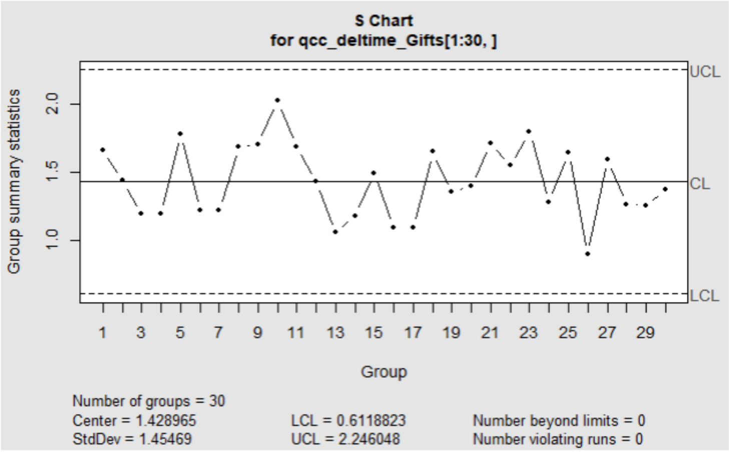


Figure 30: Gifts - S-Chart

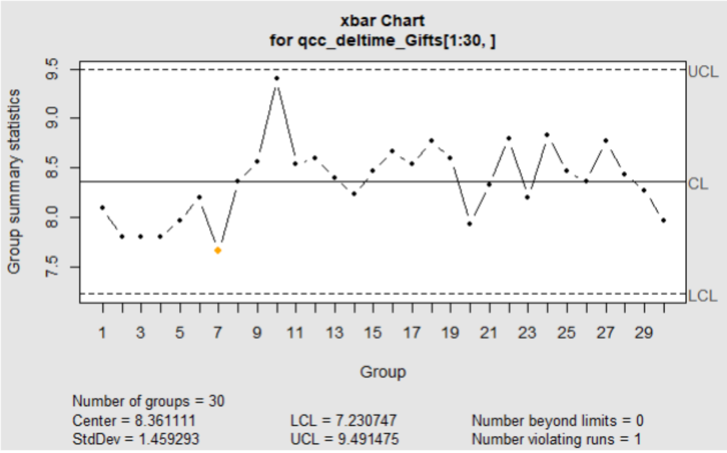


Figure 31: Gifts – Xbar-Chart

From figure 30 the S-chart is in control, since all the variations fall between the limits. Therefore, the X-chart's control limits are accurate.

The Xbar-chart in figure 31 Shows variation and some of the samples are near the centreline while others are situated near the limits. The Xbar-chart also has one violating run. The S-chart also has variation with some samples near the centreline and others near the limits. Both the charts are still in control even though they have variation.

Sweets

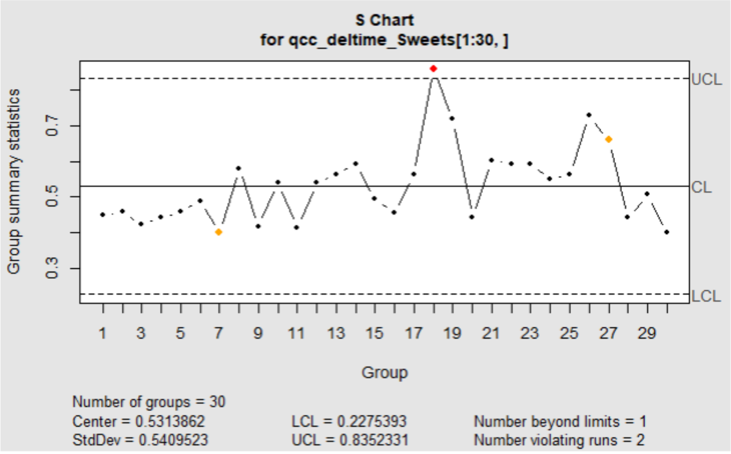


Figure 32: Sweets - S-Chart

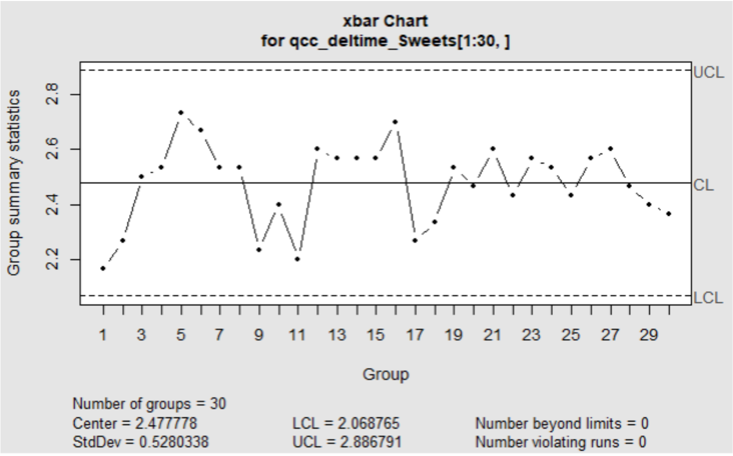


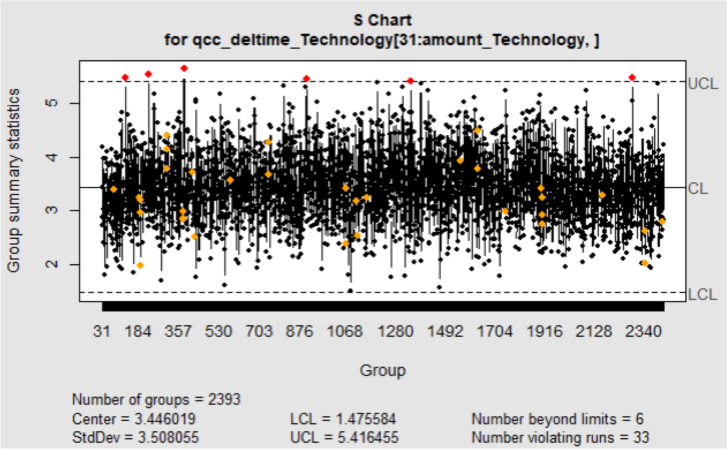
Figure 33: Sweets - Xbar-Chart

From figure 32 the S-chart is not in control, since there is one variation that fall beyond the limits. Therefore, the X-chart's control limits are not accurate.

The Xbar-chart in figure 33 Has a lot of variation with most of the points laying closer to the limits than the centreline. But the chart is still in control and stable which is verified by the small standard deviation of the S-chart. Most of the samples on the S-chart are situated near the centreline except for one that is beyond the limits. The S-chart also has 2 violating runs, but this does not make the Xbar-chart less in control or less stable.

