ECSA GRADUATE ATTRIBUTE PROJECT

Quality Assurance 344

Miss C Whiteboyd
Stellenbosch University
23959479

Table of Contents

Table of Figures

List of Tables

Introduction

A dataset containing information pertaining to the sales of an online business is given and is analysed through this report. This analysis will allow for the understanding and extraction of useful information. There are many steps to this analysis, beginning with data preparation. Data preparation is the process of cleaning the data by removing invalid entries and ordering the data chronologically. Once the valid data is prepared descriptive statistics methods are applied to gain a better insight into the meaning of the data. Thereafter the process capability indices are calculated in order to measure the ability of process to produce within certain specification limits. Following this, the x-bar and s Statistical Process Control charts along with the respective graphs are constructed. Further analysis into the graphs is done to determine whether the process is out of control. In order to make specific conclusions about the delivery process, different aspects of it are discussed. A MANOVA test is performed to the determine the effect of the feature “Class” of items on “Age”, “Price” and “Delivery Time”. The aim of this analysis is to provide the business with information it can use to improve their current processes.

1. Data Wrangling

Client data for an online business is given in the dataset, 'SalesTable2022'. The dataset set does not contain complete data; therefore, it must be analysed before it can be used. In the analysis, the data must be cleaned and any invalid data should be removed. This results in two datasets, namely a Valid and an Incomplete dataset. The incomplete dataset contains missing values and negative values.

1.1 Incomplete Data

The incomplete dataset contains all the missing values ("NA") as well as the negative price values. This dataset is created to separate the valid, complete instances from the invalid ones, in order to make the analysis of valid data viable. The data is stored in such a way that it may be repeated quickly should new data be provided. The incomplete data is shown below.

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

Figure 1: Missing Values Dataset

1.2 Valid Data

The valid dataset is shown below. This dataset is used to perform accurate analysis on the online business. It is crucial that the incomplete data be removed before any analysis is performed.

	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	1	19966	54	Sweets	246.21	2021	7	3	1.5	Recommended
2	2	34006	36	Household	1708.21	2026	4	1	58.5	Website
3	3	62566	41	Gifts	4050.53	2027	8	10	15.5	Recommended
4	4	70731	48	Technology	41843.21	2029	10	22	27.0	Recommended
5	5	92178	76	Household	19215.01	2027	11	26	61.5	Recommended
6	6	50586	78	Gifts	4929.82	2027	4	24	14.5	Random
7	7	73419	35	Luxury	108953.53	2029	11	13	4.0	Recommended
8	8	32624	58	Sweets	389.62	2025	7	2	2.0	Recommended
9	9	51401	82	Gifts	3312.11	2025	12	18	12.0	Recommended
10	10	96430	24	Sweets	176.52	2027	11	4	3.0	Recommended
11	11	87530	33	Technology	8515.63	2026	7	15	21.0	Browsing
12	12	14607	64	Gifts	3538.66	2026	5	13	13.5	Recommended
13	13	24299	52	Technology	27641.97	2024	5	29	17.0	Browsing
14	14	77795	92	Food	556.83	2025	6	3	3.0	Random
15	15	62567	73	Clothing	347.99	2024	3	29	8.5	Website
16	16	14839	47	Technology	54650.41	2027	12	30	18.5	Recommended
17	17	96208	44	Technology	14739.09	2028	3	17	13.0	Recommended
18	18	39674	69	Technology	22315.17	2026	8	20	20.5	Recommended
19	19	98694	74	Sweets	546.48	2025	5	9	2.0	Recommended

Figure 2: Extract of Valid Dataset

2. Descriptive Analytics

This section is dedicated to the analysis of the valid data set. This analysis will provide insight into trends, accuracy and general observations in order to gain a better understanding of the sales of the online business.

2.1 Feature Overview

There are 10 features in the valid dataset. A brief description of each feature is provided below.

Name	Description
X	The index of the original instances in the 'salesTable' datafile
ID	Unique values to identify customers
AGE	Age of customer
Class	Category (department) the item belongs to
Price	Amount paid to buy item
Year	Year item was purchased in
Month	Month item was purchased in
Day	Day item was purchased on
Delivery.time	Number of days taken to deliver item to the customer
Why.Bought	Reason why customer purchased the product

Table 1: Feature Overview

In order to grasp the contents of the dataset a summary of the features age, price and delivery time are provided. The aspects that were used are minimum, maximum, median, 1st and 3rd quartile. Once a general understanding of the features is known they can be analysed in more detail.

