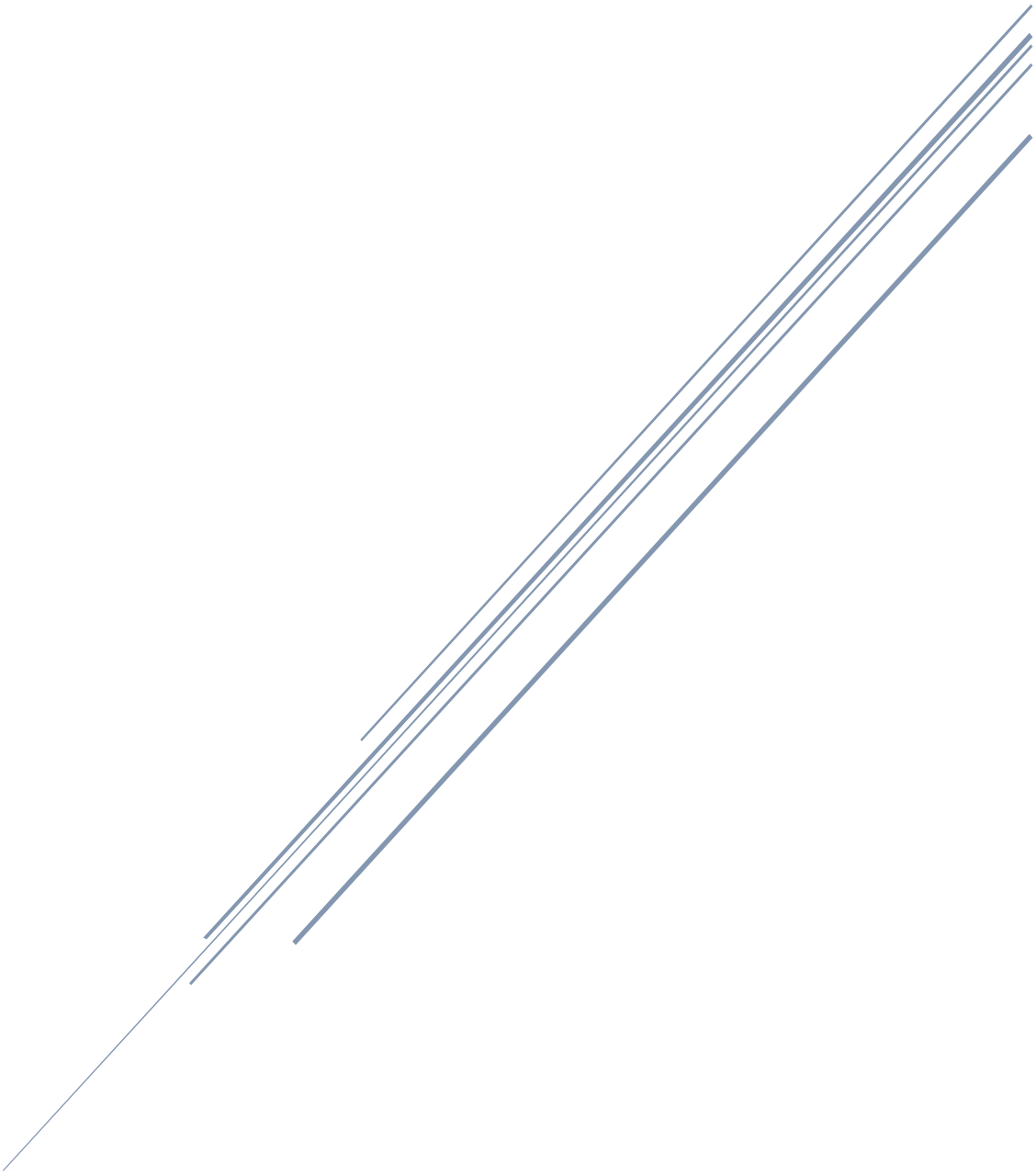
ECSA SALES DATA ANALYSIS

PB van Rhyn 23549459



Stellenbosch University
Quality Assurance 344

Abstract

The report documents and analysis data wrangling and manipulation processes of an online business sales data in R-Studio. Raw sales data is converted into valid usable data. Relationships and trends are analysed to plot and understand the data sufficiently. Statistical analysis is performed to compute all control limits. Control charts are plotted to identify out of time deliveries. Further type I and type II error probabilities are calculated. MANOVA's are performed and identifies how different descriptive features impacts one another. The reliability of services and products from different subsidiaries are also analysed.

Table of Contents

List of Figures

List of Tables

Introduction

The report analyses the sales data and investigates the quality of an online business. The purpose of the report is the interpret, analyse, and evaluate the sales data. Thus, it is possible to predict trends and probabilities. The outcomes are presented in the form of a statistical analysis. This includes data wrangling which all valid data has been extracted from the sales data set. Descriptive statistics which include plots, table, and distributions, and provides a detailed evaluation of the data. The process capability analysis is performed. Which indicates if the processes are viable. Statistical process control is performed using the delivery time of the different classes of the sales data. Which helps to control the process of the business. Delivery times will further be optimised by calculating the type I and type II errors. Finally, the reliability of services and products from different subsidiaries are calculated and analysed.

Part 1: Data Wrangling

The data within the “salesTable2022.csv” data set contains both valid and invalid data. The invalid data must be removed for accurate prediction and further processing.

The invalid data will incorrectly influence the data. Missing values (NA) in other words, values which was incorrectly added, needs to be removed. Negative values are impossible in the sales data’s context, thus must also be removed. The negative values are converted to NA values and is then removed with the rest of the NA values.

All the invalid data is added to a new variable called invalid.

All the invalid data can be viewed below (There are 22 instances):

X <int>	ID <int>	AGE <int>	Class <chr>	Price <dbl>	Year <int>	Month <int>	Day <int>	Delivery.time <dbl>	Why.Bought <chr>
12345	18973	93	Gifts	NA	2026	6	11	15.5	Website
16320	44142	82	Household	NA	2023	10	2	48.0	EMail
16321	81959	43	Technology	NA	2029	9	6	22.0	Recommended
19540	65689	96	Sweets	NA	2028	4	7	3.0	Random
19541	71169	42	Technology	NA	2025	1	19	20.5	Recommended
19998	68743	45	Household	NA	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
54321	62209	34	Clothing	NA	2021	3	24	9.5	Recommended
56789	63849	51	Gifts	NA	2024	5	3	10.5	Website
65432	51904	31	Gifts	NA	2027	7	24	14.5	Recommended
76543	79732	71	Food	NA	2028	9	24	2.5	Recommended
87654	40983	33	Food	NA	2024	8	27	2.0	Recommended
98765	64288	25	Clothing	NA	2021	1	24	8.5	Browsing
144443	37737	81	Food	NA	2022	12	10	2.5	Recommended
144444	70761	70	Food	NA	2027	9	28	2.5	Recommended
155554	36599	29	Luxury	NA	2026	4	14	3.5	Recommended
155555	33583	56	Gifts	NA	2022	12	9	10.0	Recommended
166666	60188	37	Technology	NA	2024	10	9	21.5	Website
177777	68698	30	Food	NA	2023	8	14	2.5	Recommended

Figure 1: invalid data

All valid data is used to create a new data frame and is assigned to new variable called valid. Further predictions and calculation will use valid as its dataset.

Part 2: Descriptive Statistics

The valid data set is analysed using different descriptive statistic methods. The results of the data analysis will indicate trends and help summarise the data. Accurate predictions will be possible after evaluation.

5-point summary

The 5-point summary indicates the minimum, 1st quartile, median, mean, 3rd quartile and the maximum of the valid data set. The 5-point summary of the following features is given below: Age, Price, Year, Month, Day, and Delivery.time

R output:

```
      AGE              Price               Year           Month      
 Min.   : 18.00   Min.   :    35.65   Min.   :2021   Min.   : 1.000  
 1st Qu.: 38.00   1st Qu.:   482.31   1st Qu.:2022   1st Qu.: 4.000  
 Median : 53.00   Median :  2259.63   Median :2025   Median : 7.000  
 Mean   : 54.57   Mean   : 12294.10   Mean   :2025   Mean   : 6.521  
 3rd Qu.: 70.00   3rd Qu.: 15270.97   3rd Qu.:2027   3rd Qu.:10.000  
 Max.   :108.00   Max.   :116618.97   Max.   :2029   Max.   :12.000  

      Day        Delivery.time  
 Min.   : 1.00   Min.   : 0.5  
 1st Qu.: 8.00   1st Qu.: 3.0  
 Median :16.00   Median :10.0  
 Mean   :15.54   Mean   :14.5  
 3rd Qu.:23.00   3rd Qu.:18.5  
 Max.   :30.00   Max.   :75.0  
```

Figure 2: 5-point summary

The min, max and mean of all the features are easily identifiable. It is easier to understand the distributions of each of the above-mentioned features. We can see that the youngest customer is 18 years old and the oldest 108 years old. The median and mean age of the customers is around 53 to 54. Similar assumptions can be made of the features from using the 5-point summary. Further analysis is needed.

Numerical summaries of important averages

The following numerical summaries contain important averages of the valid data set, indicating important classes and features for the online business.

Average sales price of different sales classes

The figure below indicates the average price and all the instances of clients (count) of each class:

R output:

Class <chr>	Count <int>	Avg_Price <dbl>
Clothing	26403	640.5253
Food	24583	407.7747
Gifts	39149	2961.8414
Household	20067	11008.1179
Luxury	11869	64857.1241
Sweets	21565	304.0290
Technology	36347	29508.0626

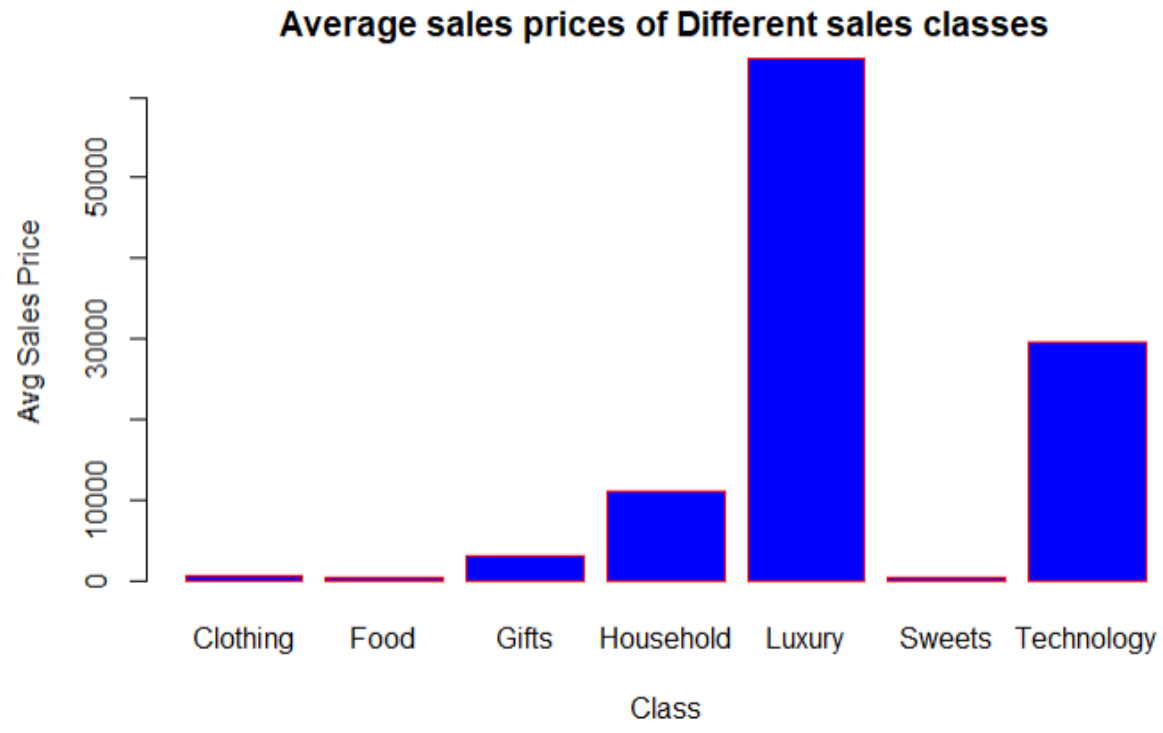


Figure 3: Average sales price of classes

Luxury items and technology are the most expensive on average. Those items should also be the most expensive to manufacture. Whilst sweets and food are the least expensive items on average. Food, sweets, and clothing should be the least expensive items, they are the least expensive to manufacture.

Average ages of the different sales classes clients

The figure below indicates the average age and all the instances of clients (count) of each class:

R output:

Class <chr>	count <int>	Avg_Age <dbl>
Clothing	26403	47.46980
Food	24583	65.37213
Gifts	39149	60.82559
Household	20067	51.92794
Luxury	11869	51.33743
Sweets	21565	57.15493
Technology	36347	46.64399

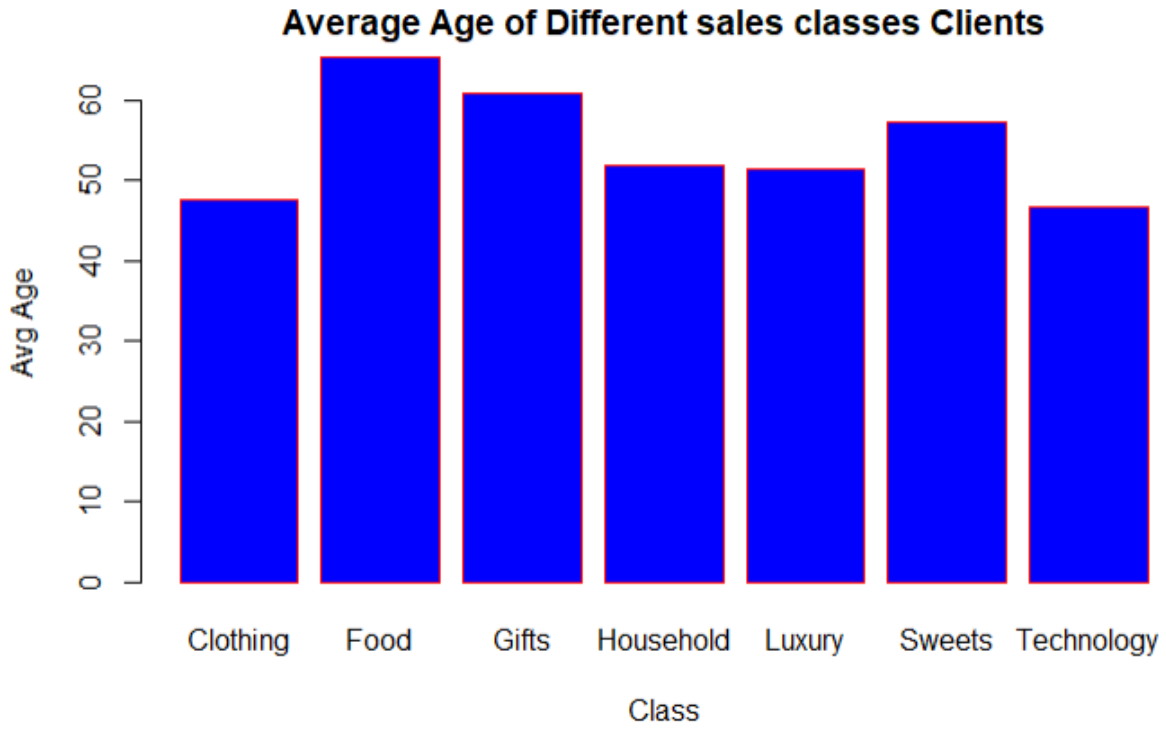


Figure 4: Average age of classes' clients

On average the class that has the youngest clients on average is technology and clothing. This indicates that clothing and technology has a younger target audience than the other classes. Older customers are less likely to buy technology. Food has the highest average age. Food is needed across all age groups, caregivers of families usually also buy food, which will also raise the average age of the specific class.