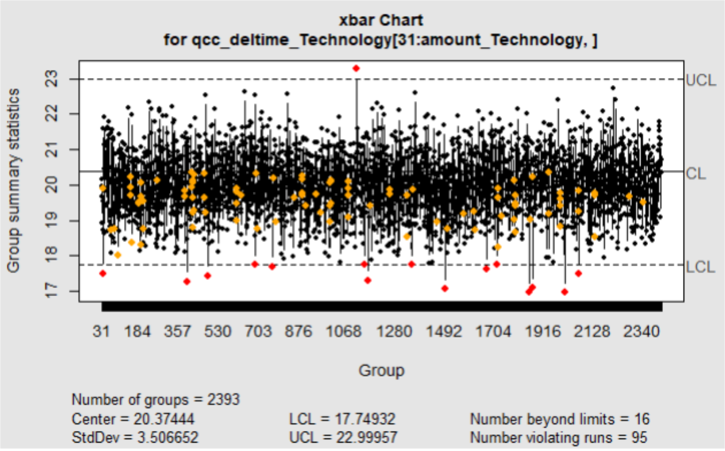
3.2 Control the process for sample 31 and onwards

Technology



Number of groups = 2393
Center = 3.446019
StdDev = 3.508055
LCL = 1.475584
UCL = 5.416455
Number beyond limits = 6
Number violating runs = 33

Figure 34: Technology (31 onwards) - S-Chart

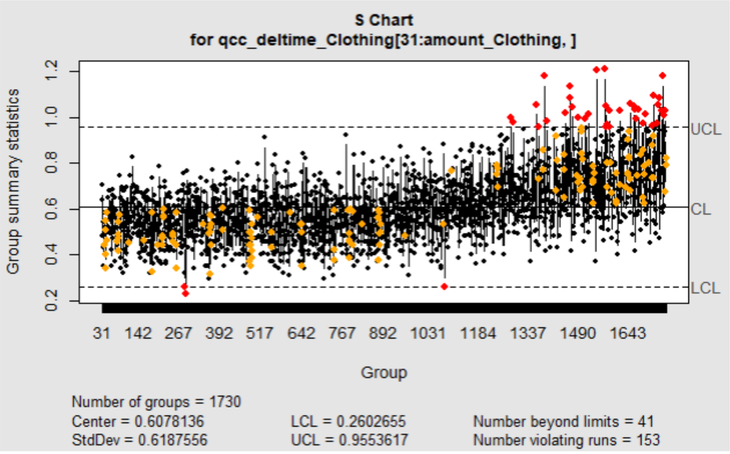


Number of groups = 2393
Center = 20.37444
StdDev = 3.506652
LCL = 17.74932
UCL = 22.99957
Number beyond limits = 16
Number violating runs = 95

Figure 35: Technology (31 onwards) - Xbar-Chart

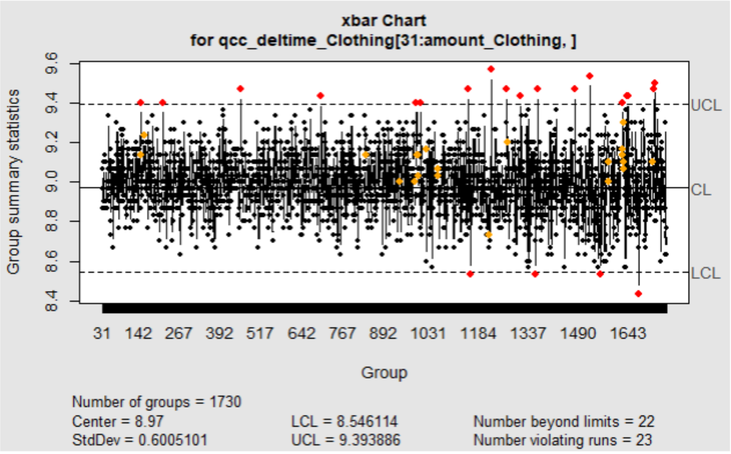
The S-Chart in figure 34 has 6 samples beyond the limits and 33 violating runs. The large standard deviation is an indication that the process is unstable. The Xbar-Chart in figure 35 has 16 samples beyond the limits and 95 violating runs. This is an indication that the process is in control but not completely. There are therefore minor problems with the delivery of technology.

Clothing



Number of groups = 1730
Center = 0.6078136
StdDev = 0.6187556
LCL = 0.2602655
UCL = 0.9553617
Number beyond limits = 41
Number violating runs = 153

Figure 36: Clothing (31onwards) - S-Chart



Number of groups = 1730
Center = 8.97
StdDev = 0.6005101
LCL = 8.546114
UCL = 9.393886
Number beyond limits = 22
Number violating runs = 23

Figure 37: Clothing (31onwards) - Xbar-Chart

The S-Chart in figure 36 has 41 samples beyond the limits and 153 violating runs. The small standard deviation seen in the S-chart is an indication that the process is stable. The Xbar-Chart in figure 37 has 22 samples beyond the limits and 23 violating runs. This can be an indication that the process is not totally in control. There are therefore minor problems with the delivery of clothing.

Household

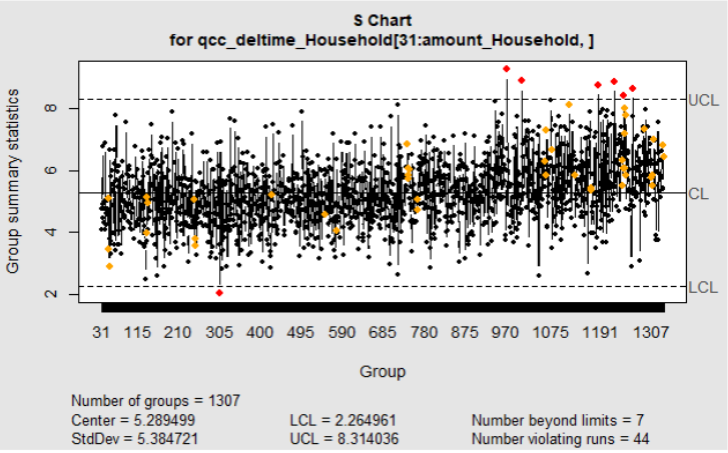


Figure 38: Household (31 onwards) - S-Chart

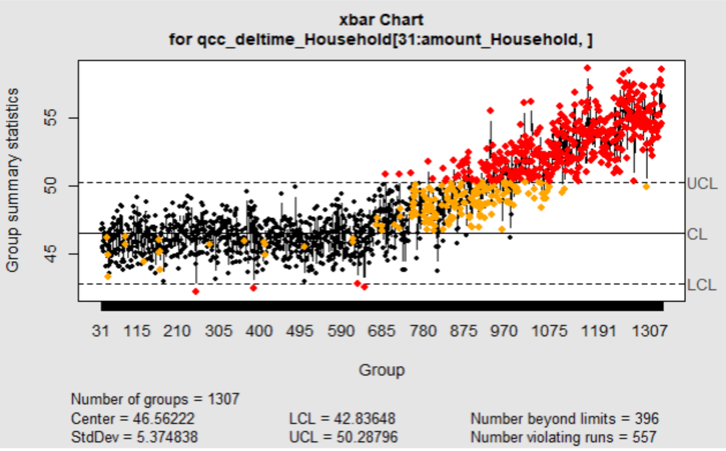


Figure 39: Household (31 onwards) - Xbar-Chart

The S-Chart in figure 38 has 7 samples beyond the limits and 44 violating runs. The Xbar-Chart in figure 39 shows an increase in the process average as time passed. This is indicated by the consecutive number of samples that lay above the UCL. The process is thus out of control and there are major problems with the delivery of the household products.