Feature	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Age	18	38	53	54	70	108
Price	35.65	482.31	2259.63	12294.10	15270.97	116618.97
Delivery.time	0.5	3	20	14.5	18.5	75

Table 2: Summary of Age, Price, Delivery Time

2.2 Continuous Features

The valid dataset contains 8 continuous features, namely X, ID, AGE, Price, Year, Month, Day and Delivery.time.

Index

The index is not an important feature in the analysis of data. It should be noted that this is a uniform distribution as expected.

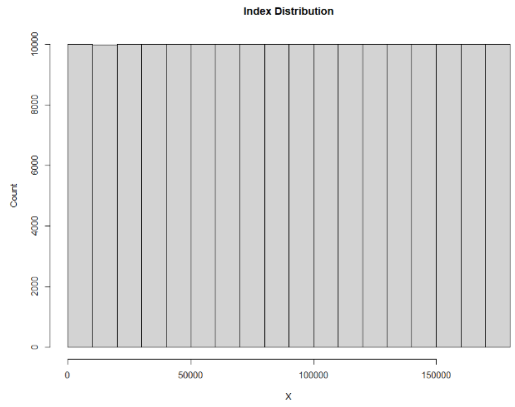

Figure 3: Index Distribution

ID

A similar reasoning is found for the ID feature in that it is not important in the analysis of data. It should be noted that ID follows a uniform distribution. The small deviations could be as a result of the incomplete data that was removed from the original dataset

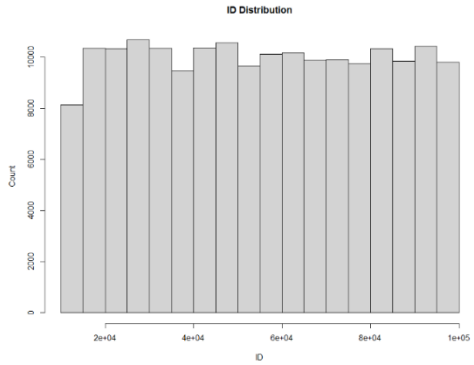

Figure 4: ID Distribution

Age

The feature age follows a skewed-right distribution. This indicates that the most frequent purchases are made by younger customers. It should also be noted that there is an age of 108 present. This is highly unlikely and could be as a result of an error when inputting the data. There are also very few purchases under the age of 20 which indicates that this age margin may not have a steady income, or rely on parents. This could be the reason for the higher density in the 25-45 age bracket.

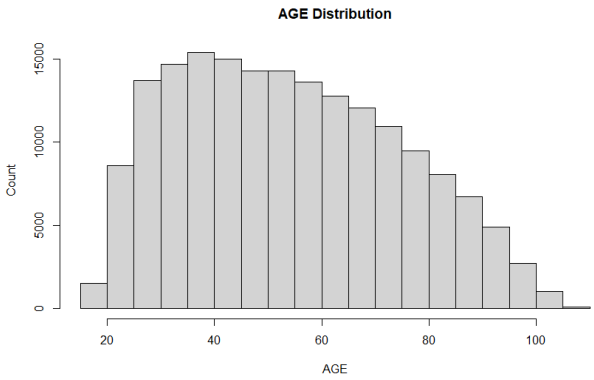

Figure 5: AGE Distribution

Price

The price feature follows an exponential distribution. This indicates that more of the cheaper products were purchased. This could provide information on whether it is worth it to sell products with very high prices.

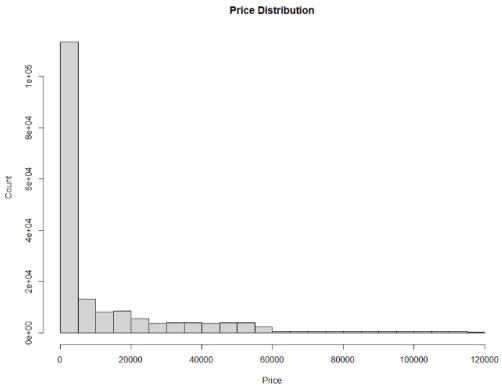

Figure 6: Price Distribution

Year

The graph below indicates that a large percentage of sales were made in 2021. It is seen that as the years progress, the amount of sales increase. This is a good trend to see as it shows that the business is improving sales each year.