Average delivery time of the different sales classes

The figure below indicates the average Delivery time and all the instances of clients (count) of each class:

R output:

Class <chr>	count <int>	Delivery_time <dbl>
Clothing	26403	8.999527
Food	24583	2.502014
Gifts	39149	12.890546
Household	20067	48.719365
Luxury	11869	3.971480
Sweets	21565	2.501229
Technology	36347	20.010950

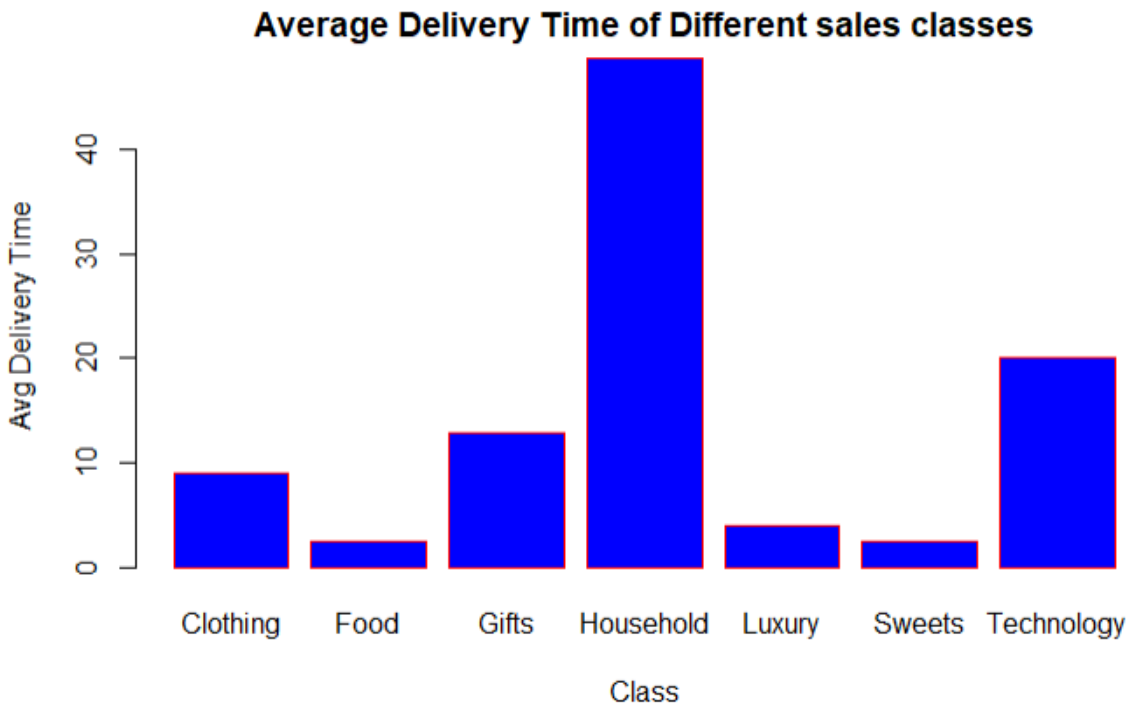


Figure 5: Average delivery time of classes

On average household items have the longest delivery time. Which indicates that they are the most cumbersome to manage, pack and ship and is most probably the largest. Sweets and food have the lowest delivery time. This indicates that they are easy to manage and are small and easy to pack and ship.

Average age of clients using the different why bought methods

The figure below indicates the average age and reasoning of all the instances of clients (count) buying a product from the online business:

R output:

Why.Bought <chr>	count <int>	Avg_Age <dbl>
Browsing	18994	53.84874
EMail	7225	55.75543
Random	13121	56.96250
Recommended	106988	54.48060
Spam	4208	54.65898
Website	29447	53.96509

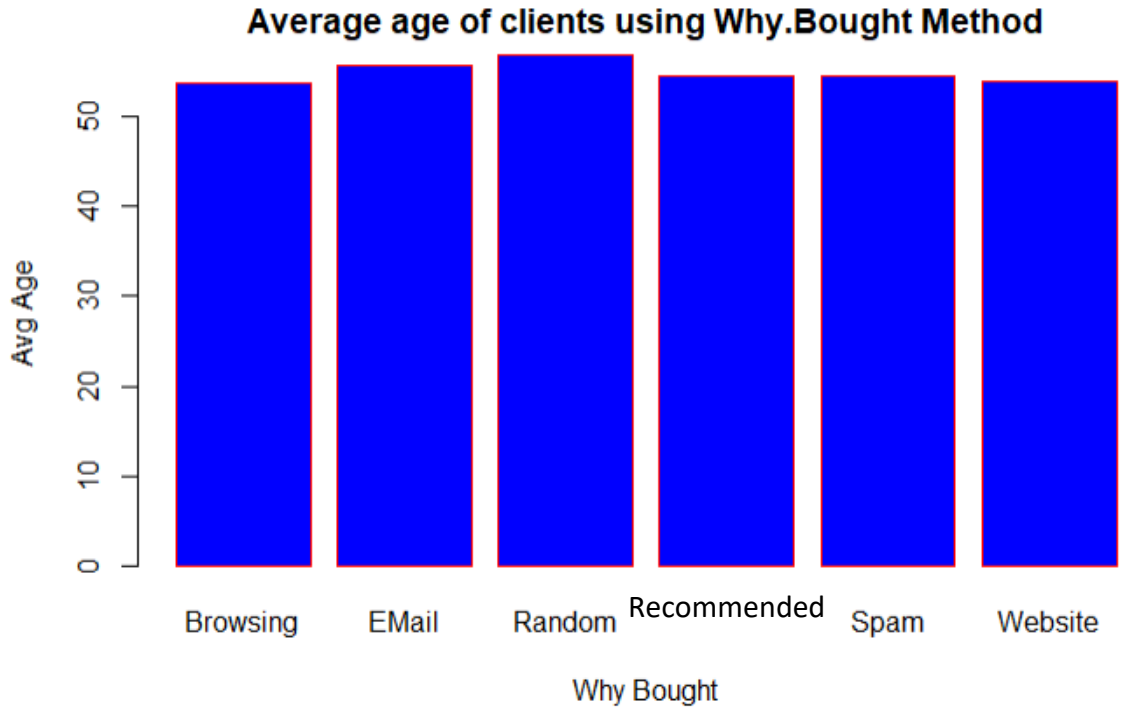


Figure 6: Average age of buying methods

Website and browsing has the lowest average age. This indicates that younger clients are more likely to buy products through browsing the internet and the companies' websites than older clients. Random and email has the highest average age. This indicates that older clients are more likely to make a purchase through receiving an email or though random means.

Numerical Summary on why clients bought a certain class

The following numerical summaries contain plots and figures indicating why client is bought certain classes. This contains vital information regarding on how each class should be advertised to customers. All the classes main source of purchasing is through recommendations. The focus shall be shifted to the other instances.

Why clients bought clothing

The figure below indicates the reasons why clients bought clothes and its instances (count) from the online business:

R output:

Why.Bought <chr>	count <int>
Browsing	1774
EMail	1042
Random	2222
Recommended	17345
Spam	602
Website	3418

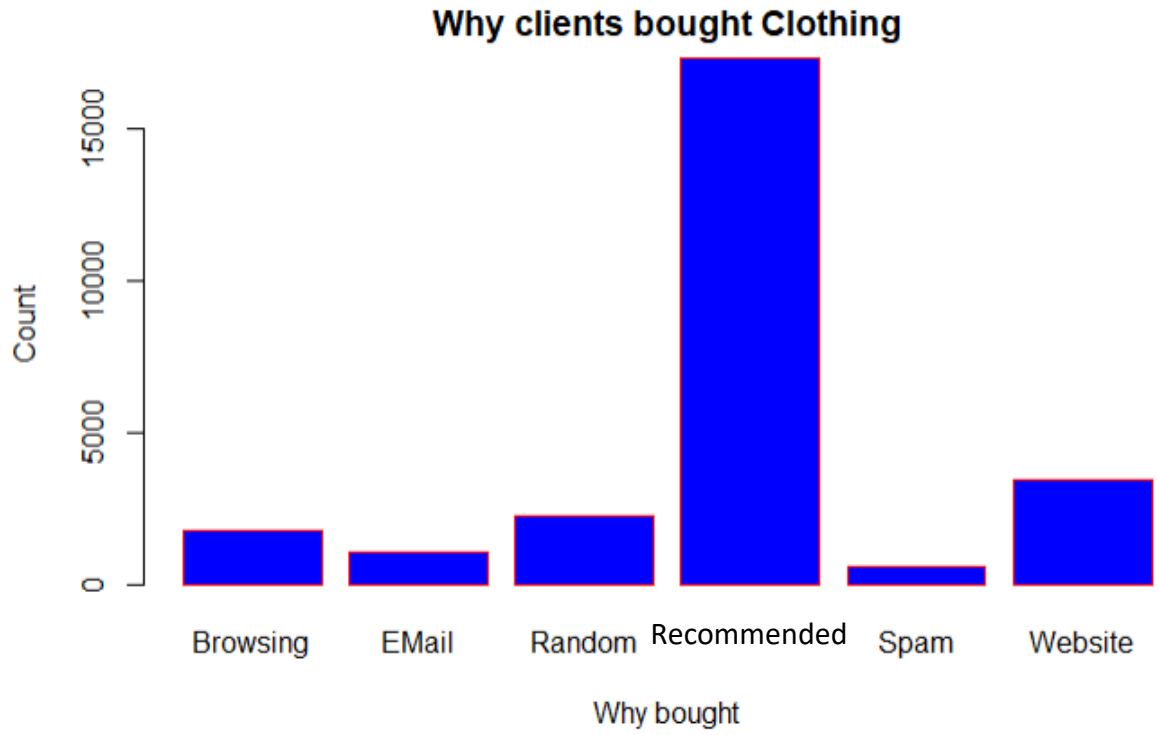


Figure 7: Why clients bought clothing

Most clothes were bought through recommendation. Indicating that people are the most likely to buy clothes though other people recommending it to them. All other instances are equally low

Why clients bought food

The figure below indicates the reasons why clients bought food and its instances (count) from the online business:

R output:

Why.Bought <chr>	count <int>
Browsing	1544
EMail	1056
Random	1986
Recommended	16466
Spam	509
Website	3022

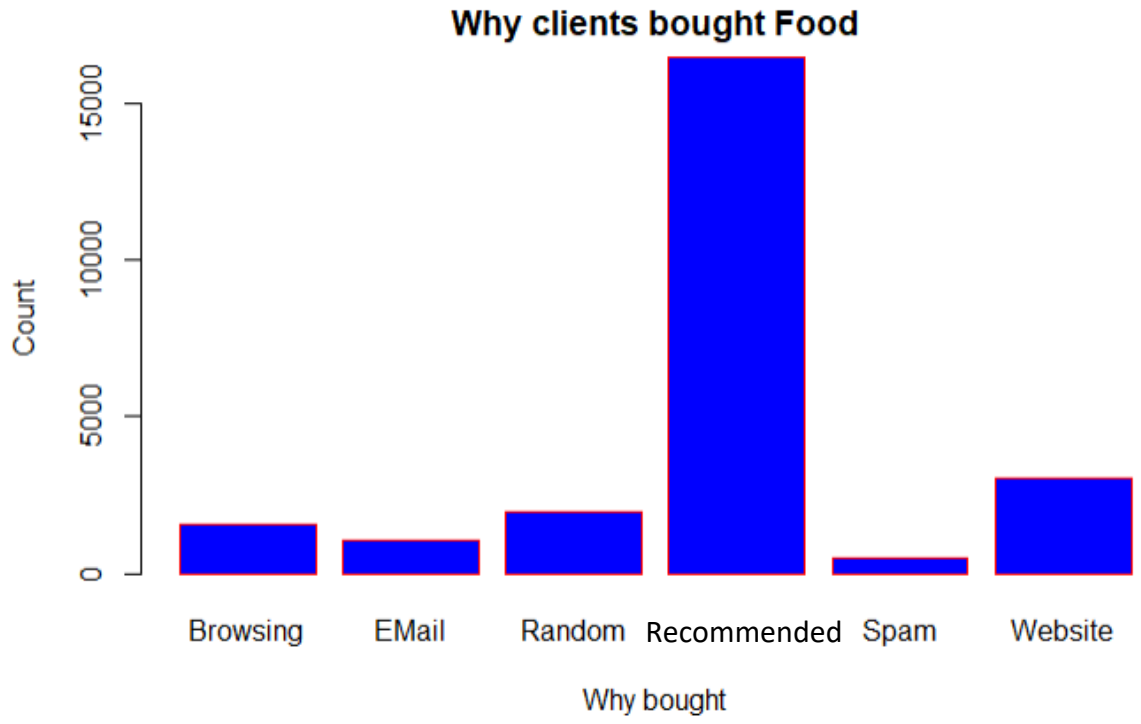


Figure 8: Why clients bough food

Most food were bought through recommendation. It follows a similar distribution to that of clothing. This indicates that people are most likely to buy food through recommendation. All other instances are equally low in comparison with food.