Luxury

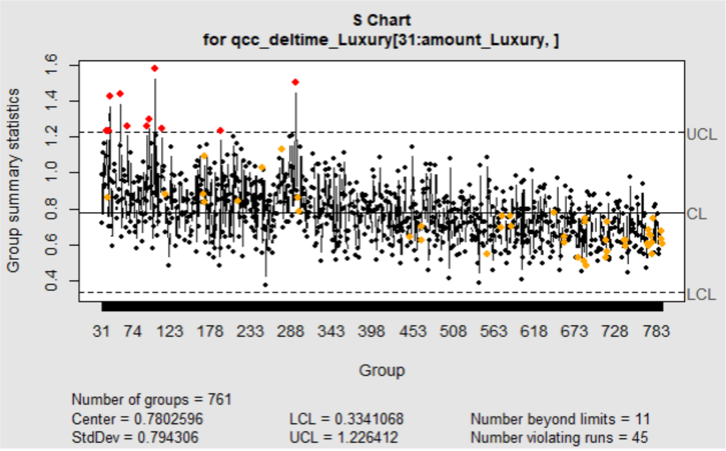


Figure 40: Luxury (31 onwards) - S-Chart

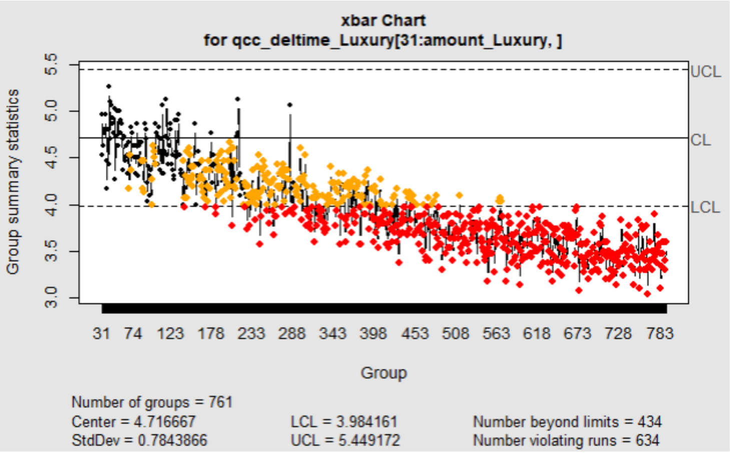


Figure 41: Luxury (31 onwards) - Xbar-Chart

The S-Chart in figure 40 has 11 samples beyond the limits and 45 violating runs. The small Standard deviation seen in the S-chart is an indication that the process is stable. The Xbar-Chart in figure 41 shows a decrease in the process average as time passed. This is indicated by the consecutive number of samples that lay below the LCL. The process is thus out of control and there are major problems with the delivery of luxury products.

Food

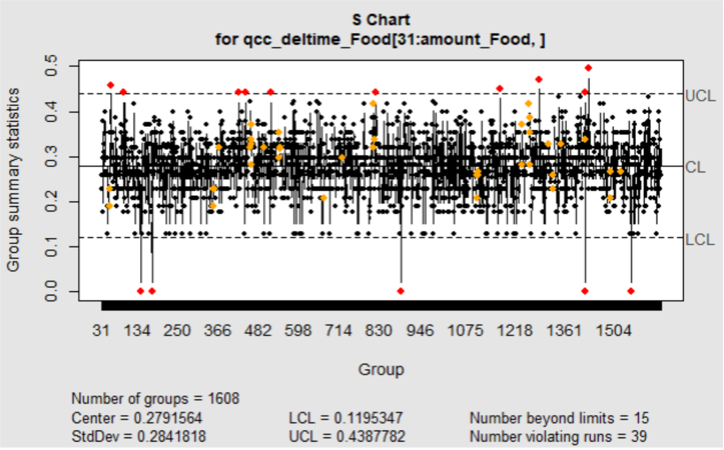


Figure 42: Food (31 onwards) - S-Chart

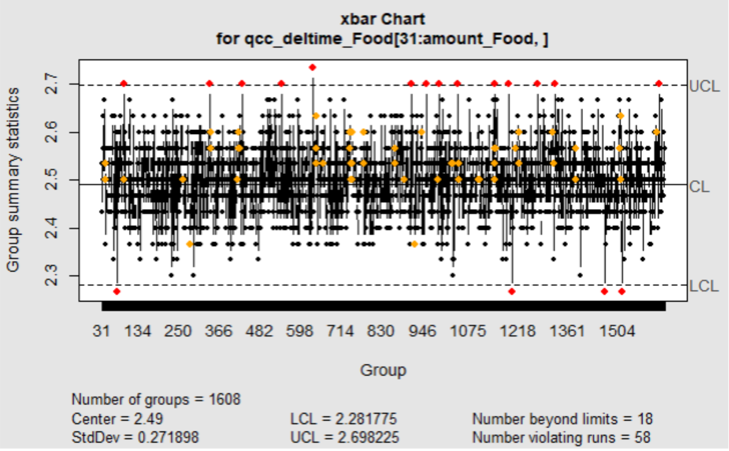


Figure 43: Food (31 onwards) - Xbar-Chart

The Xbar-chart in figure 43 does not have a significant number of outliers but it still has a lot. The process is therefore not in complete control. The process however has a small standard deviation, indicated in figure 42 (the S-chart), which means that the process is stable. There is therefore only minor problems with the delivery of food.

Gifts

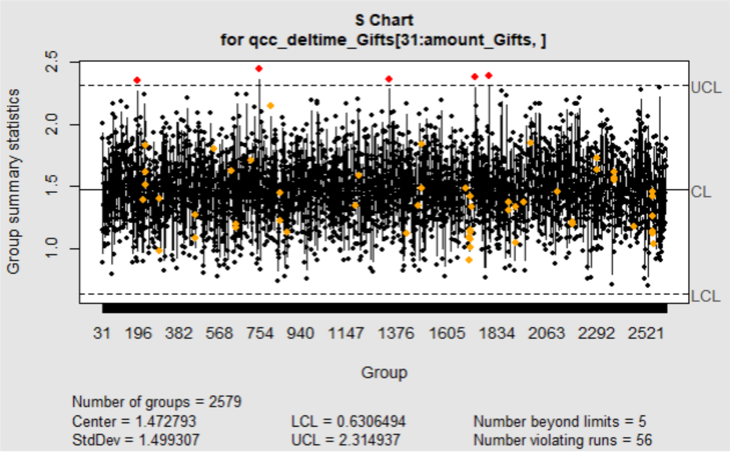


Figure 44: Gifts (31 onwards) - S-Chart

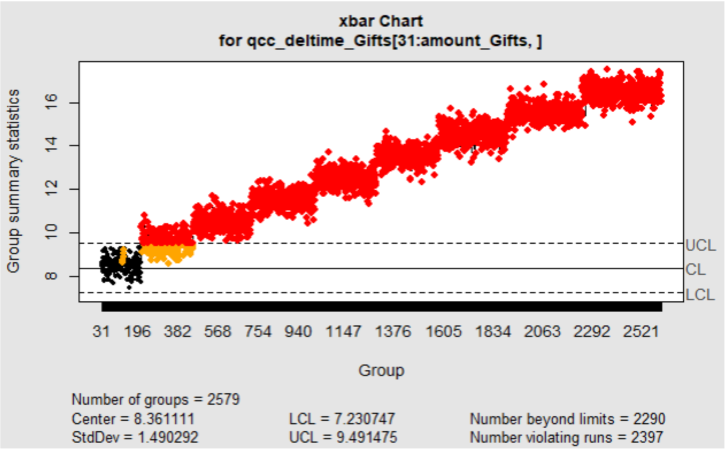


Figure 45: Gifts (31 onwards) - Xbar-Chart

The S-Chart in figure 44 has 5 samples beyond the limits and 56 violating runs. The Xbar-Chart in figure 45 shows an increase in the process average as time passed. This is indicated by the consecutive number of samples that lay above the UCL. The process is thus out of control and there are major problems with the delivery of gifts.