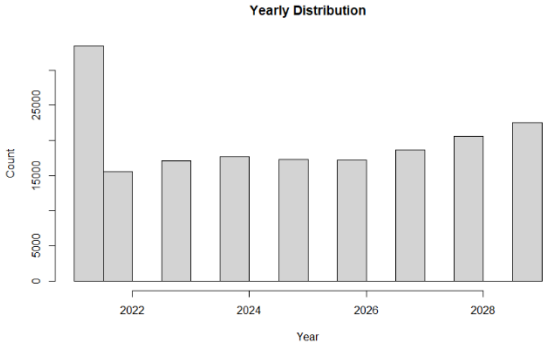

Figure 7: Yearly Distribution

Month

The monthly sales follow a uniform distribution, meaning the frequency of sales remains fairly consistent throughout the year. This is important because it shows that the sales are not seasonal.

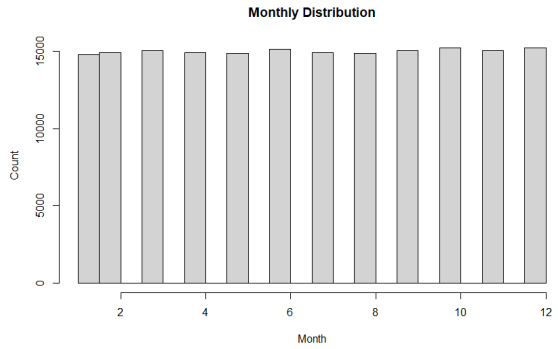

Figure 8: Monthly Distribution

Day

Similar conclusions can be made for the daily distribution. The daily distribution is uniform, meaning that the sales remain at a constant level throughout the month.

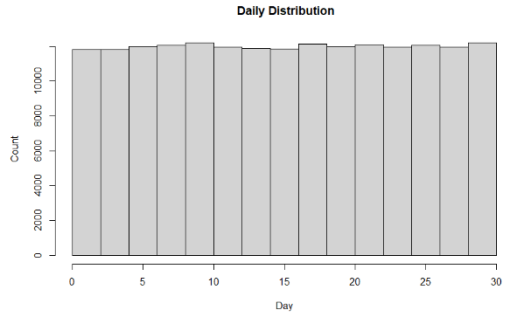

Figure 9: Daily Distribution

Delivery Time

The distribution seen in the delivery time feature is multimodal. This is different to the previous continuous features as it contains three peaks.

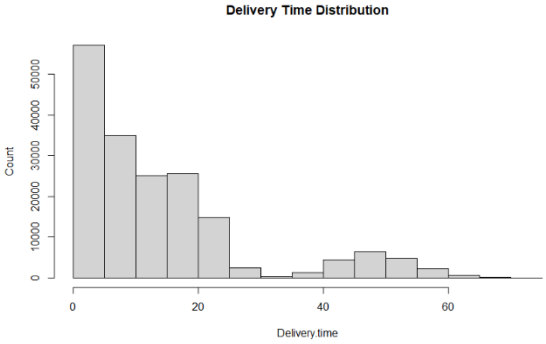

Figure 10: Delivery Time Distribution

2.3 Categorical Features

The valid dataset contains 2 categorical features, namely Class and Why.Bought.

Class Analysis

A graph of the number of items purchased per class as well as the income generated per class is shown below.

From Figure 11 it can be seen that Gifts and Technology classes have sold the most items and Luxury has sold the least items. It should be noted that although Luxury sold the least number of items, it generated the second highest income of the different classes. The opposite is true for Gifts. The class Gifts has one of the lowest generated incomes, along with Clothing, Food, Household and Sweets. This is useful because these results could be taken to the finance department so that a solution to the low profit can be found.