Why clients bought gifts

The figure below indicates the reasons why clients bought gifts and its instances (count) from the online business:

R output:

Why.Bought <chr>	count <int>
Browsing	4320
EMail	2133
Random	4244
Recommended	21233
Spam	1011
Website	6208

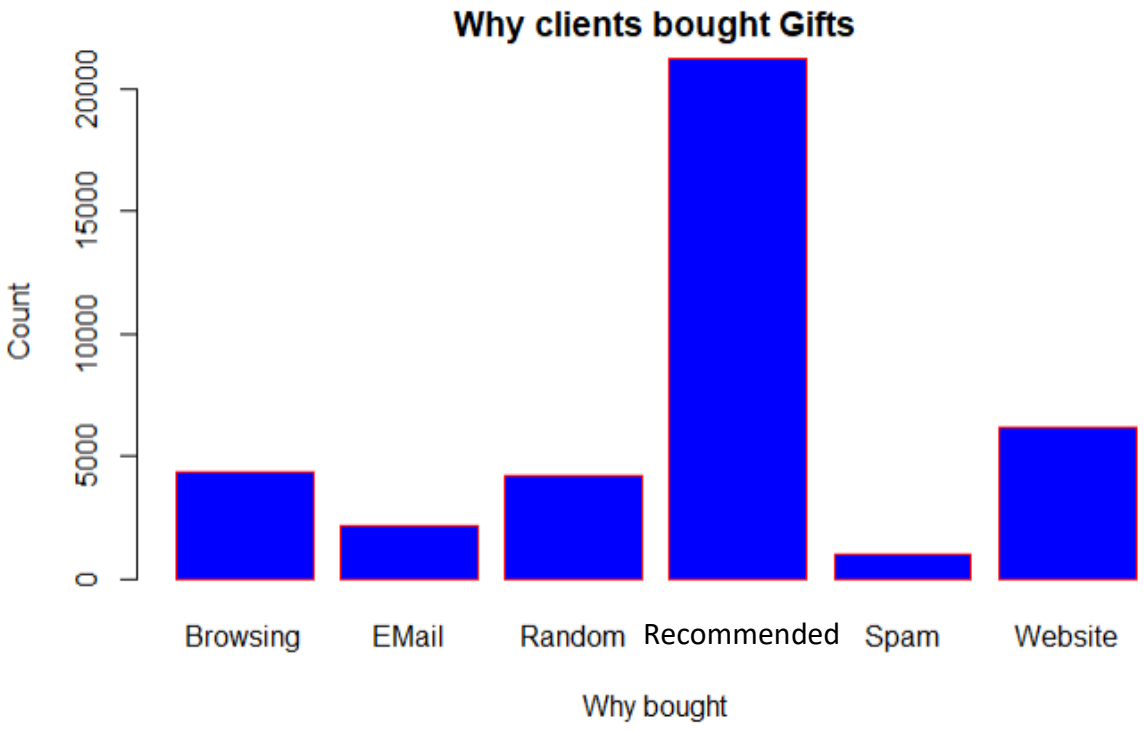


Figure 9: Why clients bought gifts

Most gifts were bought through recommendation. Gifts differs from the food and clothing distribution. Recommendation is more closely related to the other methods. More gifts were bought from website and browsing in proportion to the other distributions. This indicates that clients might not have a specific gift in mind and would like to browse multiple options before selecting a certain product as a gift.

Why clients bought household products

The figure below indicates the reasons why clients bought household and its instances (count) from the online business:

R output:

Why.Bought <chr>	count <int>
Browsing	1387
EMail	944
Random	1902
Recommended	9141
Spam	495
Website	6198

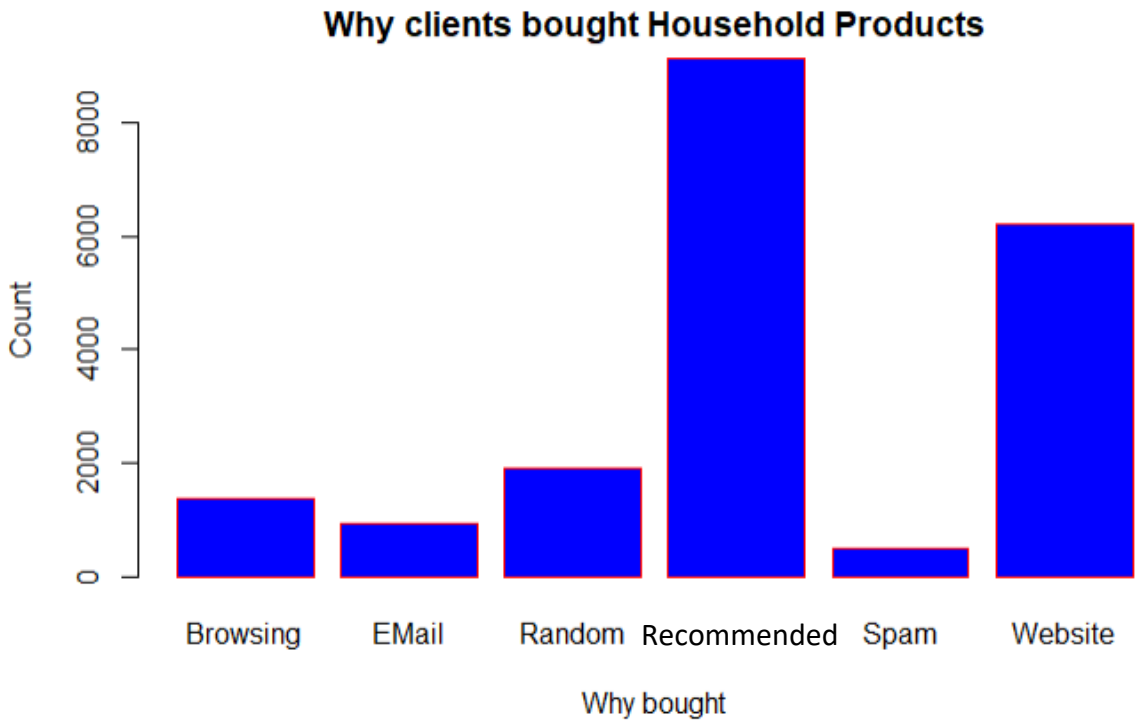


Figure 10: Why clients bought household products

Most household products were bought through recommendation. In comparison with the other distributions, a lot more is bought through website. This indicates that clients are more likely to view multiple options of the household products on the company website. To see which would fit there project the best.

Why clients bought luxury items

The figure below indicates the reasons why clients bought luxury items and its instances (count) from the online business:

R output:

Why.Bought <chr>	count <int>
Browsing	1394
EMail	118
Random	137
Recommended	9451
Spam	61
Website	708

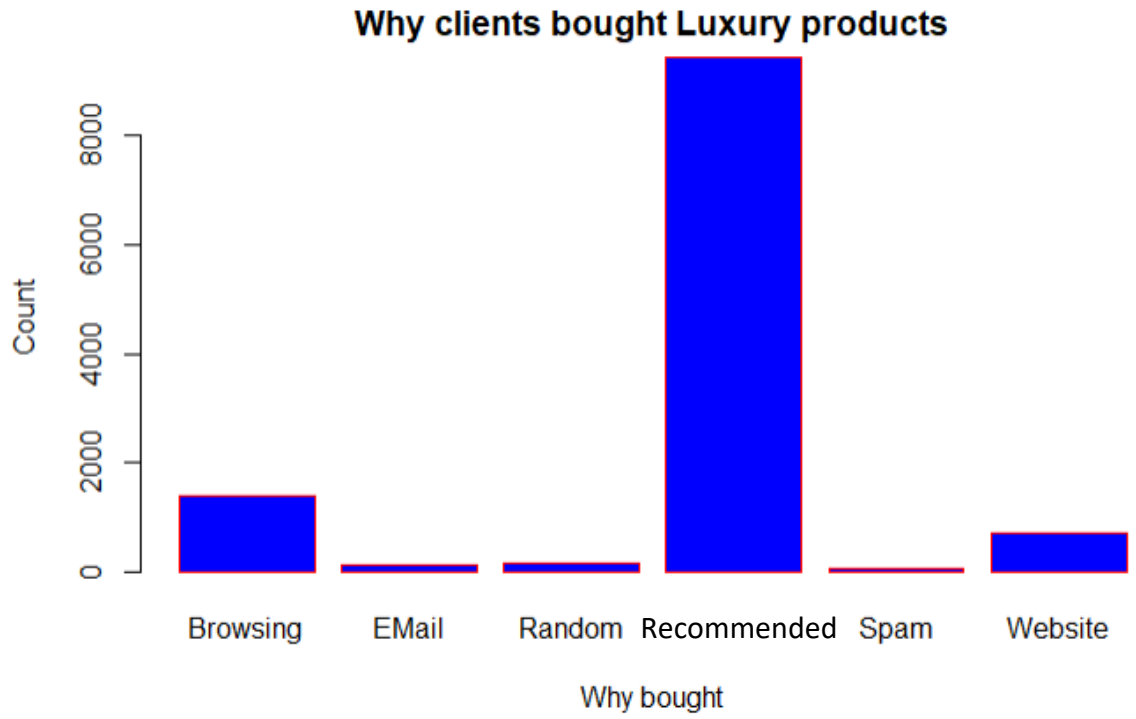


Figure 11: Why clients bought luxury items

Most luxury items were bought through recommendation. Very few luxury items were bought from emails, random and spam in comparison with the other distributions. This indicates that customers are less likely to buy the expensive luxury items from less trusted sources such as email, spam and random.

Why clients bought sweets

The figure below indicates the reasons why clients bought sweets and its instances (count) from the online business:

R output:

Why.Bought <chr>	count <int>
Browsing	2344
EMail	1163
Random	2253
Recommended	11681
Spam	598
Website	3526

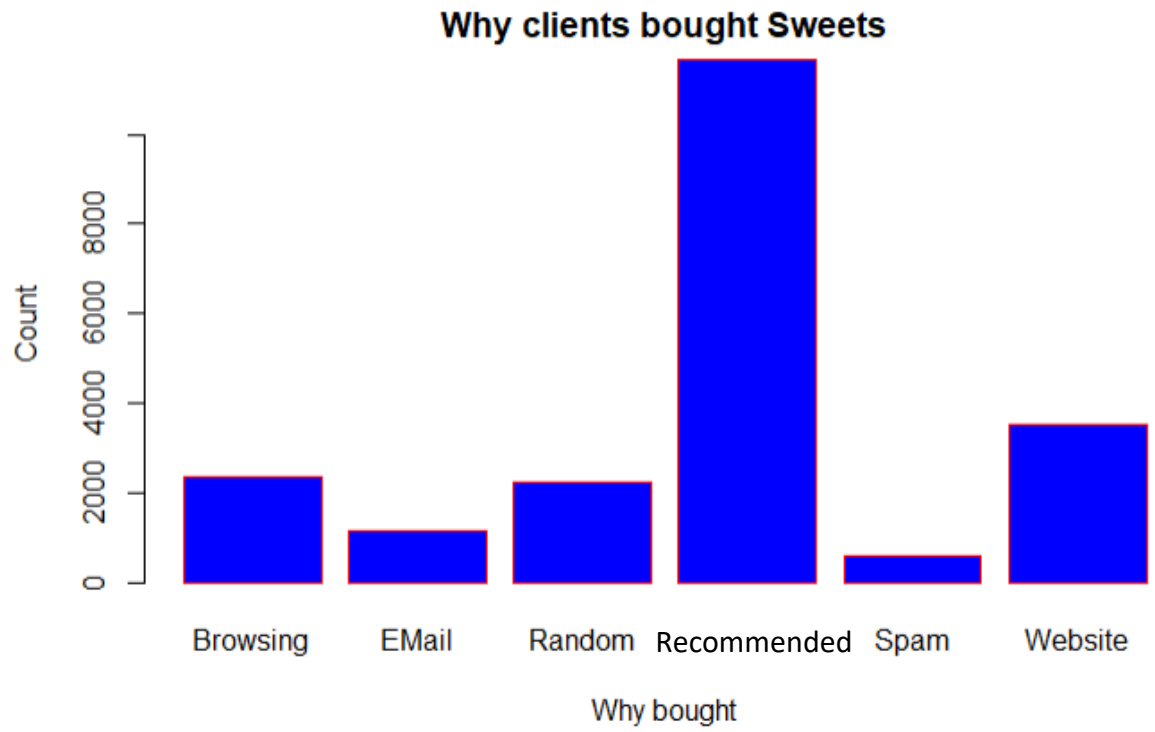


Figure 12: Why clients bought sweets

Most sweets items were bought through recommendation. In comparison with the other distributions sweets has been more likely to be bought randomly.

Why clients bought technology

The figure below indicates the reasons why clients bought technology and its instances (count) from the online business:

R output:

Why.Bought <chr>	count <int>
Browsing	6231
EMail	769
Random	377
Recommended	21671
Spam	932
Website	6367

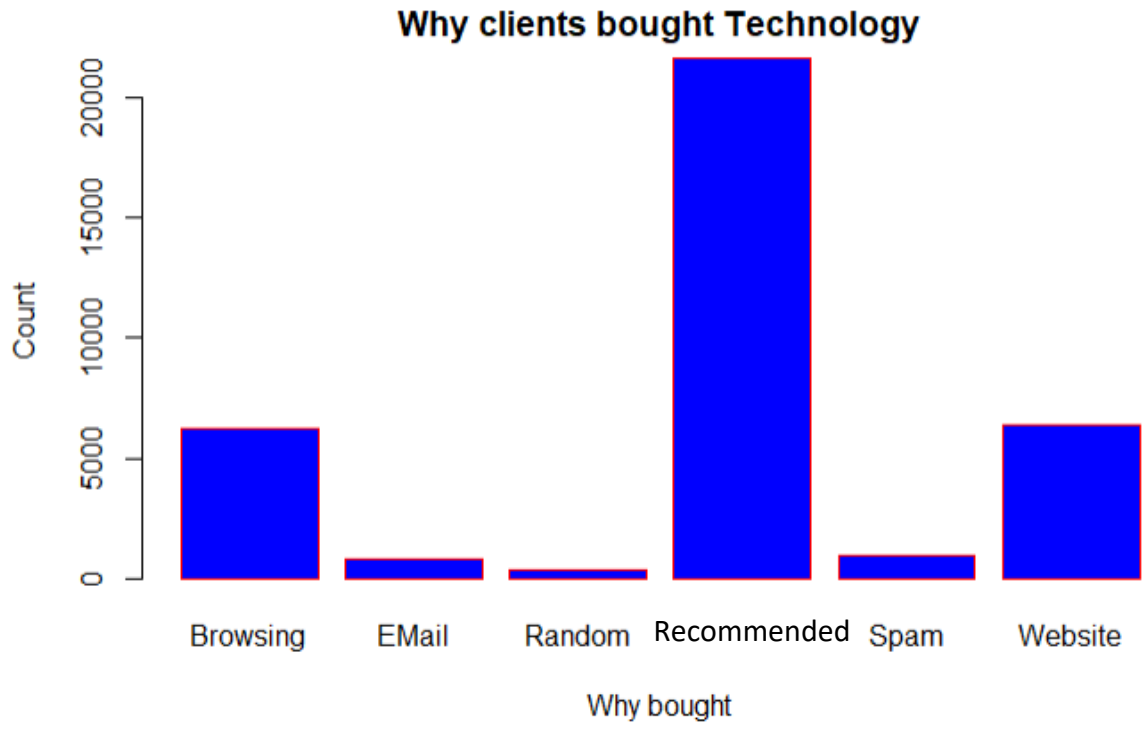


Figure 13: Why clients bought Technology

Most technology has been bought through recommendation. In comparison with the other distribution very few technology products are bought randomly. Proportionately more technology is bought through browsing than the other distributions. Indicating that clients could be searches a certain product on the web or certain specification of a product on the web then purchasing it from the online business on the shopping tab of the browser

Age distributions of the different classes' clients

The following plots indicate the age distributions of clients buying the assorted products that the online business offers. It would be possible to indicate customer trends and to cater more efficiently to specific customers and their needs.