Sweets

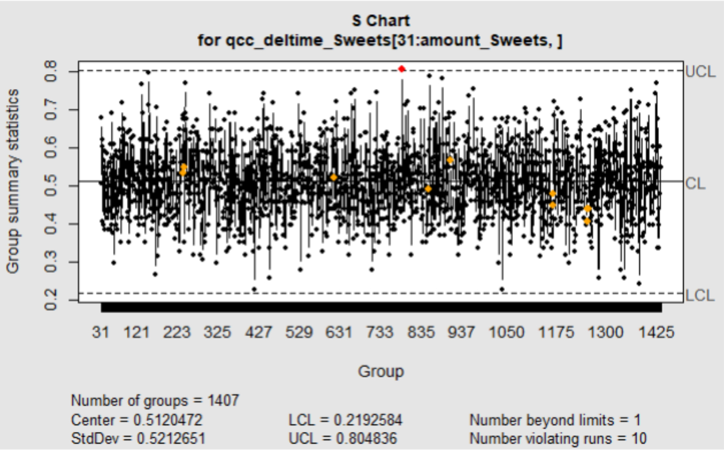


Figure 46: Sweets (31 onwards) - S-Chart

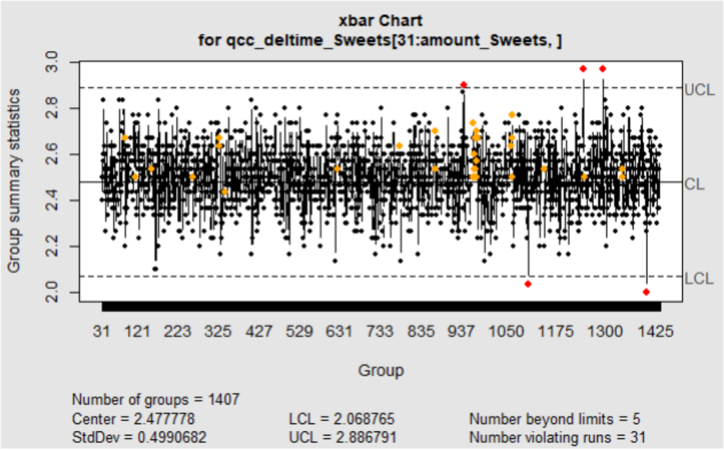


Figure 47: Sweets (31 onwards) - Xbar-Chart

The Xbar-Chart in figure 47 shows that the sweets data is relatively in control with only a couple of outliers. This indicates that there are very minor problems with the delivery of sweets. The small standard deviation in figure 46 (the S-chart) also indicates that the process is stable.

Part 4 – Optimizing the Delivery process

4.1 Processing and inspection

A – X-bar samples outside the outer control limits

Table 11: Outliers

Class	Total found	1st	2nd	3rd	3rd Last	2nd Last	Last
Technology	16	37	398	483	1872	2009	2071
Clothing	22	148	217	455	1677	1723	1724
Household	396	252	387	629	1335	1336	1337
Luxury	434	142	171	184	789	790	791
Food	18	75	93	338	1467	1515	1621
Gifts	2290	213	216	218	2607	2608	2609
Sweets	5	942	1104	1243	1294	1403	-

As seen in table 11 Gifts, Luxury and Household items have very large number of outliers. This indicates excessive variation in the process and it also means that the process average has shifted. This is also seen when looking at figure 45, 41 and 39 in part 3.2.

To improve these inconsistencies, the company should look at identifying the reasons for the changes.

B – S-bar samples between -0.3 and +0.4 sigma control limits

Table 12: Between -0.3 and +0.4 sigma limits

Class	Max between sigma lengths	Last sample position of first	Last sample position of last
Technology	0	0	0
Clothing	191	701	701
Household	5	14	14
Luxury	11	28	28
Food	1638	1638	1638
Gifts	10	75	75
Sweets	163	164	164

Table 12 can be used to show instances of where the data was completely in control and when it occurred. But it is better to look at the full picture instead of making conclusions from only the longest number of samples between the limits. This can easily create a false perception of the data being in control when it’s not.

4.2 Probability of making a Type I error

A Type I error means rejecting the null hypothesis when it's actually true, while a Type II error mean failing to reject the null hypothesis when it's actually false. (Scribbr, 2021)

Table 13: Type I and Type II errors

	Process is fine	**Process is not fine**
SPC indicated the process is not fine	Type I Error OR Manufacturer's Error	Correct to fix process
SPC indicated the process is fine	Correct to do nothing	Type II Error OR Consumer's Error

A – X-bar samples outside the outer control limits

Table 14 shows the probability of a sample being outside the UCL and LCL.

Table 14: Outer limits - Type I probability

Class	**Total Found**	**Probability of Type I Error**
Technology	16	$7.96700988024311 \times 10^{-42}$
Clothing	22	$3.08518431079213 \times 10^{-57}$
Household	396	0
Luxury	434	0
Food	18	$5.80707286218083 \times 10^{-47}$
Gifts	2290	0
Sweets	5	$1.43434888013109 \times 10^{-13}$

B – S-bar samples between -0.3 and +0.4 sigma control limits

Table 15 shows the probability of a sample being between -0.3 and +0.4 sigma control limits.

Table 15: Between -0.3 and +0.4 - Type I probability

Class	**Max above centreline lengths**	**Probability of Type I Error**
Technology	0	1
Clothing	701	$6.172712 \times 10^{-117}$
Household	14	4.776668×10^{-3}
Luxury	28	2.281656×10^{-5}
Food	1638	$2.868346 \times 10^{-272}$
Gifts	75	3.687579×10^{-13}
Sweets	164	6.495428×10^{-28}

4.3 Optimize delivery process

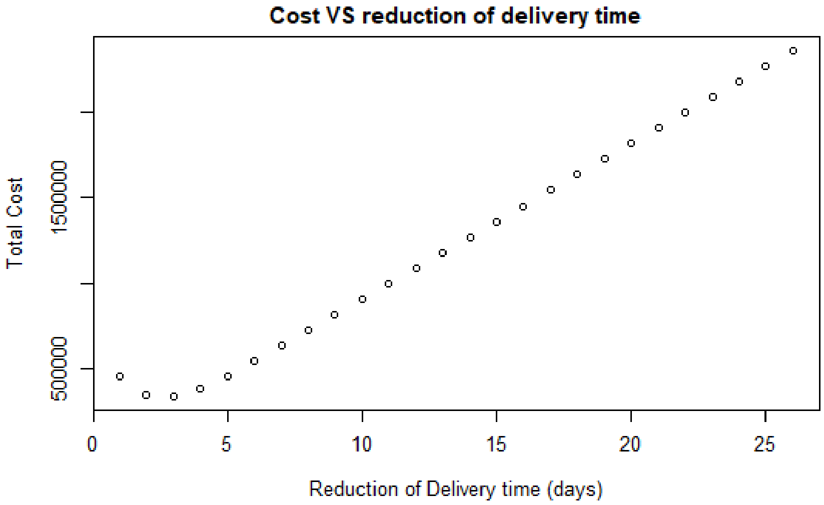


Figure 48:Cost VS Reduction of delivery time

As seen in figure 48 the cost reduces until about a 3 day reduction in delivery time, and then the total cost increases significantly when reducing the delivery time.