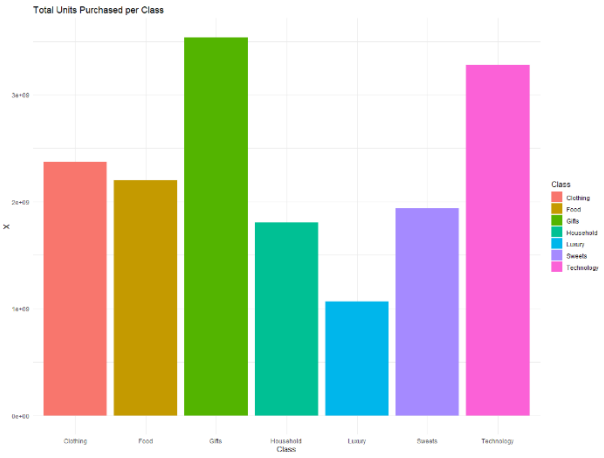

Figure 11: Total Units Purchased per Class

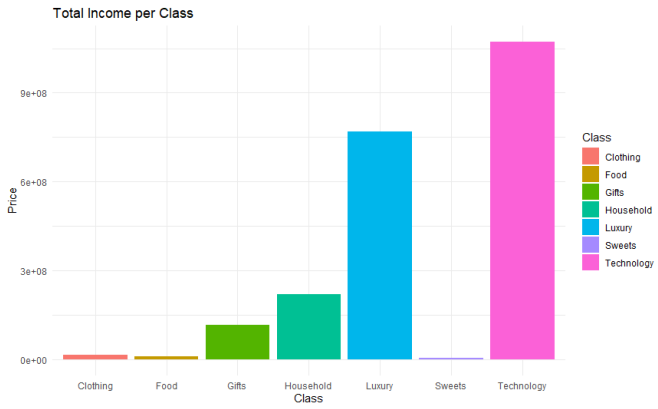

Figure 12: Total Income per Class

The figure below shows the distribution of age for the different classes. It is seen that Food is purchased mostly by older customers whereas Clothing is purchased more frequently by younger customers. There is a wider range of customers for the Sweets class.

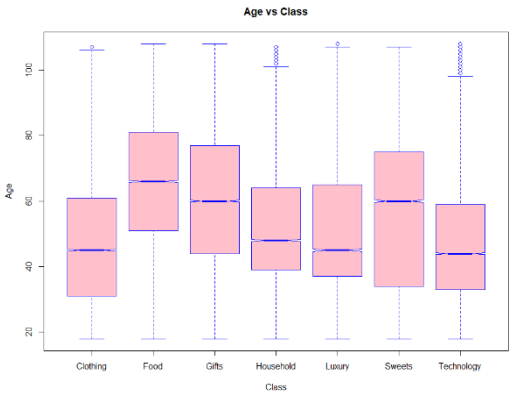

Figure 13: Age vs Class

Why.Bought Analysis

A graph of the number of items purchased per reason for purchase as well as the graph of income generated per reason for purchase is shown below. From the figures 16 and 17 below, it is seen that Recommended generated the highest number of sales as well as the largest income among the different reasons. This is useful to show the business that the methods or algorithms they are using are working. It also gives them an indication that there is room or improvement in the other reasons.