Age distribution customers buying clothing

R output:

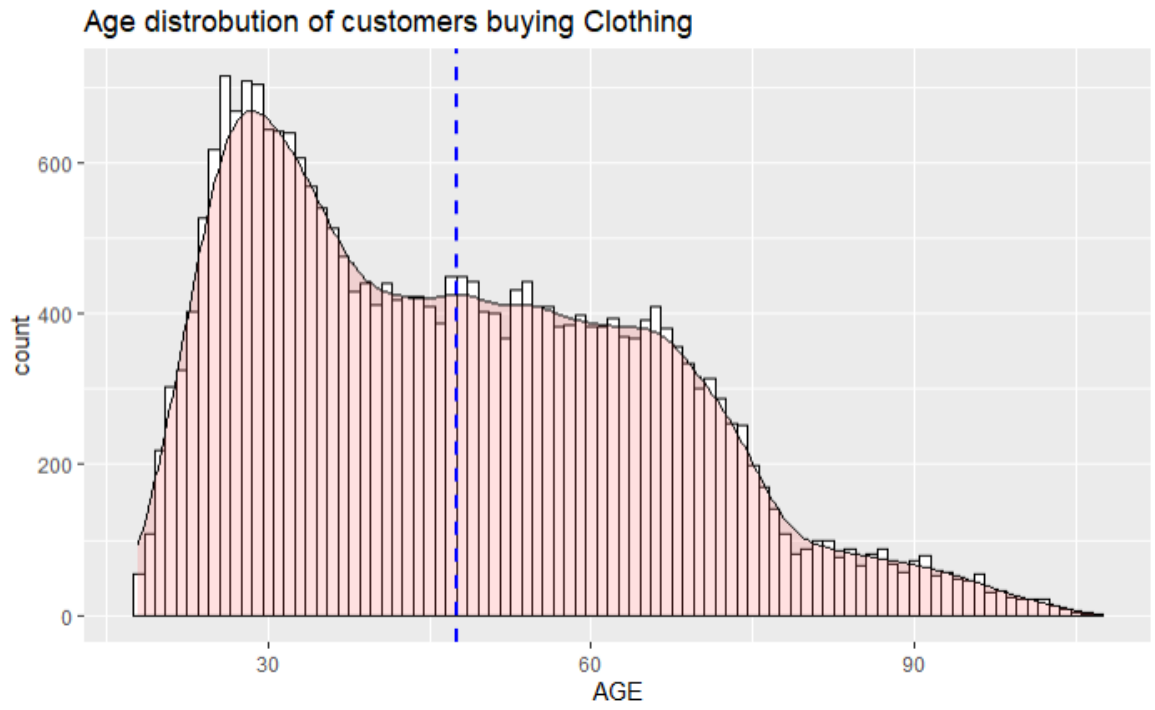


Figure 14: Age distribution of clothing

The customers purchasing clothing from the online business age distribution is skewed to the right. This indicates that there is a higher concentration of data to the left region of the x axis. This indicates that a younger demographic of clients buys clothing more frequently. The data has a fat tail to the right side. Which indicates that the frequency of customers buying clothing decreases slightly as age increases (there is not a significant drop in frequency between the ages of 40 and 67). After the age of 67 years, there is a significant drop of in purchase frequency. We can assume that after the age of 67 years clients are less likely and interested in buying clothes more frequently. The blue line indicates the mean age of customers buying clothing

Age distribution of customers buying food

R output:

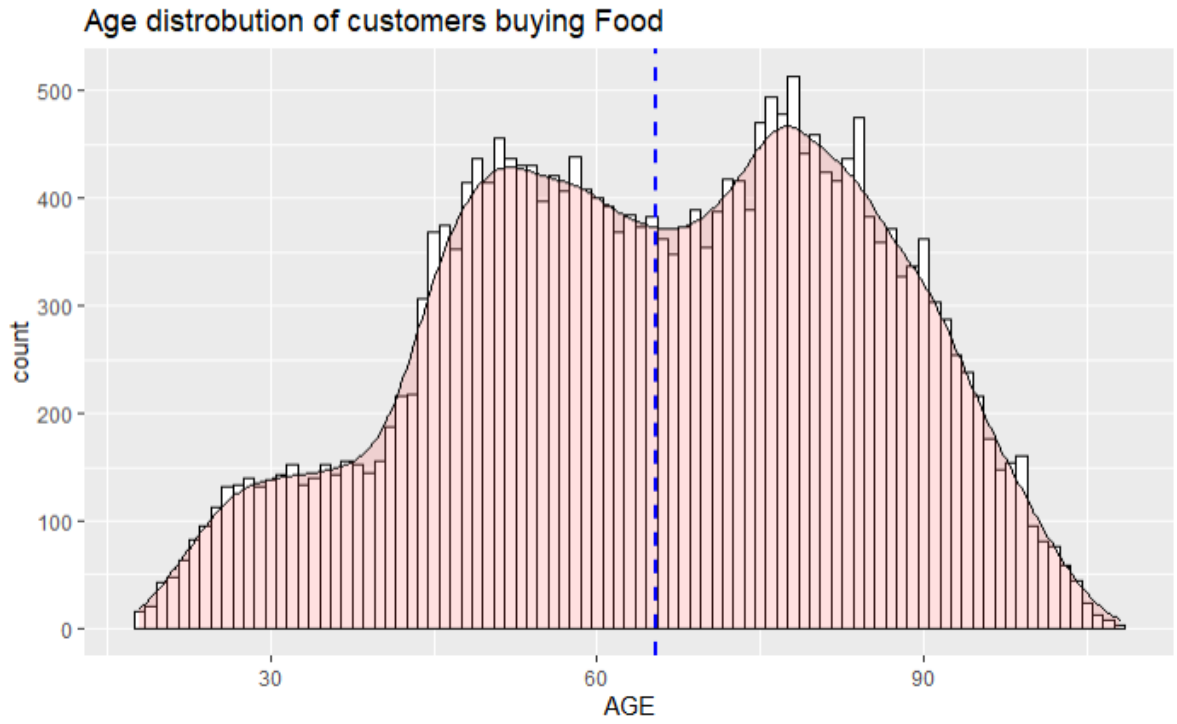


Figure 15: Age distribution of food

The customers purchasing food from the online business age distribution has a bimodal distribution (Bimodal, 2022). This indicates that the data has two modes and local maxima. The data has a first local maxima between the age groups of 45 and 55. The second local maxima is between the age groups of 75 and 85. The company should investigate why the frequency of food purchases decreases after the age of 55 and increases around the age of 70. The blue line indicates the mean age of customers buying food

Age distribution of customers buying gifts

R output:

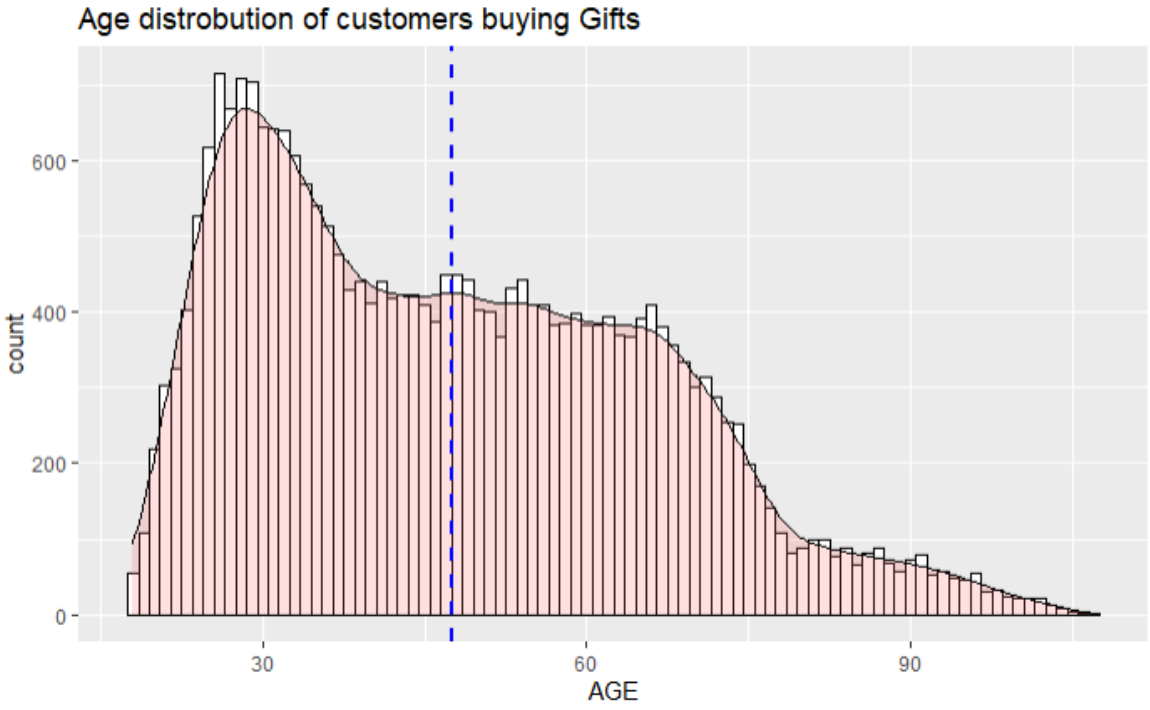


Figure 16: Age distribution of gifts

The customers purchasing gifts from the online business age distribution is skewed to the right. This indicates that there is a higher concentration of data to the left region of the x axis. This indicates that a younger demographic of clients buys gifts more frequently. The data has a long fat tail to the right side. Which indicates that the frequency of customers buying gifts decreases slightly as age increases (there is not a significant drop in frequency between the ages of 35 and 70). After the age of 80 years, there is a significant drop of in purchase frequency. We can assume that after the age of 80 years clients are less likely and interested in buying clothes more frequently. Customers between the ages of 24 and 35 are the most likely to buy a gift from the online business. The blue line indicates the mean age of customers buying gifts

Age distribution of customers buying household products

R output:

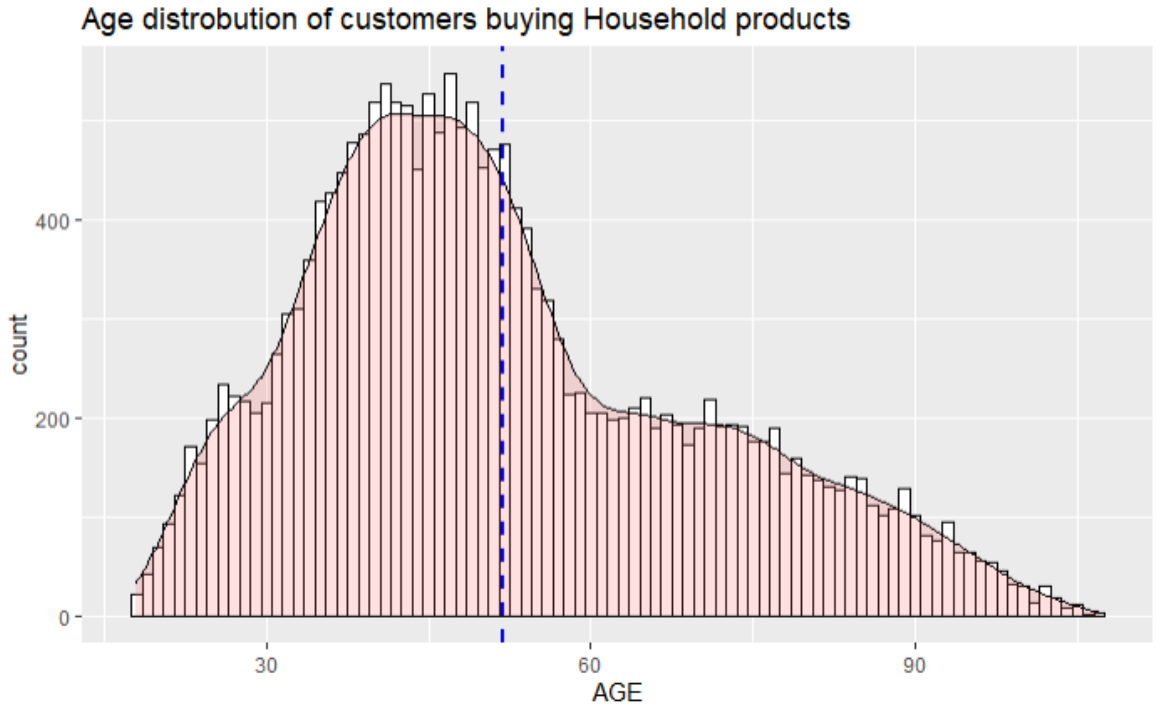


Figure 17: Age distribution of household products

The customers purchasing household products from the online business age distribution is skewed to the right. This indicates that there is a higher concentration of data to the left region of the x axis. This indicates that a younger demographic of clients buys household products more frequently. The age distribution of the customers buying household products does not have a fat tail in comparison with the age distributions of the gifts and clothing classes. There is a steeper drop of in purchasing frequency after the mean age (indicated by the blue line)

Age distribution of customers buying luxury products

R output:

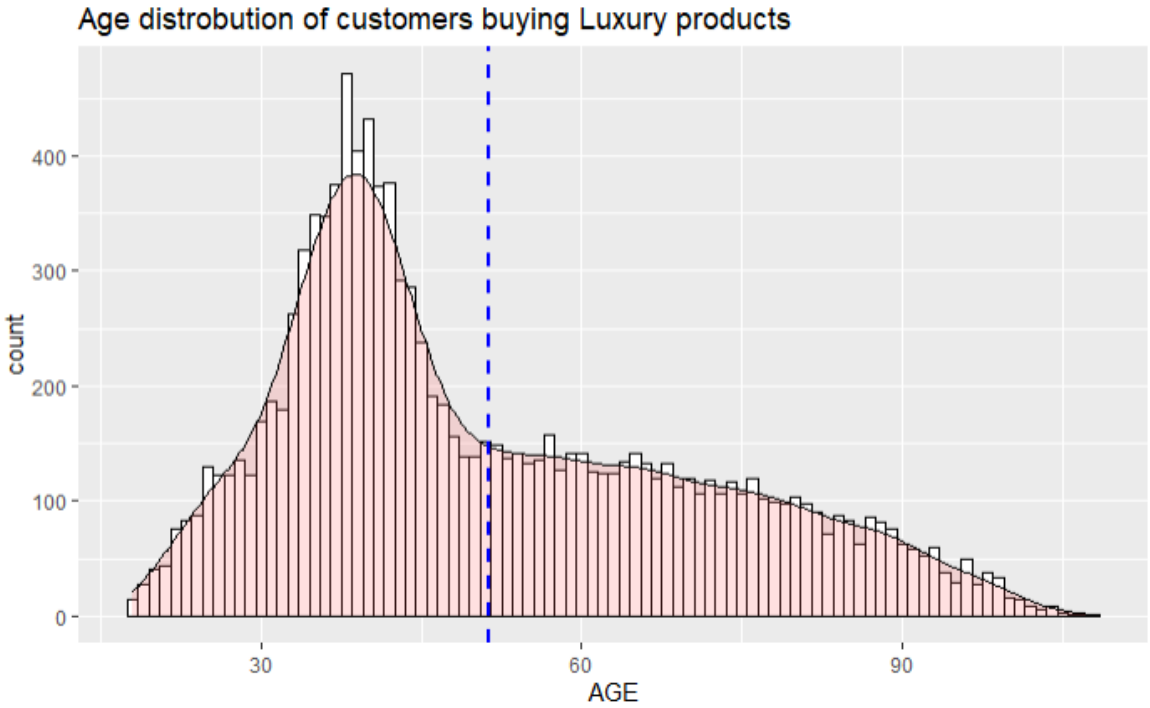


Figure 18: Age distribution of luxury products

The customers purchasing luxury products from the online business age distribution is skewed to the right. This indicates that there is a higher concentration of data to the left region of the x axis. This indicates that a younger demographic of clients buys luxury products more frequently. The age distribution of the customers buying luxury products does not have a fat tail in comparison with the age distributions of the gifts and clothing classes. There is a steeper drop of in purchasing frequency after the mean age (indicated by the blue line). Proportionately the difference in frequency between the local maxima and the right-side skewed data is larger than the previous data that had similar distributions. Which indicates that there is a higher concentration of a younger demographic buying the luxury products.