4.4 Probability of making a Type II Error

A Type II Error is an error where the SPC indicated that the process is fine, while it was faulty. This will lead to unsatisfied customers.

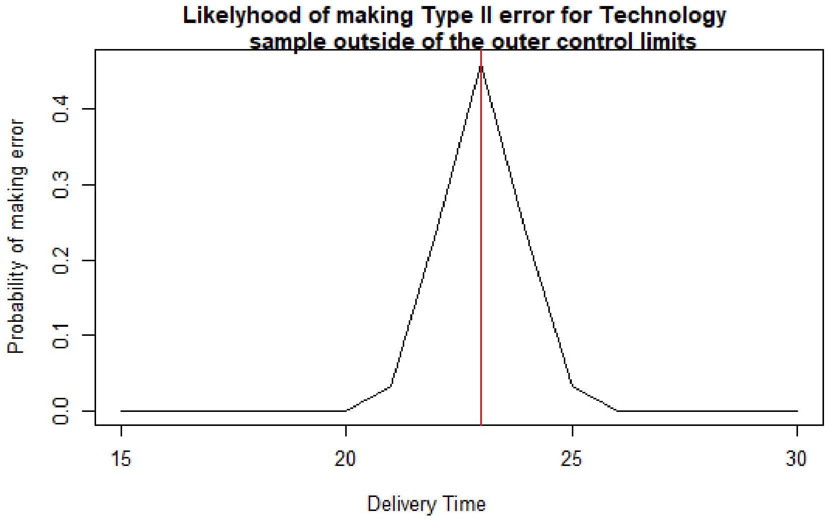


Figure 49: Technology delivery times outside outer control limits - Type II Error

As seen in figure 49 at a mean delivery time of 23, there is a probability of 0.4883177 for a Type II error to occur.

Part 5 – DOE and MANOVA

5.1 First Hypothesis

The first hypothesis is to test if certain ages prefer certain classes of products. This hypothesis will help the company to determine whether they should focus the marketing of certain classes only to certain ages or to all age groups.

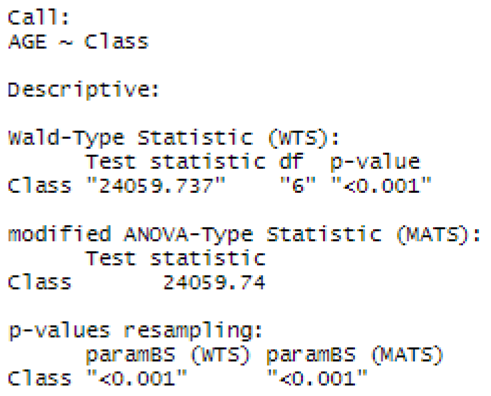
The null hypothesis is that the age of customers has no impact on the class that is bought.

The alternative hypothesis is that the age of customers has an impact on the class that is bought.

This hypothesis test gives the following output as seen in table 16 and figure 50 below.

Table 16: Summary - Hypothesis test 1

Class	n	Age
Technology	36347	46.644
Clothing	26403	47.470
Household	20065	51.927
Luxury	11868	51.339
Food	24582	65.371
Gifts	39149	60.826
Sweets	21564	57.153



```
Call:
AGE ~ Class

Descriptive:

Wald-Type Statistic (WTS):
      Test statistic df  p-value
Class "24059.737"    "6" "<0.001"

modified ANOVA-Type Statistic (MATS):
      Test statistic
Class       24059.74

p-values resampling:
      paramBS (WTS) paramBS (MATS)
Class "<0.001"      "<0.001"
```

Figure 50: ANOVA table - Hypothesis test 1

From the ANOVA table in figure 50 we see that the p value is extremely small. This indicates that there are specific age groups that buys specific classes of products. The null hypothesis is therefore rejected with high certainty and it is confirmed that the age of customers has an impact on the class that is bought.

However, looking at table 16 we see that the mean age for the different classes is somewhat the same. This indicates that the age of customers does not have an impact on the class that is bought. The null hypothesis is therefore not rejected and the hypothesis test is not accurate.

The graph below in figure 51 is plotted to validate the accuracy of the hypothesis test.

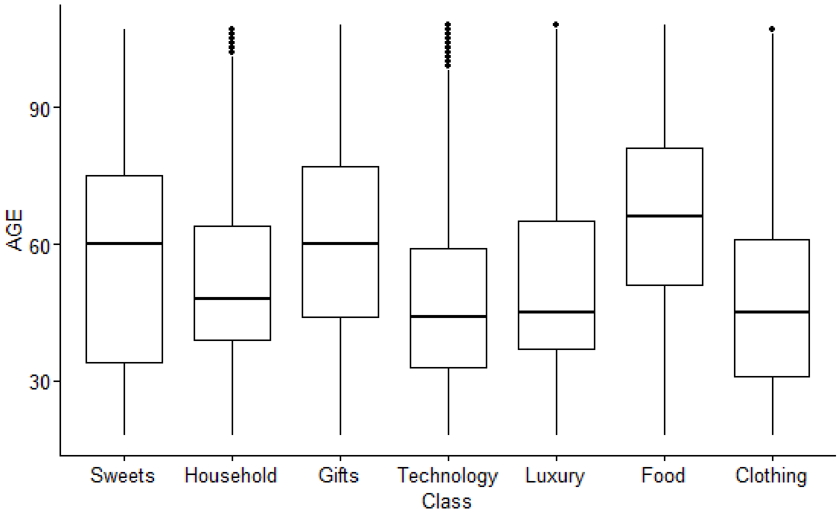


Figure 51: Boxplot - Hypothesis test 1

From the graph it is very clear that the age of customers does not have an impact on the class that is bought, since the mean age of all the classes is somewhat the same. This hypothesis test is therefore not very accurate.

5.2 Second Hypothesis

The second hypothesis is to test if delivery time and sales price is different for all the different classes. The hypothesis will test whether the mean for delivery time and price is the same for al the classes. There are two ways how this can be critical for the company. The company's reliability will be impacted if the assume that the mean delivery time for all the classes is the same. The company will promise the customer a certain delivery time. If the hypothesis is therefore incorrect, the company could lose customers due to low reliability and this will cause the service level to drop.

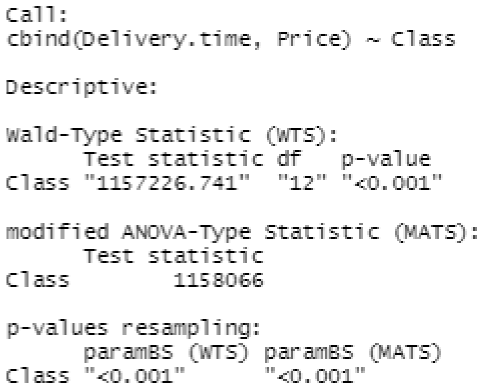
The null hypothesis is that the class has no impact on the price or delivery time of the products.