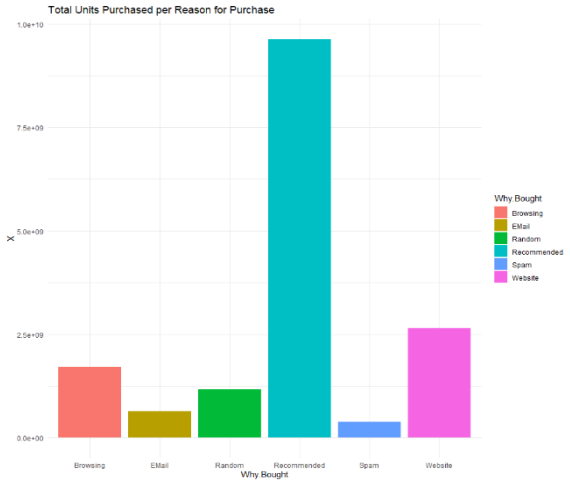

Figure 14: Total Units Purchased per Reason for Purchase

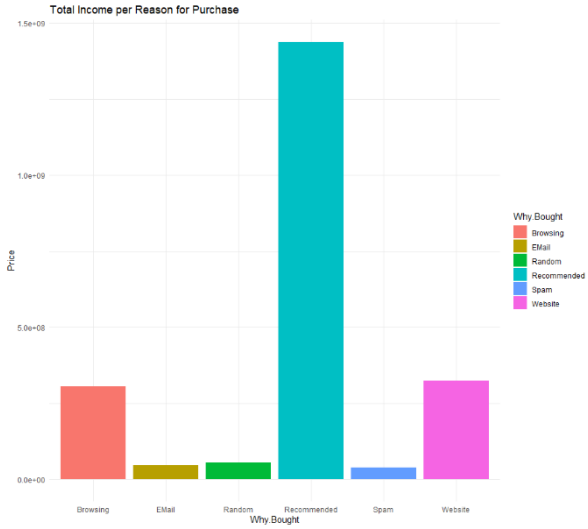

Figure 15: Total Income per Reason for Purchase

Figure 16 below represents the distribution of age throughout the different reasons for purchase. From this figure it can be concluded that there is a uniform distribution and that customers of all ages buy products for the same reasons.

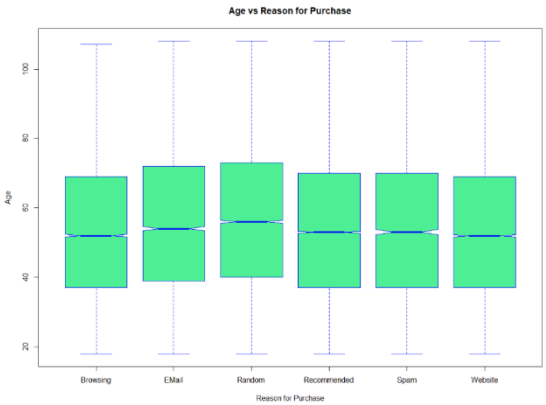

Figure 16: Age vs Reason for Purchase

2.4 Process Capability Indices

Process capability index are used to measure and determine the ability of process to produce within certain specification limits.

The process capabilities were calculated by assuming the following:

USL (Upper Specification Limit) = 24 Hours

LSL (Lower Specification Limit) = 0 Hours

A Lower Specification Limit of 0 is logical because it is not possible to have a delivery time lower than 0 Hours.

The mean and standard deviation were obtained by calculations (based on assumption that Delivery.time is given in hours):

σ = 3.501993

$\bar{X}$= 20.01095

By using the values mentioned above, the following process capability indices were calculated:

Indices	C_p	C_{pu}	C_{pl}	C_{pk}
Formula	$\frac{USL - LSL}{6\sigma}$	$\frac{USL - \bar{X}}{3\sigma}$	$\frac{\bar{X} - LSL}{3\sigma}$	$\min(C_{pu}, C_{pl})$
Value	1.142207	0.3796933	1.90472	0.3796933

Table 3: Capability Indices

A C_{pk} of 0.38 indicates the process is marginally capable of producing within the specification limits. In order to improve this, the business should decrease variability within the processes used. Further detail into this ability to remain within specification limits is discussed further in the sections following.

3. Statistical Process Control (SPC) for the X&s-charts

3.1 Initialise X&s-charts

The valid dataset is chronologically ordered according to Year, Month and Day from the oldest transaction to the newest and named ordData. The new dataset, ordData is then divided into each class of sale. The subsets of data are divided into 30 samples of 15 instances each. From this, 450 instances are used to generate the control charts.

3.1.1 X-Chart

The first 30 samples are used to determine centre lines, outer control limits, the 2-sigma-control limits and the 1-sigma-control limits. The following table provides a summary of the different delivery times for the seven different classes.

CLASS	LCL	L2SIGMA	L1SIGMA	CL	U1SIGMA	U2SIGMA	UCL
CLOTHING	8.535066	8.680044	8.825022	8.970000	9.114978	9.259956	9.404934
HOUSEHOLD	42.876117	44.104818	45.333520	46.562222	47.790924	49.019626	50.248328
FOOD	2.270542	2.343695	2.416847	2.490000	2.563153	2.636305	2.709458
TECHNOLOGY	17.774273	18.640997	19.507721	20.374444	21.241168	22.107892	22.974616
SWEETS	2.058514	2.198269	2.338023	2.477778	2.617532	2.757287	2.897042
GIFTS	7.233658	7.609475	7.985293	8.361111	8.736929	9.112747	9.488565
LUXURY	3.977146	3.977146	4.482752	4.735556	4.988359	5.241162	5.493965

Table 4: X-Chart

Graphs

The following graphs provide a graphical representation to the data shown in the X-charts above. The centre line is shown by the red line. The upper and lower limits are shown by the blue lines. The upper and lower one-sigma lines are shown by the orange lines and similarly the two-sigma lines are shown by the green lines.