Age distribution of customers buying sweets

R output:

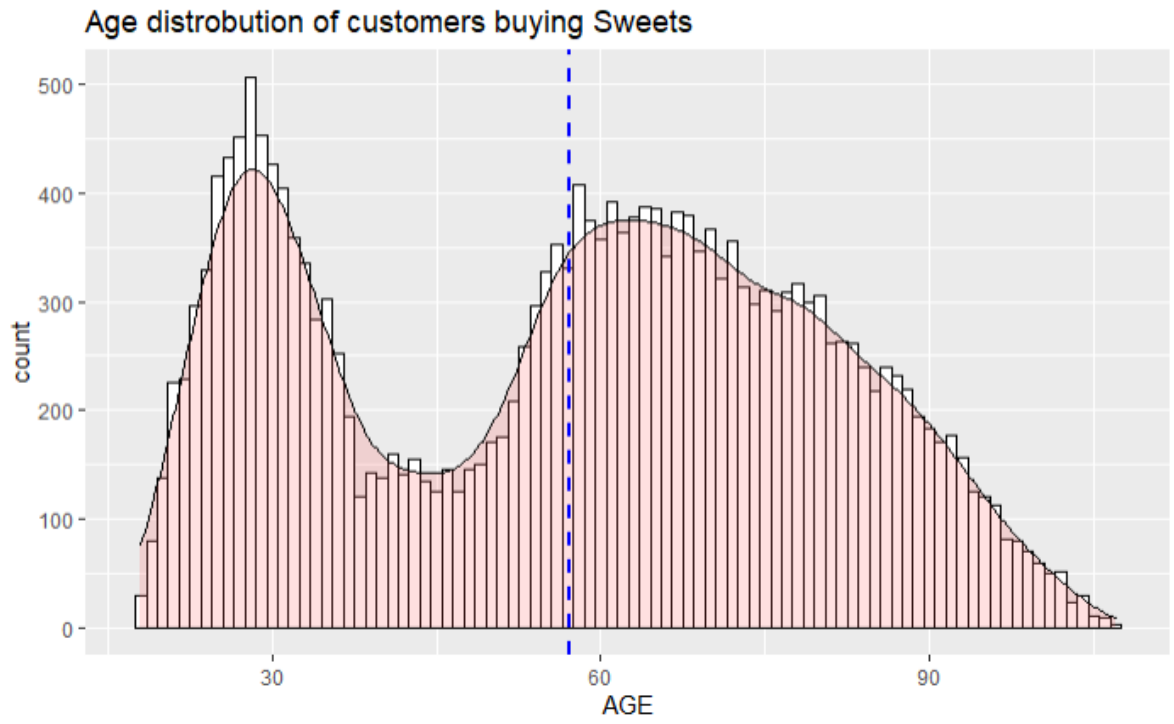


Figure 19: Age distribution of sweets

The customers purchasing sweets from the online business age distribution has a bimodal distribution (Bimodal, 2022). This indicates that the data has two modes and local maxima. The data has a first local maxima between the age groups of 25 and 35. The second local maxima is between the age groups of 55 and 75. The company should investigate why the frequency of sweets purchases decreases after the age of 35 and increases around the age of 55. The frequency of purchase decreases at a less rate after the 2nd modes in comparison with the 1st modes. The blue line indicates the mean age of customers buying sweets

Age distribution of customers buying technology

R output:

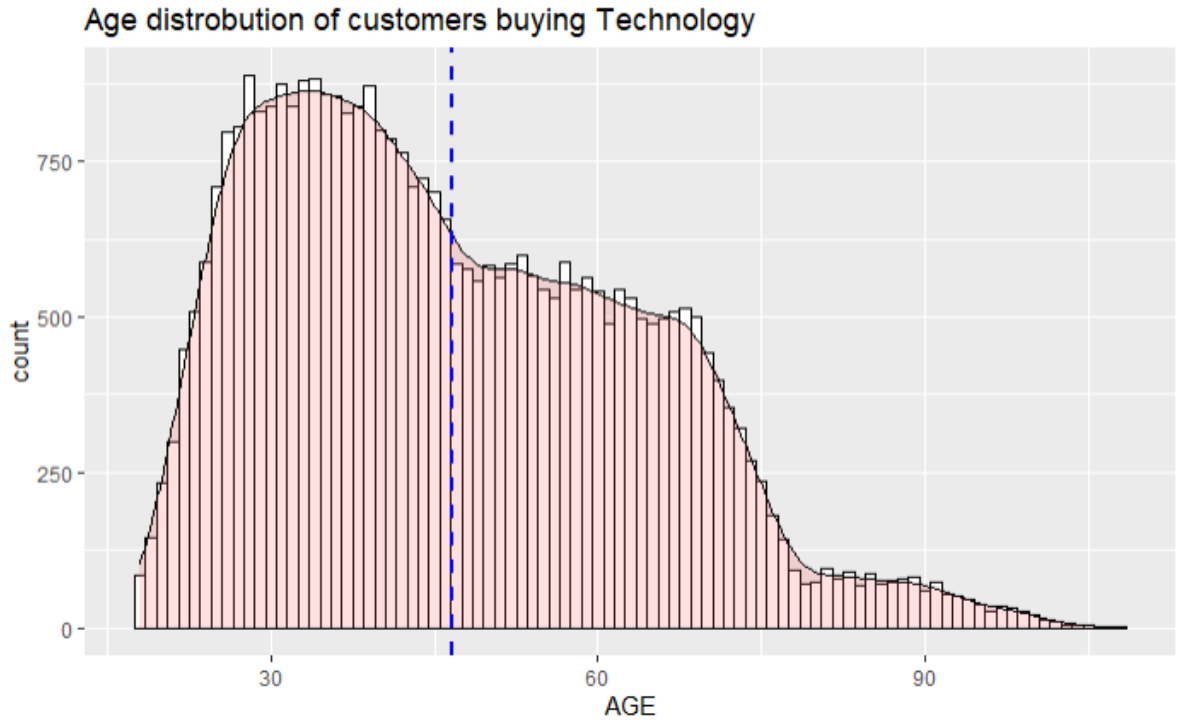


Figure 20: Age distribution of technology

The customers purchasing technology from the online business age distribution is skewed to the right. This indicates that there is a higher concentration of data to the left region of the x axis. This indicates that a younger demographic of clients buys technology more frequently. The data has a fat tail to the right side. Which indicates that the frequency of customers buying technology decreases slightly as age increases (there is not a significant drop in frequency between the ages of 40 and 70). After the age of 70 years, there is a significant drop of in purchase frequency. We can assume that after the age of 70 years clients are less likely and interested in buying technology more frequently. Clients after the age of 70 do not buy technology at a rate similar to other products in the different classes. The blue line indicates the mean age of customers buying clothing

Process Capability indices

A USL of 24 and LSL of 0 is given. A LSL (Lower spesification limit) of 0 is logical in this context.

The Process Capability indices were calculated and rounded to 3 significant digits.

$$Cp = \frac{(USL - LSL)}{6\sigma}$$

$$Cpu = \frac{(USL - \mu)}{3\sigma}$$

$$Cpl = \frac{(\mu - LSL)}{3\sigma}$$

$$Cpk = \min(Cpl, Cpu)$$

(Six Sigma Study Guide, 2022)

The process cpability indices where calculated in R. The R output is as follow:

```
[1] "The Cp  value is:  1.14"
[1] "The Cpu value is:  0.38"
[1] "The Cpl value is:  1.9"
[1] "The Cpk value is:  0.38"
```

Figure 21: R output - Process Capability indices

Cp value

The Cp value is a measurement of the potential capability of a process. The process has a potential capability of 1.14. This indicates the capability of the process if all drifts and process shifts were eliminated.

Cpu value

The Cpu value is the potential of the process based on its upper specification limit, which is 0.38. The process is not capable in the upper tail of its distribution. Improvement is necessary

Cpl value

The Cpl value is the capability of the of the process based on its lower limits. The Cpl of the process is 1.9. This indicates that the process is capable in the lower tail of its distribution.

Cpk value

The calculated Cpk value is 0.38. It is less than 1. Thus, we can assume that the process is not capable. Since the Cpk value is poor. The product can be classified as bad because of the low Cpk value.

Part 3: Statistical process Control (SPC)

The purpose of statistical process control is to monitor and control outliers and problems that could influence the behaviour of a system. This is possible by using different tools, techniques and procedures that plot section control charts that will plot the sales data on a plot with three reference lines and sectors. Firstly, the UCL – the upper control limit, secondly the CL – Centreline and lastly the LSL – the lower control limit. A control chart will be plotted of each class.

Table Analysis:

The delivery time of all the classes is plotted in the control charts. To plot the data accurately the dates should be order in ascending order. In other words. The oldest data first. The following tables contain the values of the X & S charts.

Table: X-Chart

The following figures and tables contain the UCL, U2Sigma, U1Sigma, CL, L1Sigma, L2Sigma, LCL values of all the X – Charts for the different classes

X-Chart							
Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22.97462	22.10789	21.24117	20.37444	19.50772	18.641	17.77427
Clothing	9.404934	9.259956	9.114978	8.97	8.825022	8.680044	8.535066
Household	50.24833	49.01963	47.79092	46.56222	45.33352	44.10482	42.87612
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

Table 1: X-chart

R output:

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22.974616	22.107892	21.241168	20.374444	19.507721	18.640997	17.774273
Clothing	9.404934	9.259956	9.114978	8.970000	8.825022	8.680044	8.535066
Household	50.248328	49.019626	47.790924	46.562222	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	2.490000	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

Figure 22: X-chart

Table: S-Chart

The following figures and tables contain the UCL, U2Sigma, U1Sigma, CL, L1Sigma, L2Sigma, LCL values of all the S – Charts for the different classes

S-Chart							
Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.18057	4.552222	3.923875	3.295528	2.667181	2.038833	1.410486
Clothing	0.86656	0.761455	0.656351	0.551247	0.446142	0.341038	0.235934
Household	7.34418	6.45341	5.56264	4.67187	3.7811	2.89033	1.999561
Luxury	1.511052	1.327778	1.144503	0.961229	0.777955	0.59468	0.411406
Food	0.437247	0.384213	0.33118	0.278147	0.225113	0.17208	0.119047
Gifts	2.246333	1.973877	1.701421	1.428965	1.156509	0.884053	0.611597
Sweets	0.835339	0.734022	0.632704	0.531386	0.430069	0.328751	0.227433

Table 2: S-chart

R output:

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

Figure 23: S-chart

The UCL (Upper control limit) is three-Sigma standard deviations above the mean. The U2Sigma control limit is two standard deviations above the mean. The U1Sigma control limit is one standard deviations above the mean. The CL (Centre line) represents the mean of the samples of the class. The L1Sigma control limit is one standard deviations below the mean or centre line. The L2Sigma control limit is two standard deviations below the mean or centre line. The LCL (Lower control limit)) is three-Sigma standard deviations below the mean or centre line.

Initialization Example of a X-Chart:

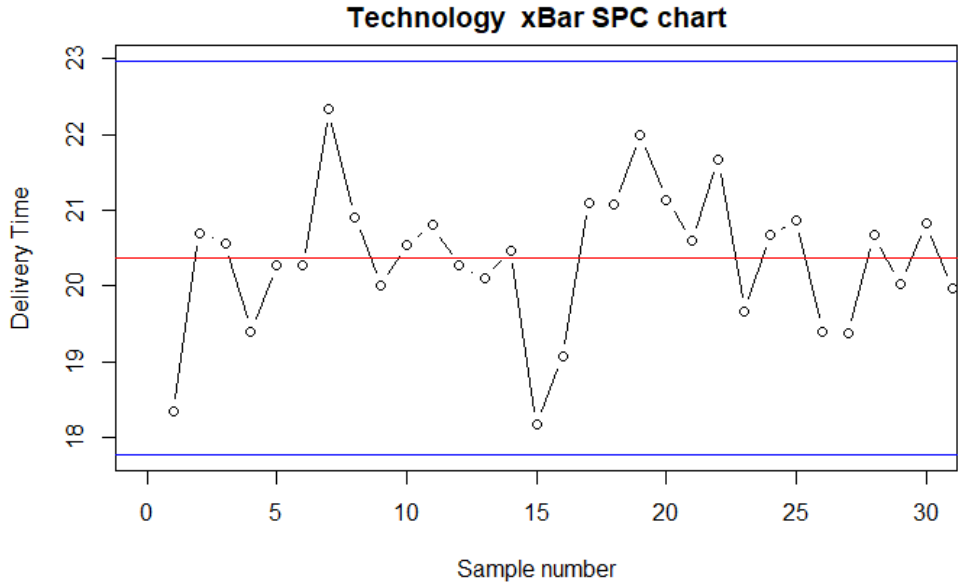


Figure 24: Tech x bar 30 samples

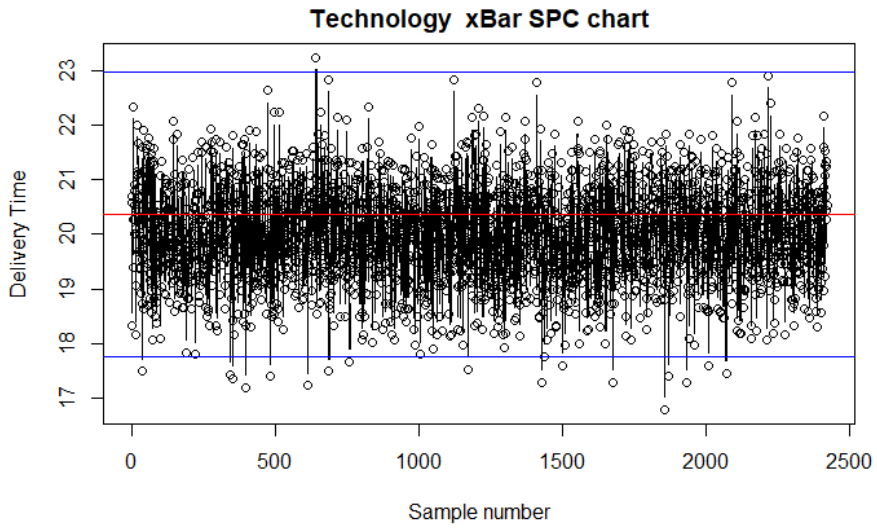


Figure 25: Tech x bar 2500 samples

Initialization Example of a S-Chart:

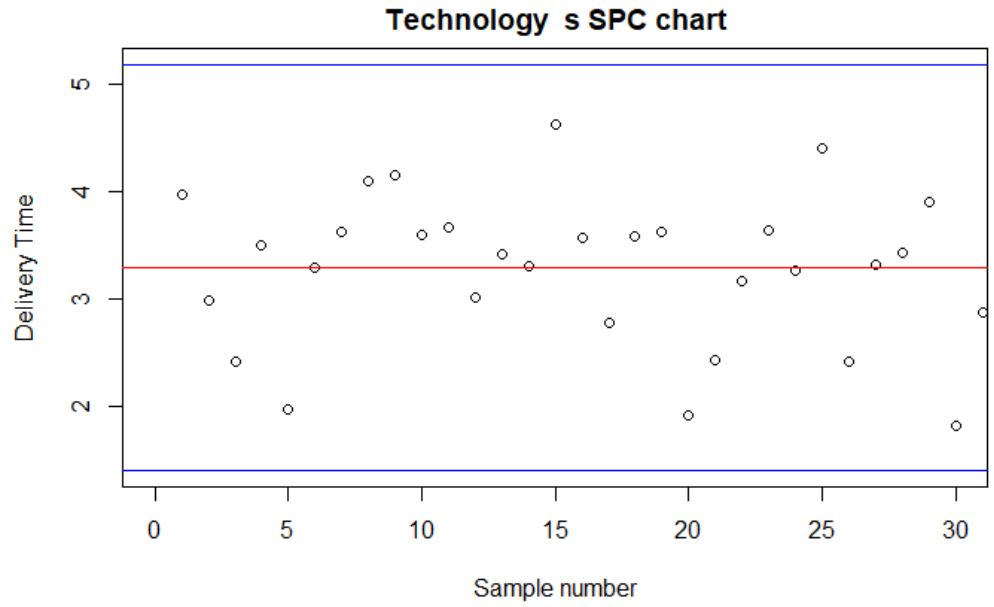


Figure 26: S chart 30 samples

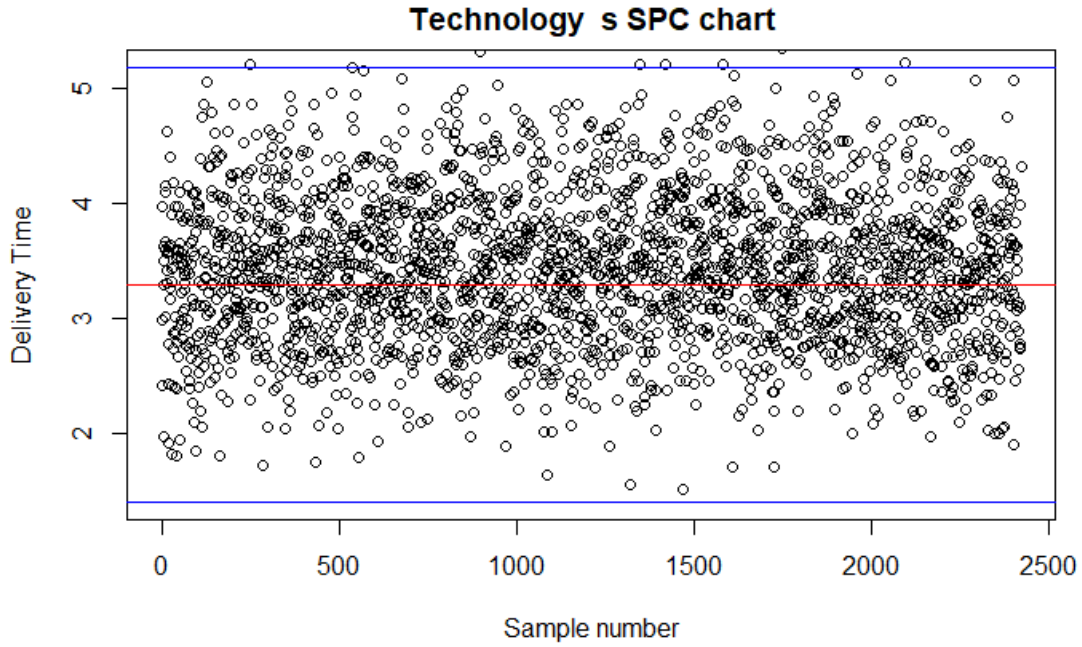


Figure 27: S chart 2500 samples

Part 4: Optimisation of the delivery process

All instances where sample numbers give indication to out of control instance will be plotted and listed.

Part 4.1a: 1 X-Bar outside control limits

The instances where its falls outside 1 X-bar or samples means outside of the outer control limits will be plotted and listed. In the case where many instances are possible, the first 3 and last three 3 numbers will be given.

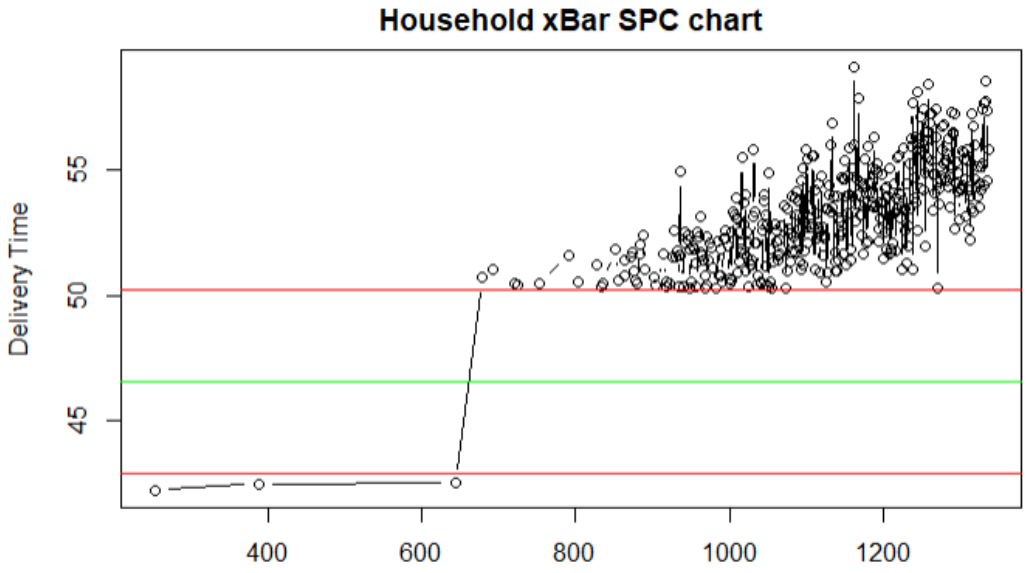


Figure 28: Outside 1 X-Bar Household

The plot plots all instances, where points fall 1 X-Bar outside the delivery time outer control limits of the household products. The first three instances have a Delivery time of 42.23333, 42.46667, 42.50000 and the last three have a Delivery time of 57.36667 54.56667 55.80000. There is a total of 395 instances that fall 1 X-Bar outside the control limits.

List of Delivery times that fall out of the 1 – X Bar control limits:

- First Three: 42.23333, 42.46667, 42.50000
- Last Three: 57.36667, 54.56667, 55.80000
- Total: 395

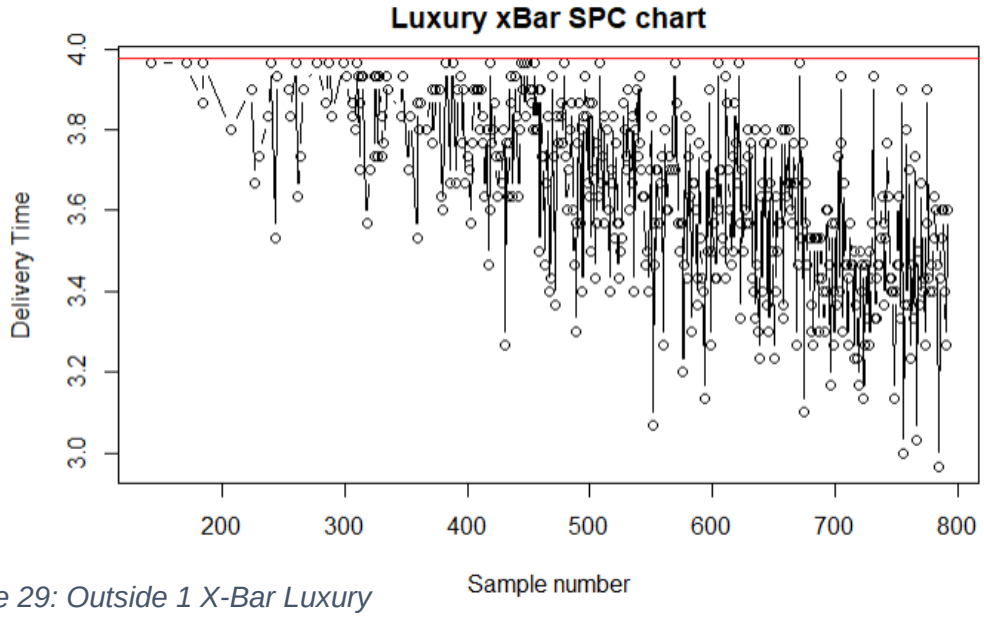


Figure 29: Outside 1 X-Bar Luxury

The plot plots all instances, where points fall 1 X-Bar outside the delivery time outer control limits of the luxury products. The first three instances have a Delivery time of 3.400000, 3.266667, 3.600000 and the last three have a Delivery time of 3.966667, 3.966667, 3.866667. There is a total of 440 instances that fall 1 X-Bar outside the control limits.

All instances of data outside the control limits, are all below the lower control limit.

List of Delivery times that fall out of the 1 – X Bar control limits:

- First Three: 3.400000, 3.266667, 3.600000
- Last Three: 3.966667, 3.966667, 3.866667
- Total: 440

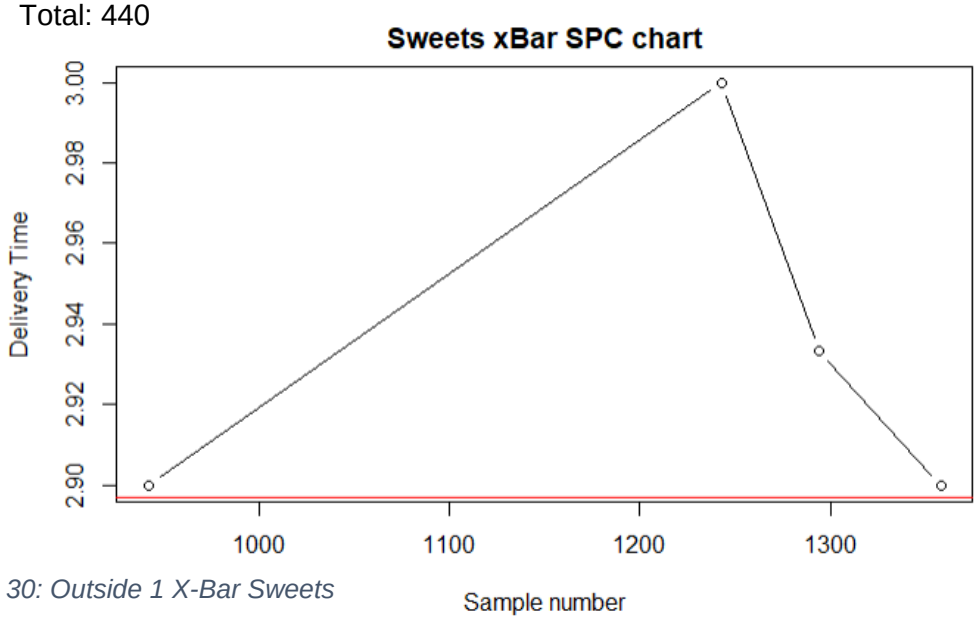


Figure 30: Outside 1 X-Bar Sweets

The plot plots all instances, where points fall 1 X-Bar outside the delivery time outer control limits of the sweets. There are only 4 instances of data breaching the control limits and fall 1 X-Bar outside the control limits.

All instances of data outside the control limits, are above the upper control limit.

List of Delivery times that fall out of the 1 – X Bar control limits:

- Values: 2.900000, 3.000000, 2.933333, 2.900000
- Total: 4

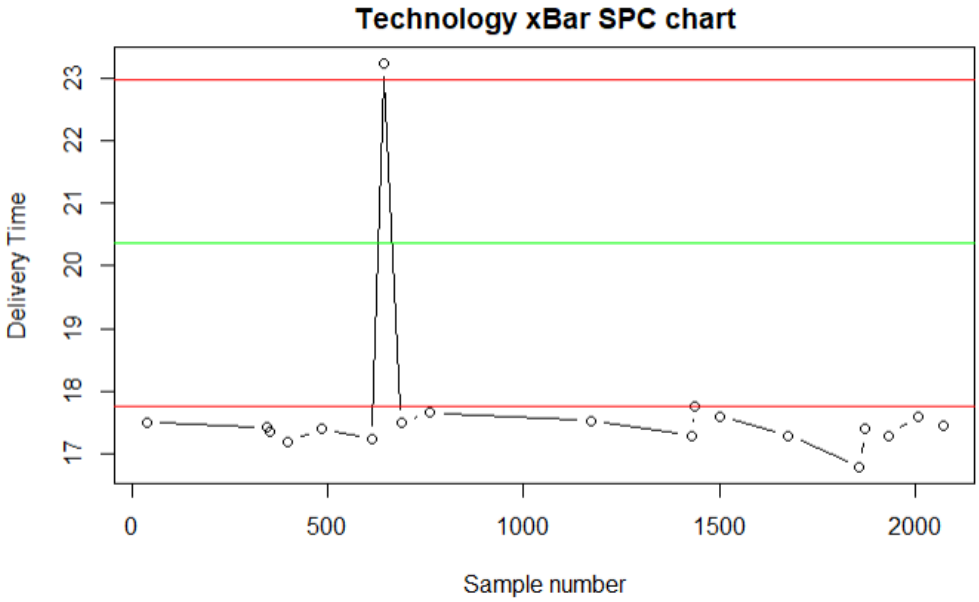


Figure 31: Outside 1 X-Bar Technology

The plot plots all instances, where points fall 1 X-Bar outside the delivery time outer control limits of Technology. The first three instances have a Delivery time of 17.50000, 17.43333, 17.36667 and the last three have a Delivery time of 17.30000 17.60000 17.46667. There is a total of 19 instances that fall 1 X-Bar outside the control limits.

There is one instance of data outside the upper control limit. There 18 instances below the lower control limit.

List of Delivery times that fall out of the 1 – X Bar control limits:

- First Three: 17.50000 17.43333 17.36667
- Last Three: 17.30000 17.60000 17.46667
- Total: 19

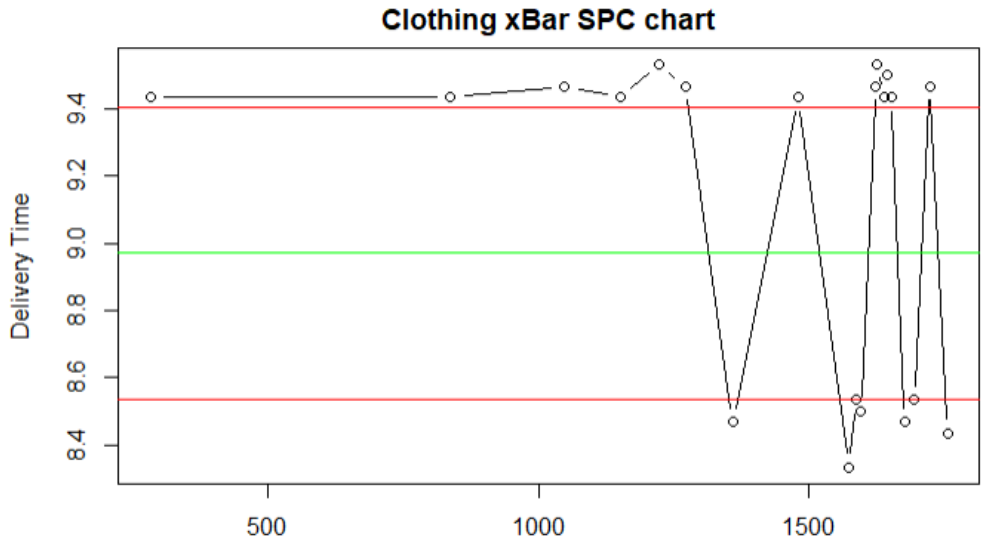


Figure 32: Outside 1 X-Bar Technology

The plot plots all instances, where points fall 1 X-Bar outside the delivery time outer control limits of clothing. The first three instances have a Delivery time of 9.433333, 9.433333, 9.466667 and the last three have a Delivery time of 8.533333, 9.466667, 8.433333. There is a total of 20 instances that fall 1 X-Bar outside the control limits.