The alternative hypothesis is that the class has an impact on the price as well as the delivery time of the products.

This hypothesis test gives the following output as seen in table 17 and figure 52 below.

Table 17: Summary - Hypothesis test 2

Class	n	Delivery time	Price
Technology	36347	20.011	29508.063
Clothing	26403	9	640.525
Household	20065	48.720	11009.274
Luxury	11868	3.972	64862.639
Food	24582	2.502	407.815
Gifts	39149	12.891	2961.841
Sweets	21564	2.501	304.070



```
Call:
cbind(Delivery.time, Price) ~ Class

Descriptive:

Wald-Type Statistic (WTS):
      Test statistic df   p-value
Class "1157226.741"  "12" "<0.001"

modified ANOVA-Type Statistic (MATS):
      Test statistic
Class        1158066

p-values resampling:
      paramBS (WTS) paramBS (MATS)
Class "<0.001"      "<0.001"
```

Figure 52: ANOVA table - Hypothesis test 2

From the ANOVA table in figure 52 it is clear that the p value is extremely small. This indicates that the mean of the delivery times and prices is not the same for all the classes. This is confirmed by looking at table 17 the hypothesis test is therefore accurate.

The null hypothesis is therefore rejected with high certainty since the class of a products has significant impact on the price as well as the delivery times.

The company can thus not promise the same delivery time for products in different classes. To be reliable they should give the customers the average delivery times per class. They can also not promise the same service level for all the classes.

The graph below in figure 53 is plotted to confirm whether the hypothesis test was accurate.

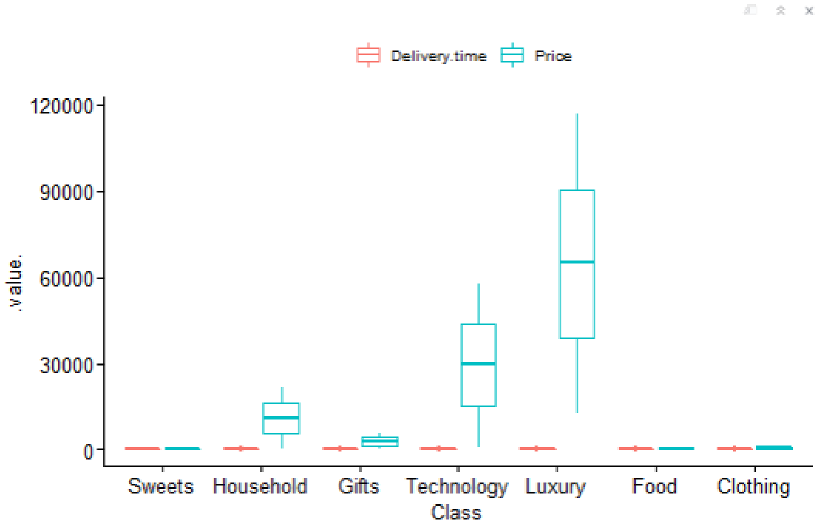


Figure 53: Boxplot - Hypothesis test 2

From the graph it is very clear that the mean for both delivery times and prices for the different classes differ.

Part 6 – Reliability of the service and products

6.1 Problem 6

Thickness specification: 0.06 $\pm$ 0.04 cm

Cost to scrap part that is outside the specifications = $45

Toguchi loss function

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

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

Thus **k = $28 125**

Figure 54 below shows the plot for the Toguchi loss function of problem 6 and the two red lines are situated at x equal to 0.005 and 0.075.

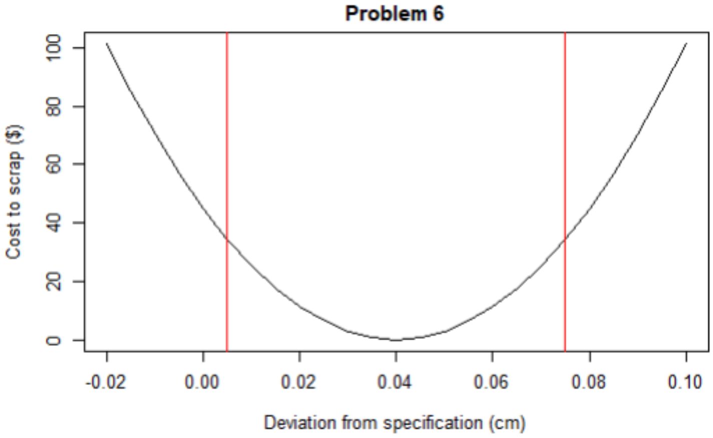


Figure 54: Toguchi Loss Function graph - Problem 6

6.1 Problem 7

a.

Thickness specification: 0.06 0.04 cm

Cost to scrap part that is outside the specifications = $35

<u>Toguchi loss function</u>

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

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

Thus **k = $21 875**

Figure 55 below shows the plot for the Toguchi loss function of problem 7A and the two red lines are situated at x equal to 0.005 and 0.075.

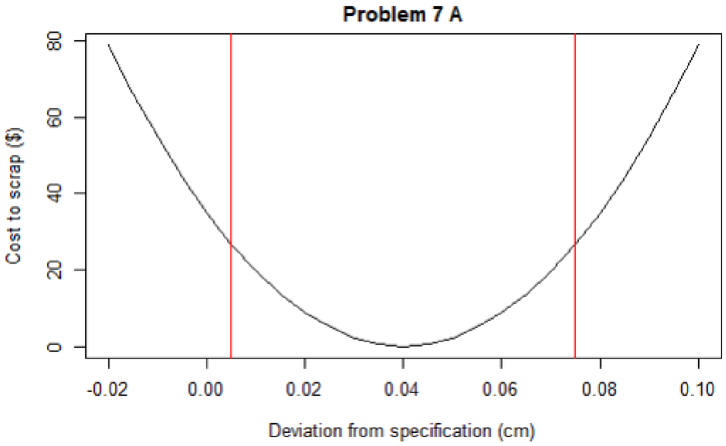


Figure 55: Toguchi Loss Function graph - Problem 7A

b.

Process deviation = 0.027 cm

<u>Toguchi loss function</u>

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

$$35 = k(0.027)^2$$

$$\text{Thus } \boldsymbol{k} = \$\mathbf{48\,010.97}$$

Figure 56 below shows the plot for the Toguchi loss function of problem 7B and the two red lines are situated at x equal to 0.005 and 0.075.