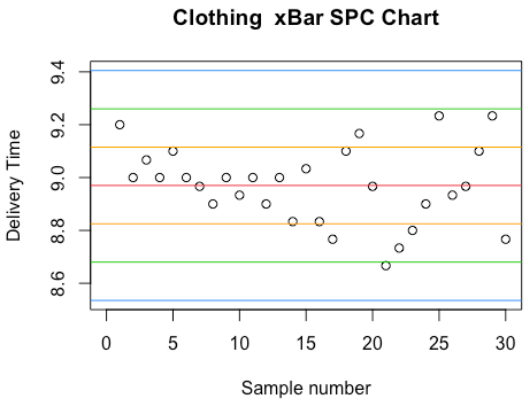

Figure 17: Clothing xBar SPC Chart

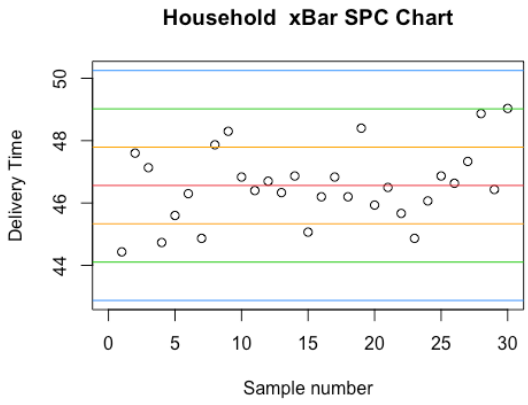

Figure 18: Household xBar SPC Chart

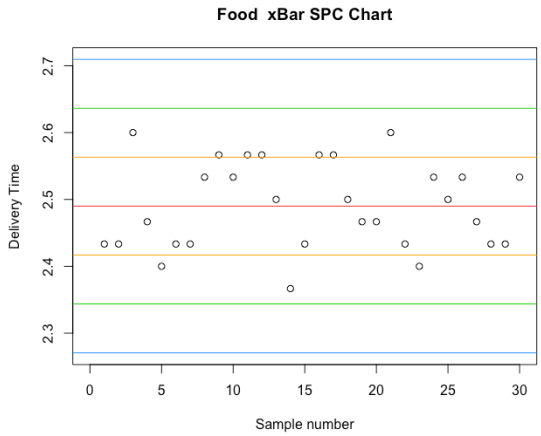

Figure 19: Food xBar SPC Chart

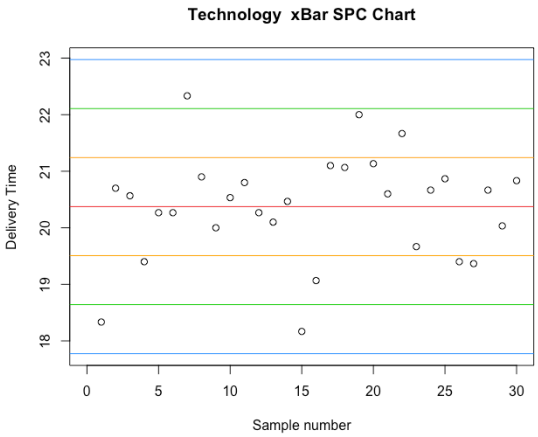

Figure 20: Technology

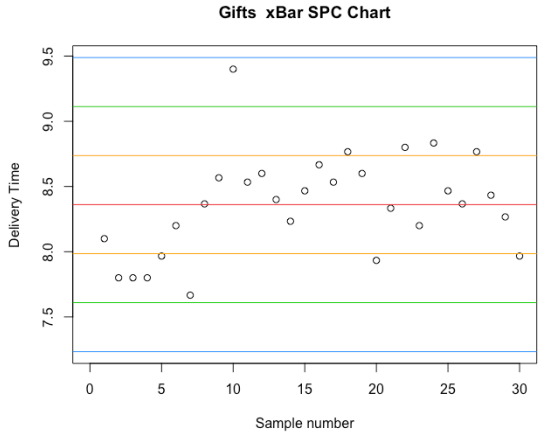

Figure 21: Gifts xBar SPC Chart

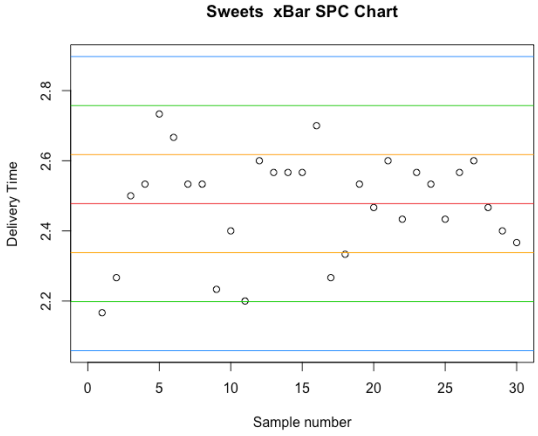

Figure 22: Sweets xBar SPC Chart

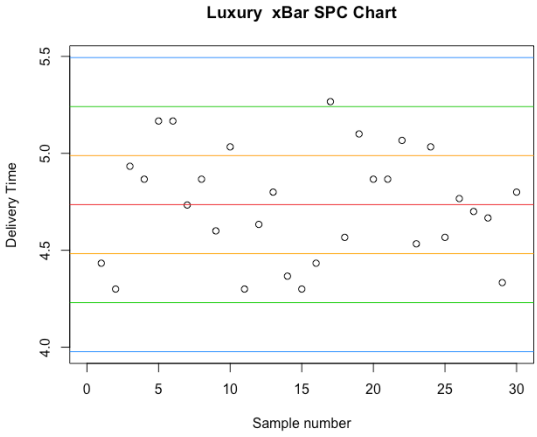

Figure 23: Luxury xBar SPC Chart

3.1.2 s-Chart

The first 30 samples are used to determine centre lines, outer control limits, the 2-sigma-control limits and the 1-sigma-control limits.

The following table provides a summary of the different delivery times for the seven different classes.

CLASS	LCL	L2SIGMA	L1SIGMA	CL	U1SIGMA	U2SIGMA	UCL
CLOTHING	0.2359335	0.3410379	0.4461422	0.5512465	0.6563509	0.7614552	0.8665596
HOUSEHOLD	1.9995605	2.8903304	3.7811003	4.6718703	5.5626402	6.4534101	7.3441801
FOOD	0.1190468	0.1720801	0.2251134	0.2781467	0.3311800	0.3842133	0.4372466
TECHNOLOGY	1.4104859	2.0388332	2.6671805	3.2955278	3.9238751	4.5522224	5.1805697
SWEETS	0.2274333	0.3287509	0.4300686	0.5313862	0.6327039	0.7340215	0.8353391
GIFTS	0.6115971	0.8840532	1.1565092	1.4289652	1.7014213	1.9738773	2.2463333
LUXURY	0.4114060	0.5946803	0.7779546	0.9612289	1.1445032	1.3277775	1.5110518

Table 5: s Chart

Graphs

The following graphs provide a graphical representation to the data shown in the S-charts above. The centre line is shown by the red line. The upper and lower limits are shown by the blue lines. The upper and lower one-sigma lines are shown by the orange lines and similarly the two-sigma lines are shown by the green lines.