There are 13 instances of data outside the upper control limit. There 7 instances below the lower control limit.

List of Delivery times that fall out of the 1 – X Bar control limits:

- First Three: 9.433333, 9.433333, 9.466667
- Last Three: 8.533333, 9.466667, 8.433333
- Total: 20

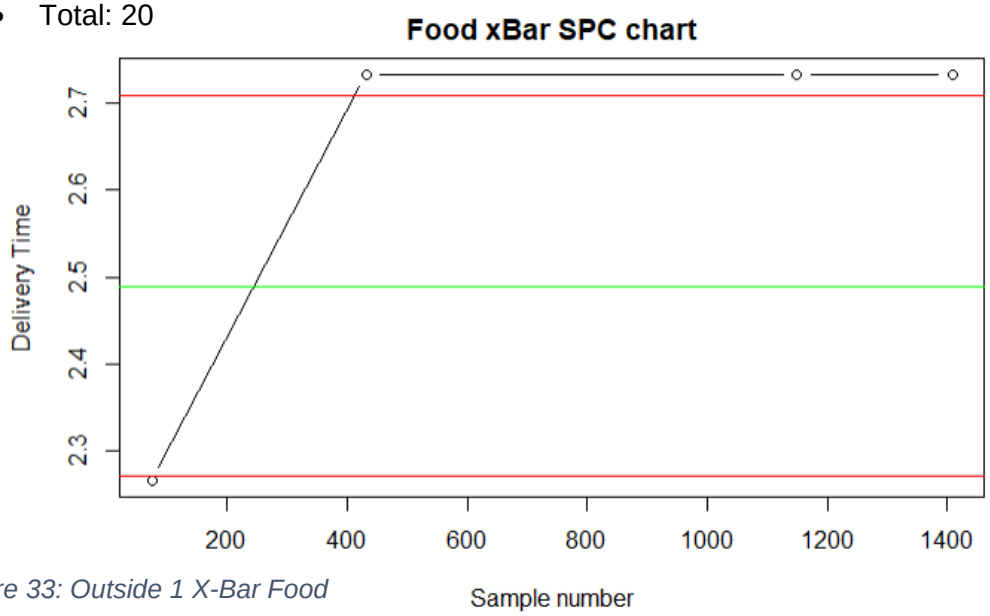


Figure 33: Outside 1 X-Bar Food

The plot plots all instances, where points fall 1 X-Bar outside the delivery time outer control limits of food. There are only 4 instances of data breaching the control limits and fall 1 X-Bar outside the control limits.

Three instances of data outside the control limits, are above the upper control limit. And one below the lower control limit

List of Delivery times that fall out of the 1 – X Bar control limits:

- Values: 2.266667, 2.733333, 2.733333, 2.733333
- Total: 4

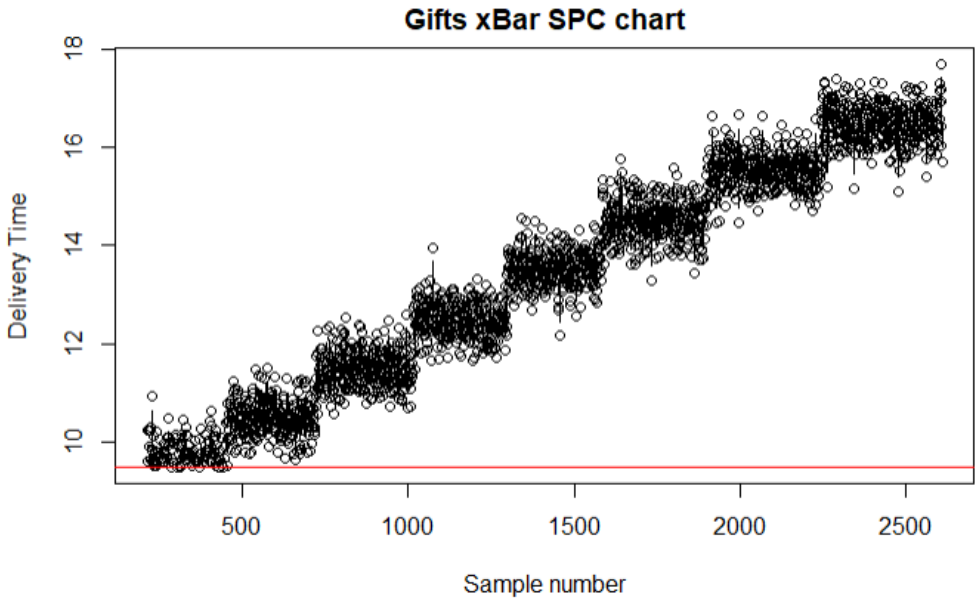


Figure 34: Outside 1 X-Bar Gifts

The plot plots all instances, where points fall 1 X-Bar outside the delivery time outer control limits of gifts. The first three instances have a Delivery time of 10.23333, 9.60000, 9.90000 and the last three have a Delivery time of 16.03333, 16.93333, 15.70000. There is a total of 2287 instances that fall 1 X-Bar outside the control limits.

There is an abnormal number of values falling outside the control limits, this must be investigated further.

List of Delivery times that fall out of the 1 – X Bar control limits:

- First Three: 10.23333, 9.60000, 9.90000
- Last Three: 16.03333, 16.93333, 15.70000
- Total: 2287

Part 4.1b: Most consecutive sample of s-bar between -0.3 and + 0.4

The instance with the most consecutive samples of s-bar or sample standard deviations needs to be identified between -0.3 and +0.4 sigma control limits and the ending sample number.

R Output:

```
[1]    6 1776
[1]   4 223
[1]  3 45
[1]  4 63
[1]   5 756
[1]    7 2477
[1]  4 94
```

Figure 35: Most Consecutive Samples and Ending sample number

Class	Max Count	Ending Number
Technology	6	1776
Clothing	4	223
Household	3	45
Luxury	4	63
Food	5	756
Gifts	7	2477
Sweets	4	94

Table 3: Most consecutive samples and ending sample number

Technology's most consecutive samples of sample standard deviations is 6. The ending sample number is 1776.

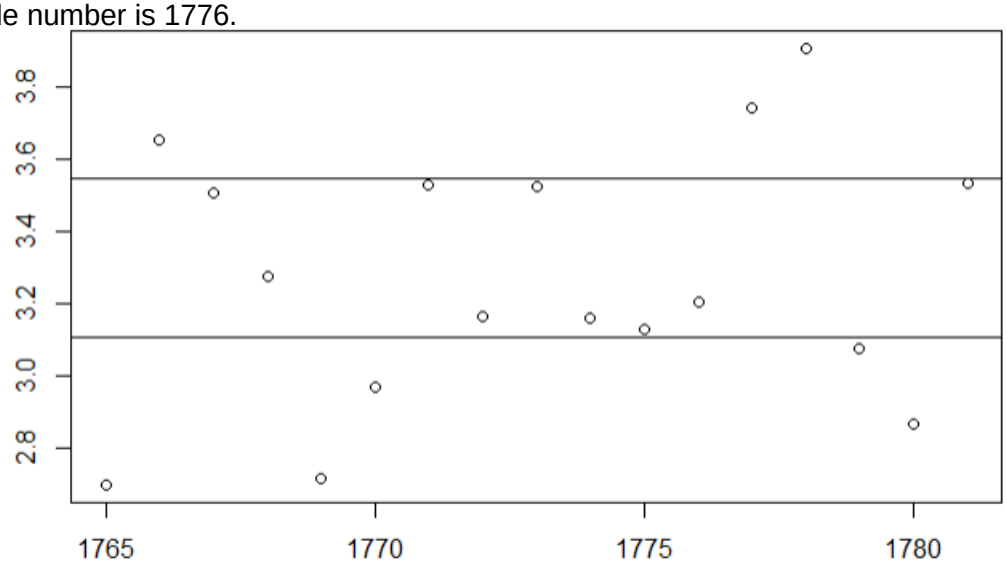


Figure 36: Representation of technology 4.1b

Clothing's most consecutive samples of sample standard deviations is 4. The ending sample number is 223

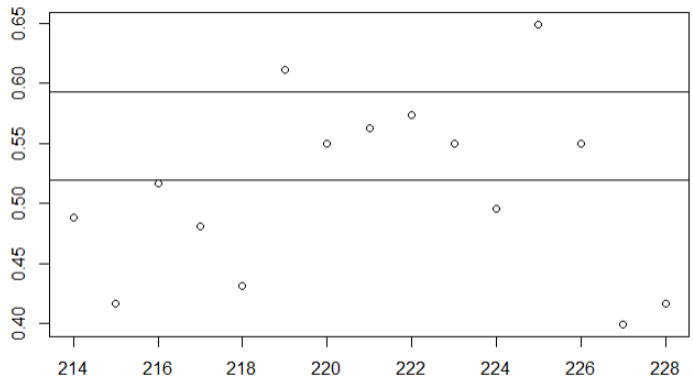


Figure 37: Representation of Clothing 4.1b

Household's most consecutive samples of sample standard deviations is 3. The ending sample number is 45

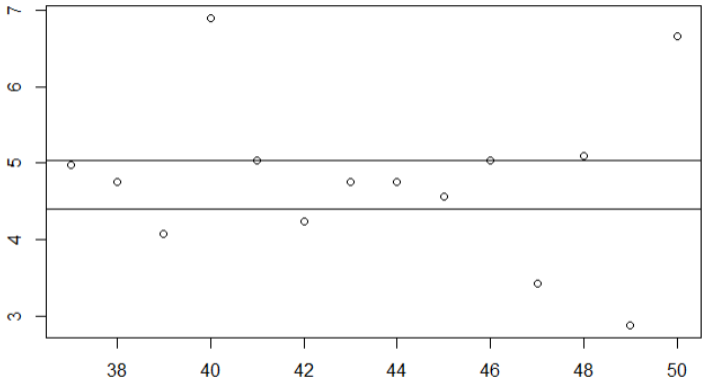


Figure 38: Representation of household 4.1b

Luxury's most consecutive samples of sample standard deviations is 4. The ending sample number is 63

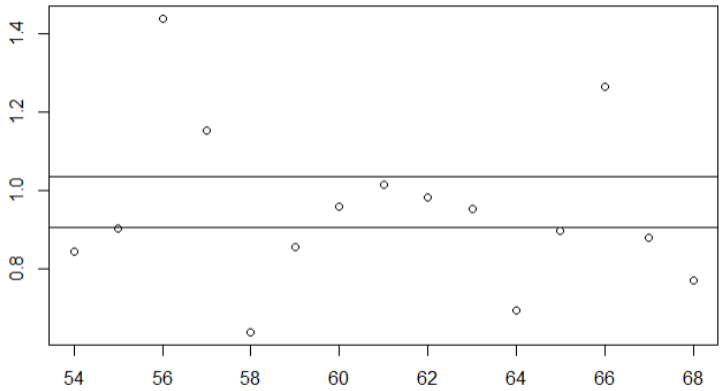


Figure 39: Representation of luxury 4.1b

Food's most consecutive samples of sample standard deviations is 5. The ending sample number is 756

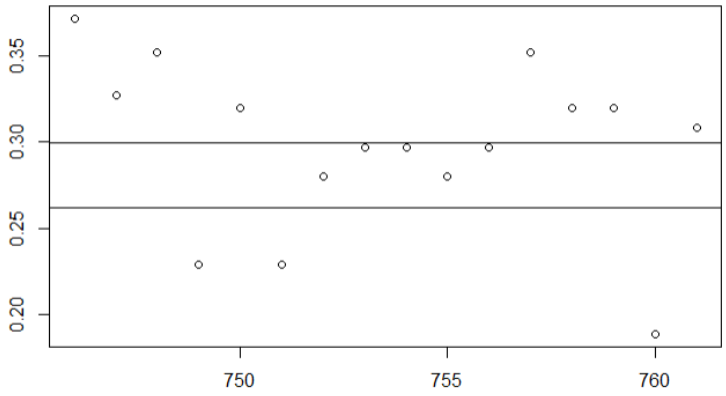


Figure 40: Representation of Food 4.1b

Gifts' most consecutive samples of sample standard deviations is 7. The ending sample number is 2477

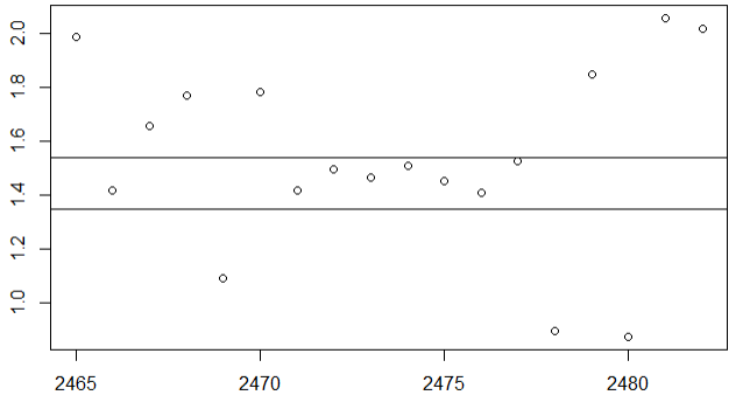


Figure 41: Representation of gift 4.1b

Sweets' most consecutive samples of sample standard deviations is 4. The ending sample number is 94

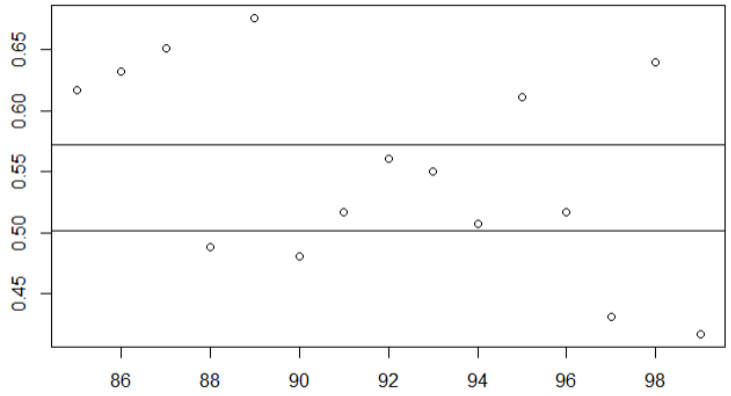


Figure 42: Representation of sweets 4.1b

Part 4.2: Type I error for A and B

A type I Error is when the null hypothesis is rejected but was true. In other words, it is the probability that the business rejects the null hypothesis (which is in control and centred) but is wrongly rejected.