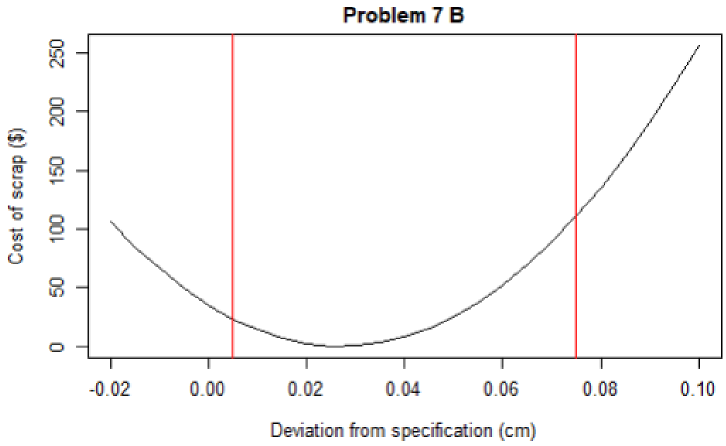


Figure 56: Toguchi Loss Function graph - Problem 7B

6.2 Problem 27

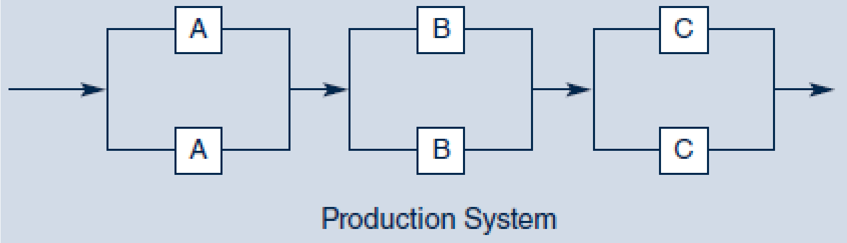


Figure 57: Production System – Problem 27

Table 18: Machine Reliability - Problem 27

Machine	Reliability
A	0.85
B	0.92
C	0.90

a.

Only one machine at each station (the backup machines are out of operation

The machines are therefore in series

$$R_s = R_A \times R_B \times R_C$$

$$R_s = 0.85 \times 0.92 \times 0.90$$

$$\boldsymbol{R_s = 0.7038}$$

Thus, the production system has a reliability of **70.38%** when there is only one machine at each stage.

b.

Two machines at each stage

$$R_p = \left(1-(1-R_A)(1-R_A)\right) \times \left(1-(1-R_B)(1-R_B)\right) \times \left(1-(1-R_C)(1-R_C)\right)$$

$$R_p = \left(1-(1-0.85)(1-0.85)\right) \times \left(1-(1-0.92)(1-0.92)\right) \times \left(1-(1-0.9)(1-0.9)\right)$$

$$\boldsymbol{R_p = 0.96153156}$$

Thus, the production system has a reliability of **96.15%** when there is two machines at each stage.

Thus, the reliability when adding another machine at each stage increases the reliability from 0.7038 to 0.9615

The reliability increases with $\mathbf{25.77}\% = \mathbf{96.15}\% - \mathbf{70.38}\%$

6.3 Availability

Table 19: Available days of Vehicles - 6.3

Available days of Vehicles	
Vehicles (21)	**Days Available (1560)**
20	190
19	22
18	3
17	1

Table 20: Available days of Drivers - 6.3

Available days of Drivers	
Vehicles (21)	**Days Available (1560)**
20	95
19	6
18	1

Assumptions – for the rest of the 1560 days

- 21 Vehicles available
- 21 Drivers available

For a reliable process we need:

- 19 Vehicles available
- 19 Drivers available for the 19 Vehicles

The following equation will be used:

Binomial distribution

$$f(x) = \binom{n}{x} \times (p)^x \times (1-p)^{n-x}$$

$$f(x) = \left(\frac{n!}{x!\,(n-x)!}\right) \times (p)^x \times (1-p)^{n-x}$$

p = Probability of success

n = Number of identical experiments

Part 1: 21 Vehicles and 21 Drivers

Vehicles

P(21 Vehicles) $= \frac{1560-190-22-3-1}{1560} = \binom{21}{0} \times (p)^0 \times (1-p)^{21-0}$

$$\frac{56}{65} = (1-p)^{21}$$

$$\boldsymbol{p = 0.007071808586}$$

P(20 Vehicles) $= \frac{190}{1560} = \binom{21}{1} \times (p)^1 \times (1-p)^{21-1}$

$$\frac{19}{156} = 21 \times p \times (1-p)^{20}$$

$$\boldsymbol{p = 0.0066242837}$$

P(19 Vehicles) $= \frac{22}{1560} = \binom{21}{2} \times (p)^2 \times (1-p)^{21-2}$

$$\frac{19}{156} = 210 \times p^2 \times (1-p)^{19}$$

$$\boldsymbol{p = 0.0089231847}$$

P(18 Vehicles) $= \frac{3}{1560} = \binom{21}{3} \times (p)^3 \times (1-p)^{21-3}$

$$\frac{1}{520} = 1330 \times p^3 \times (1-p)^{18}$$

$$\boldsymbol{p = 0.0121699342}$$

P(17 Vehicles) $= \frac{1}{1560} = \binom{21}{4} \times (p)^4 \times (1-p)^{21-4}$

$$\frac{1}{1560} = 5985 \times p^4 \times (1-p)^{17}$$

$$\boldsymbol{p = 0.0196856635}$$

Weighted $\boldsymbol{p}$ $= \frac{1344(0.007071808586)+190(0.0066252837)+22(0.0089231847)+3(0.0121699342)+1(0.0196856635)}{1560}$

$$\boldsymbol{p = 0.007061301392}$$

P(0 fail) $= \binom{21}{0} \times (0.007061301392)^{0} \times (1 - 0.007061301392)^{21-0}$

$= \mathbf{0.8617299351}$

Expected number of days with 0 failures $= 1560 \times 0.8617299351$

$= \mathbf{1344.298699}\ \boldsymbol{days}$

P(1 fail) $= \binom{21}{1} \times (0.007061301392)^{1} \times (1 - 0.007061301392)^{21-1}$

$= \mathbf{0.1286923662}$

Expected number of days with 1 failures $= 1560 \times 0.1286923662$

$= \mathbf{200.7600912}\ \boldsymbol{days}$

P(2 fail) $= \binom{21}{2} \times (0.007061301392)^{2} \times (1 - 0.007061301392)^{21-2}$

$= \mathbf{0.009151980739}$

Expected number of days with 2 failures $= 1560 \times 0.009151980739$

$= \mathbf{14.27708995}\ \boldsymbol{days}$

P(3 fail) $= \binom{21}{3} \times (0.007061301392)^{3} \times (1 - 0.007061301392)^{21-3}$

$= \mathbf{4.122016777 \times 10^{-4}}$

Expected number of days with 3 failures $= 1560 \times (4.122016777 \times 10^{-4})$

$= \mathbf{0.6430346172}\ \boldsymbol{days}$

P(4 fail) $= \binom{21}{4} \times (0.007061301392)^{4} \times (1 - 0.007061301392)^{21-4}$

$= \mathbf{1.319120836 \times 10^{-5}}$

Expected number of days with 4 failures $= 1560 \times (1.319120836 \times 10^{-5})$

$= \mathbf{0.02057828504}\ \boldsymbol{days}$

Expected percentage of reliable days:

$$\frac{1344.298699 + 200.7600912 + 14.27708995 + 0.6430346172 + 0.02057828504}{1560} \times 100$$