Question 4	Error Type I
A	0.27%
B	72.67%

Table 4: Type I Error

In the case of A, the probability for making a Type I error is 0.27% when the samples are outside the outer control limits. The probability is exceedingly small. It is not likely making this error. (Falsely rejecting that a sample is outside the outer control limits when the null hypothesis is true)

In the case of B, the probability of making a Type I error is 72.67%.

Thus, we can state, if method B is used to determine if a process is in control, it will not be accurate. There is a probability of 72.67% for a type I error. Thus, stopping the process was not necessary 72.67% of the time.

Method A is a better and more accurate method to determine if the process is in control.

Part 4.3: Centre the delivery process

The optimal delivery times should be determined. All individual delivery times are considered.

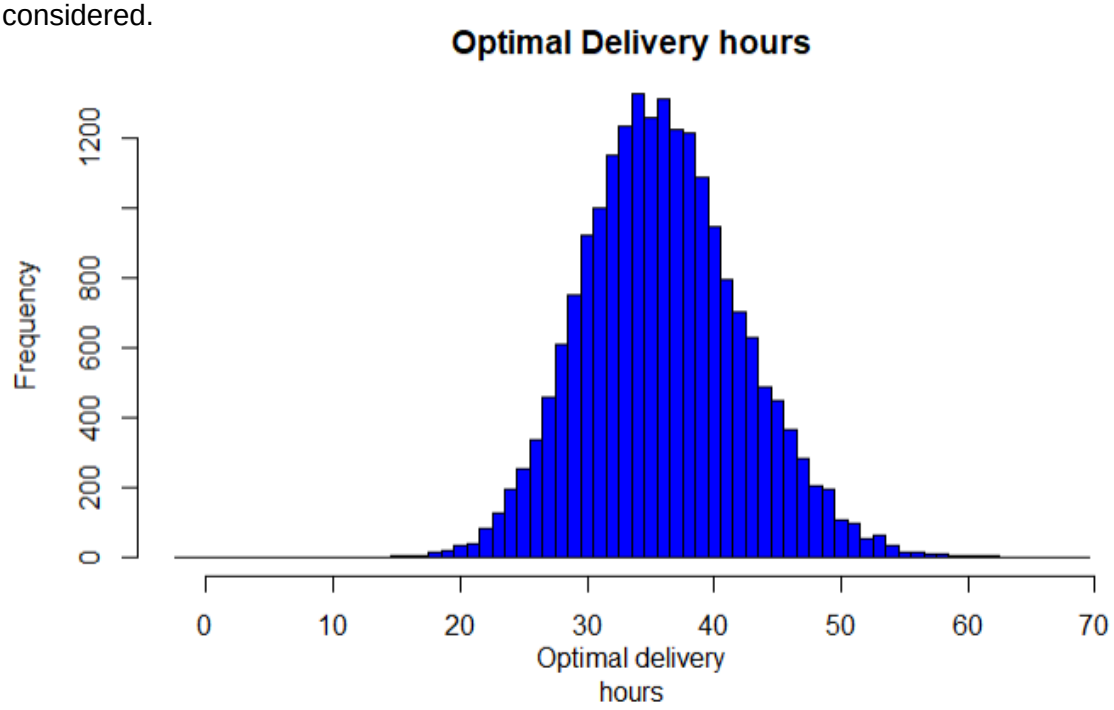


Figure 43: Optimal delivery hours

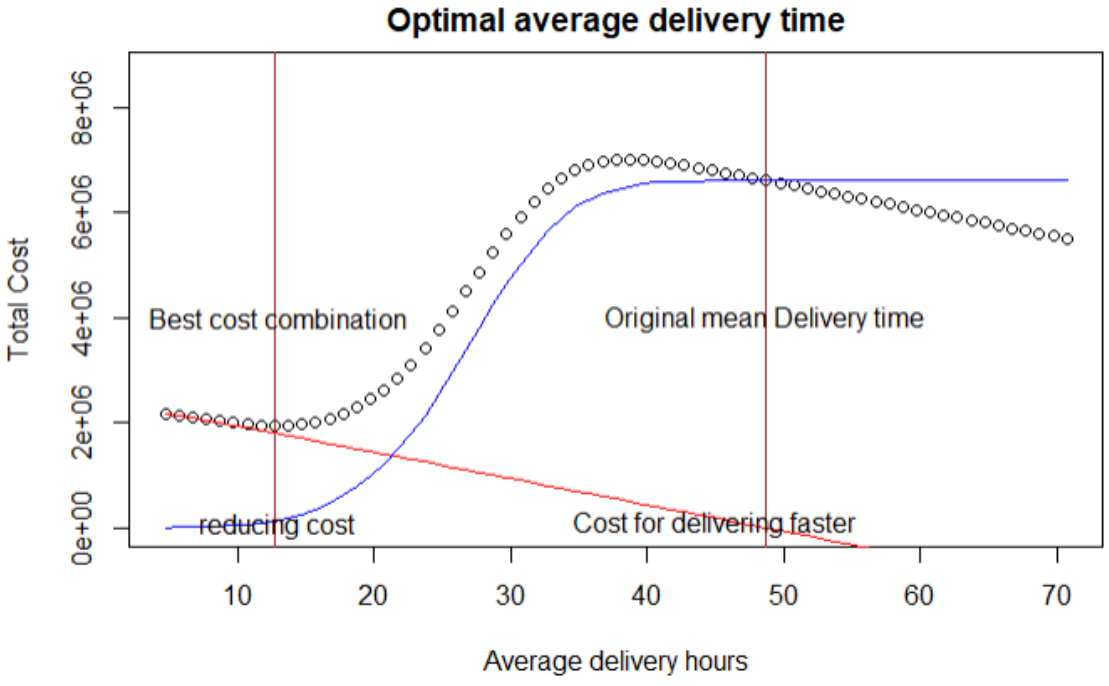


Figure 44: Optimal delivery time distribution

The optimal delivery hours to centre the delivery process is 12.71859 hours. This will ensure the maximisation of profit.

Part 4.4: Likelihood of making a Type II error for A

A type II Error is a false negative. When the null hypothesis when the process is in control and cantered is not rejected, but it should be. In other words, the process is assumed to be in control and cantered when it is not, the null hypothesis should be rejected.

The probability of A in class technology making a Type II Error is 48.83177%

R output:

```
[1] 0.4883177
```

This is prominent level of error type II.

Part 5: DOE and MANOVA

The results of part 2, 3 and 4 is consulted. The MANOVA is as follows:

R output:

```
              Df Pillai approx F num Df den Df    Pr(>F)
Class          6 1.6797   157291     12 359942 < 2.2e-16 ***
Residuals 179971
---
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 5.7168e+13 9.5281e+12   80258 < 2.2e-16 ***
Residuals   179971 2.1366e+13 1.1872e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Delivery.time :
               Df   Sum Sq Mean Sq F value    Pr(>F)
Class           6 33458565 5576427  629429 < 2.2e-16 ***
Residuals   179971  1594452       9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 45: MANOVA

Ho: Class has no impact on the price or the delivery time of the products

Ha: Class has an impact on price and the delivery time of the products

After the hypothesis test, the null hypothesis was rejected. The class of a product has an impact on the price and the delivery time of the product.

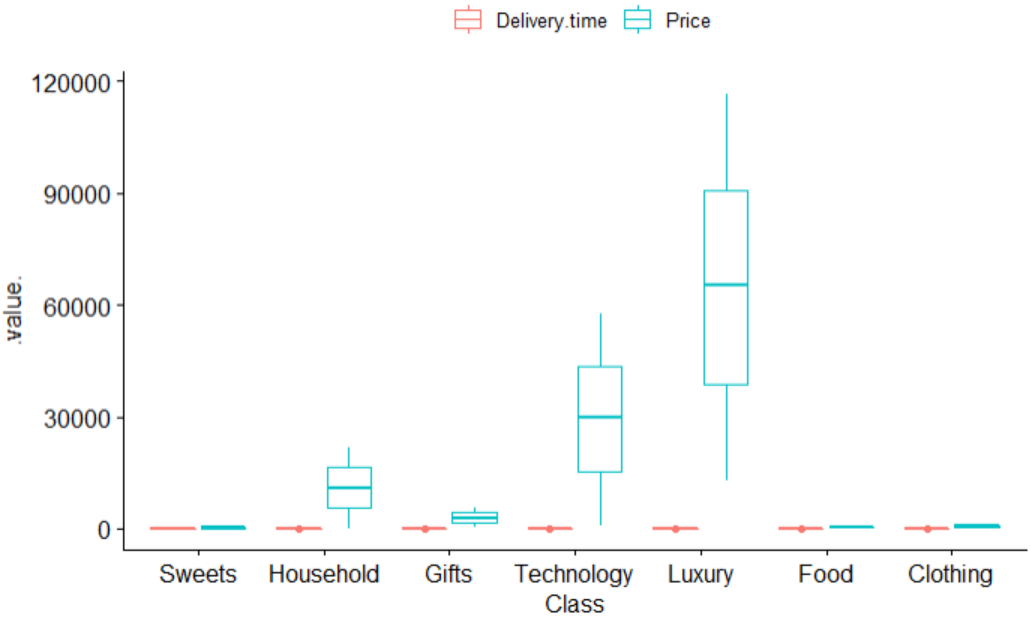


Figure 46: MANOVA results

The box plot visualizes the relation between price and class. The relation between delivery times is not viable. Due to the y axis of the graph not being compatible.

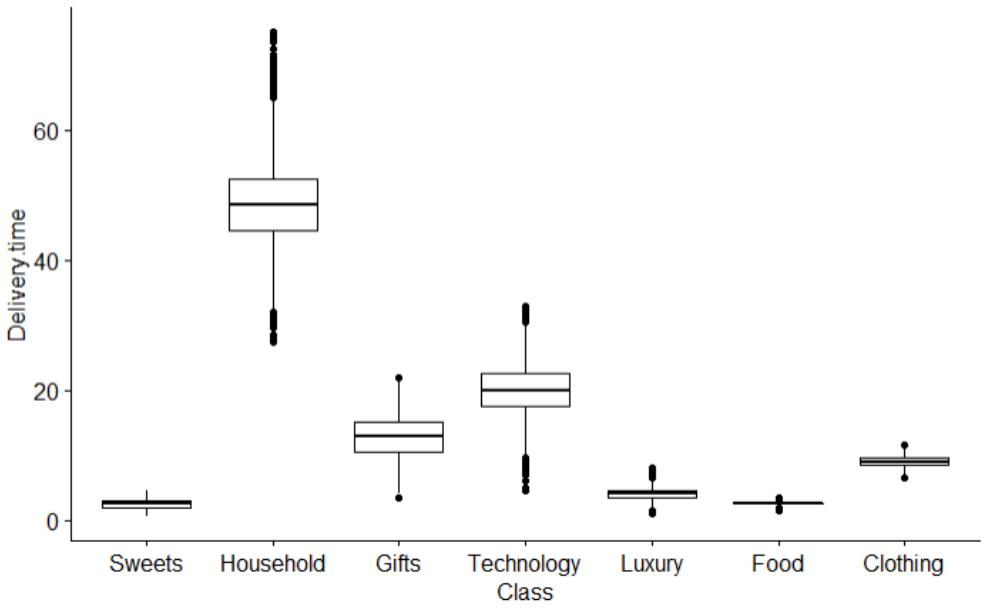


Figure 47: Delivery vs Class

Delivery time is visualized more accurately in this graph. It is clear that the class of the product has a significant impact on the delivery times of the product.

The graphs further support that the null hypothesis was rejected. It is clear that the class of a product has a significant impact on the price and delivery time of a product.

Part 6:

6.1 Results of subsidiary Lafideradora

6. Taguchi Loss Function

Cost to scrap: $45

Specification of thickness: 0.06+-0.04

$$L = k(y - m)^2$$

$$k = \frac{L}{(0.06 + 0.04 - 0.06)^2}$$

$$k = \frac{45}{0.04^2}$$

$$L = 28125(y - 0.06)^2$$

R Output:

```
[1] "The Tagushi loss function for 6 is: L = 28125 (y-0.06)^2"
```

7a. Cool Food, inc. Taguchi loss function

While continuing to find the root cause of the scrap, they found a way to reduce the price of scrap

Cost to scrap: $35

Specification of thickness: 0.06+-0.04

$$L = k(y - m)^2$$

$$k = \frac{L}{(0.06 + 0.04 - 0.06)^2}$$

$$k = \frac{35}{0.04^2}$$

$$L = 21875(y - 0.06)^2$$

R Output:

```
[1] "The Tagushi loss function for 7 is: L = 21875 (y-0.06)^2"
```

7b. If the process deviation is reduced to 0.027cm what is the Taguchi loss?

$$L = k(y - m)^2$$

$$L = 21875(0.27)^2$$

$$L = 15.946875$$

R Output:

```
[1] "The Tagushi loss is: L = 15.946875"
```

6.2 Results of subsidiary Magnaplex

Magnaplex has a complex Manufacturing process with three operations that are performed in series. The three machines' reliabilities are listed below.

Machine	Reliability
A	0.85
B	0.92
C	0.90

Table 5: Reliability of Machines

a. System reliabilty (one machine at a stage)

$$Ra * Rb * Rc = 0.85 * 0.92 * 0.90$$

$$Ra * Rb * Rc = 0.7038$$

R Output:

```
[1] "Ra*Rb*Rc = 0.7038"
```

b. How much is Reliability improved by having two machines at each stage

$$Raa = 1 - (1 - 0.85)^2$$

$$Rbb = 1 - (1 - 0.92)^2$$

$$Rcc = 1 - (1 - 0.90)^2$$

$$Raa * Rbb * Rcc = (1 - (1 - 0.85)^2) * (1 - (1 - 0.92)^2) * (1 - (1 - 0.90)^2)$$

$$Raa * Rbb * Rcc = 0.96153156$$

R Output:

```
[1] "Raa*Rbb*Rcc = 0.96153156"
```

6.3 Delivery Process

We can expect an estimated reliable delivery times 294.964 days per year with 20 vehicles. If the number of vehicles is increased by 1 to 21 vehicles. Then we can expect reliable delivery times on 346.7182 days per year. There is a substantial increase.

R output:

```
[1] 294.964
[1] 346.7182
```

Conclusion

This report has analysed and documented the data wrangling processes of the sales data of an online business. All invalid data types were removed from the data set. Important averages of the data set were summarised and plotted. Assumptions were made based on each summary. Each classes buy methods were plotted and established. It is clear that the online business main contribution in sales was recommendation across all classes. The age distributions of all classes were plotted. The target market for each class was identified.

X & S charts were plotted and put in the correct corresponding tables. All instances of data falling outside the outer limits of the control limits were also plotted and noted. Type I and Type II errors of question 4.1a and 4.1b was calculated. MANOVA's were performed. This indicates what descriptive features are dependent on one another within the dataset. The conclusion was made that the class of the product influences the price and delivery time of the product. Lastly, the reliability of subsidiaries was determined.

References

Statistics.com: Data Science, Analytics & Statistics Courses. 2022. *Bimodal*. [online] Available at: <https://www.statistics.com/glossary/bimodal/> [Accessed 18 October 2022].

Hessing, T. (2014) *Process capability (cp & Cpk)*, *Six Sigma Study Guide*. Available at: <https://sixsigmastudyguide.com/process-capability-cp-cpk/> [Accessed: October 18, 2022].