$= \mathbf{99.9999675}\%$ **Reliability**

Number of reliable days:

$$365 \times 99.9999675\%$$

$$= 364.9998814$$

For at least 364 days per year we can expect reliable delivery times

Drivers

P(21 Drivers) $= \frac{1560-95-6-1}{1560} = \binom{21}{0} \times (p)^0 \times (1-p)^{21-0}$

$$\frac{243}{260} = (1-p)^{21}$$

$$\boldsymbol{p = 0.0032148302}$$

P(20 Drivers) $= \frac{95}{1560} = \binom{21}{1} \times (p)^1 \times (1-p)^{21-1}$

$$\frac{19}{312} = 21 \times p \times (1-p)^{20}$$

$$\boldsymbol{p = 0.0030847117}$$

P(19 Drivers) $= \frac{6}{1560} = \binom{21}{2} \times (p)^2 \times (1-p)^{21-2}$

$$\frac{1}{260} = 210 \times p^2 \times (1-p)^{19}$$

$$\boldsymbol{p = 0.0044654848}$$

P(18 Drivers) $= \frac{1}{1560} = \binom{21}{3} \times (p)^3 \times (1-p)^{21-3}$

$$\frac{1}{1560} = 1330 \times p^3 \times (1-p)^{18}$$

$$\boldsymbol{p = 0.008239491}$$

Weighted $\boldsymbol{p} = \frac{1458(0.0032148302)+95(0.0030847117)+6(0.0044654848)+1(0.008239491)}{1560}$

$$\boldsymbol{p = 0.003214937463}$$

P(0 fail) $= \binom{21}{0} \times (0.003214937463)^{0} \times (1 - 0.003214937463)^{21-0}$

$= \mathbf{0.9346132739}$

Expected number of days with 0 failures $= 1560 \times 0.9346132739$

$= \mathbf{1457.996707}\ \boldsymbol{days}$

P(1 fail) $= \binom{21}{1} \times (0.003214937463)^{1} \times (1 - 0.003214937463)^{21-1}$

$= \mathbf{0.06330270201}$

Expected number of days with 1 failures $= 1560 \times 0.06330270201$

$= \mathbf{98.75221514}\ \boldsymbol{days}$

P(2 fail) $= \binom{21}{2} \times (0.003214937463)^{2} \times (1 - 0.003214937463)^{21-2}$

$= \mathbf{0.00204170624}$

Expected number of days with 2 failures $= 1560 \times 0.00204170624$

$= \mathbf{3.185061734}\ \boldsymbol{days}$

P(3 fail) $= \binom{21}{3} \times (0.003214937463)^{3} \times (1 - 0.003214937463)^{21-3}$

$= \mathbf{4.170581482 \times 10^{-5}}$

Expected number of days with 3 failures $= 1560 \times (4.170581482 \times 10^{-5})$

$= \mathbf{0.06506107112}\ \boldsymbol{days}$

Expected percentage of reliable days:

$$\frac{1457.996707 + 98.75221514 + 3.185061734 + 0.06506107112}{1560} \times 100$$

$= \mathbf{99.99993878}\%$ **Reliability**

Number of reliable days:

$365 \times 99.99993878\%$

$= 364.9997765$

For at least 364 days per year we can expect reliable delivery times

Vehicles and Drivers

$$\boldsymbol{P}(\boldsymbol{Reliable}) = P(Vehicles) \times P(Drivers)$$

$$= 0.999999675 \times 0.9999993878$$

$$= 0.9999990628$$

Thus the Vehicles and Drivers together are 99.99990628% Reliable

Number of reliable days:

$$365 \times 99.99990628\%$$

$$= 364.9996579$$

For at least 364 days per year we can expect reliable delivery times

Part 2: 22 Vehicles and 21 Drivers

Vehicles

For vehicles the weighted p = 0.007061301392

P(0 fail) $= \binom{22}{0} \times (0.007061301392)^0 \times (1 - 0.007061301392)^{22-0}$

$= \mathbf{0.8556450003}$

P(1 fail) $= \binom{22}{1} \times (0.007061301392)^1 \times (1 - 0.007061301392)^{22-1}$

$= \mathbf{0.1338685654}$

P(2 fail) $= \binom{22}{2} \times (0.007061301392)^2 \times (1 - 0.007061301392)^{22-2}$

$= \mathbf{0.009996091429}$

P(3 fail) $= \binom{22}{3} \times (0.007061301392)^3 \times (1 - 0.007061301392)^{22-3}$

$= \mathbf{4.739158918 \times 10^{-4}}$

P(4 fail) $= \binom{22}{4} \times (0.007061301392)^4 \times (1 - 0.007061301392)^{22-4}$

$= \mathbf{1.600874154 \times 10^{-5}}$

$$\boldsymbol{P(Reliable)} = (0.8556450003) + (0.1338685654) + (0.009996091429) + (4.739158918 \times 10^{-4}) + (1.600874154 \times 10^{-5})$$

$$= \mathbf{0.9999995818}$$

Drivers

The Reliability for the drivers stays the same as in part 1 since the amount of drivers stayed 21.

Thus the reliability = 99.99993878%

Vehicles and Drivers

$$\boldsymbol{P}(\boldsymbol{Reliable}) = P(Vehicles)\ \times P(Drivers)$$

$$= 0.9999995818\ \times 0.9999993878$$

$$= 0.9999989696$$

Thus the Vehicles and Drivers together are 99.99989696% Reliable

Number of reliable days:

$$365\ \times 99.99989696\%$$

$$= 364.9996239$$

For at least 364 days per year we can expect reliable delivery times

Conclusion

This company caters to a wide variety of customers with various ages and preferences. The Gifts and technology classes dominate the sales. Luxury products is sold the least, but they are sold at the highest price. Therefore, all the classes are important. By analysing why specific groups buy from them as well as which ages prefer to purchase which class of item, marketing strategies can be easily identified.

The most common reason why people buy from this company is because they were recommended to them. This is something the business can be quite proud of since it shows how well-liked and loyal its customers are. They should therefore avoid disappointing the customers in the future, but rather try to obtain or improve their service levels.

The classes with the most stable delivery times are the sweets, food and clothing classes. The technology class's delivery times is also stable, but this class has a high variation. The gifts, luxury and household classes are out of control and should be managed as soon as possible.

The delivery times for luxury products decreased significantly, but this is not a negative aspect. The control limits should just be adapted.

The delivery times for gifts and household products should be improved by management, since it is increasing more and more.

An analysis was conducted to determine if shorter delivery times would result in higher or lower overall costs for the technology class. The conclusion was that the cost decrease op to a 3-day delivery time reduction before dramatically rising after that.

The type I errors done on the X-bar is extremely low. We can therefore assume that the outcomes are accurate. The Type II errors done on the X-bar however is large and it can become larger if the mean of the delivery time is shifted without taking it into account. We can assume that the outcomes aren't that accurate and that these calculations should be regularly renewed if the means start to shift.

Two hypothesis tests were done to evaluate the strength of the samples and give useful information. The first hypothesis was to test if certain ages prefer certain classes of products. And the second hypothesis was to test if delivery time and sales price is different for all the different classes. Both these tests gave meaningful information that the company will be able to use to grow.

Improvement projects are explored and we can see that adding an extra vehicle won't have a big effect on the reliability. The reliability will stay somewhat the same. It will therefore not be a good financial decision if the company decides to add an extra vehicle.

To conclude, data analysis can provide a lot of useful information which can be used to improve a business. These improvements can include improving marketing tactics, prices and delivery times. Therefore, as we explore data further and further, more information will become visible.

References

Scribbr. (2021). *Type I and Type II errors*. [online] Available at: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/#:~:text=What%20are%20Type%20I%20and.

support.minitab.com. (n.d.). *Interpret the key results for Xbar-R Chart*. [online] Available at: https://support.minitab.com/en-us/minitab/21/help-and-how-to/quality-and-process-improvement/control-charts/how-to/variables-charts-for-subgroups/xbar-r-chart/interpret-the-results/key-results/ [Accessed 20 Oct. 2022].