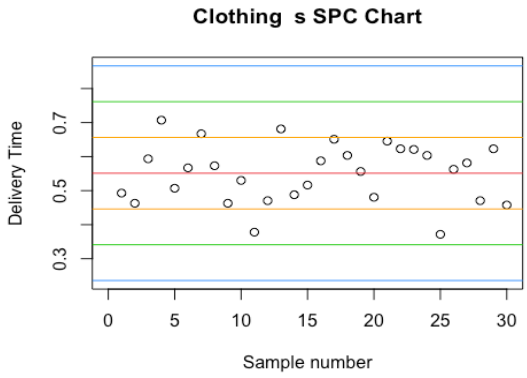

Figure 24: Clothing s SPC Chart

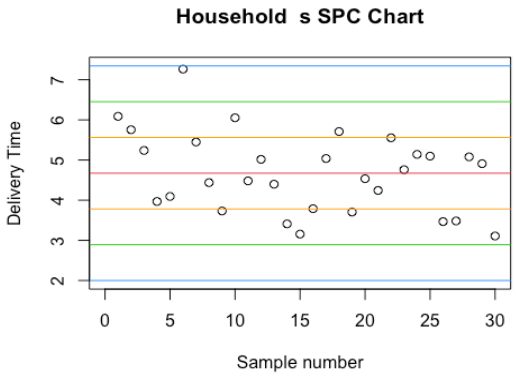

Figure 25: Household s SPC Chart

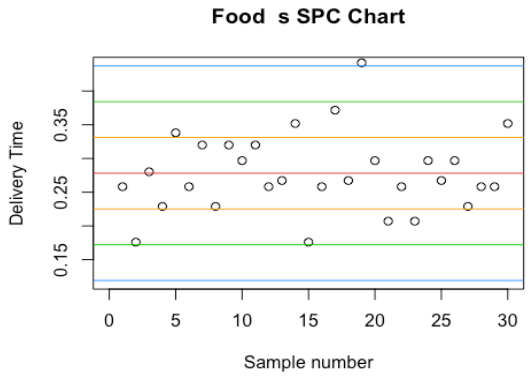

Figure 26: Food s SPC Chart

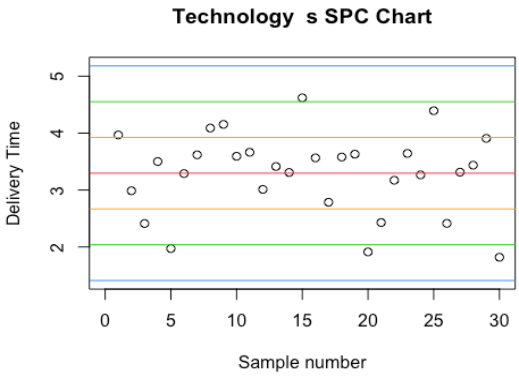

Figure 27: Technology s SPC Chart

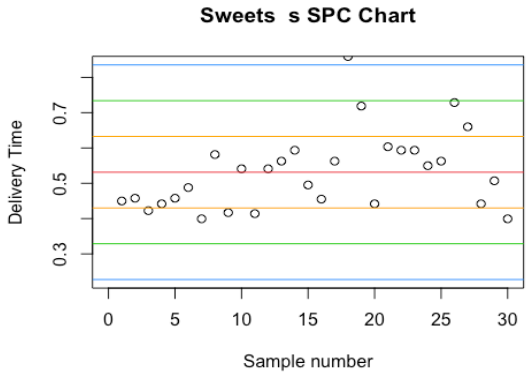

Figure 28: Sweets s SPC Chart

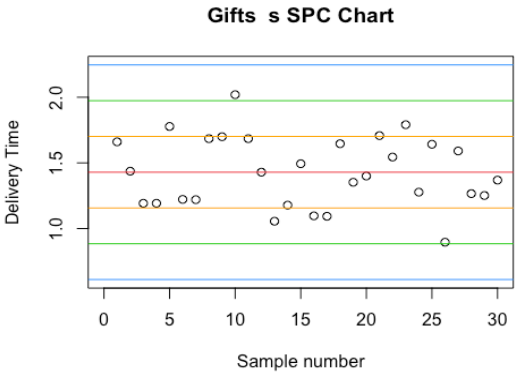

Figure 29: Gifts s SPC Chart

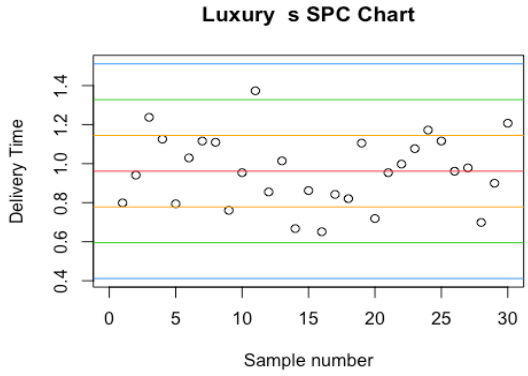

Figure 30: Luxury s SPC Chart

3.2 Remaining Samples

The remaining Delivery Time samples were drawn from the valid data and plotted according to their respective charts using Statistical Process Control. From these charts, it can be seen which instances fall outside the control limits, those that are within the limits as well as those areas that are more dense than others.

3.2.1 Remaining X and s-Charts

The graphs shown below provide a graphical representation of the remaining x-bar and s samples. The centre line is shown by the red line. The upper and lower control limits are shown by the blue lines. The green and yellow lines represent the 2-sigma and 1 sigma control limits respectively.

Clothing

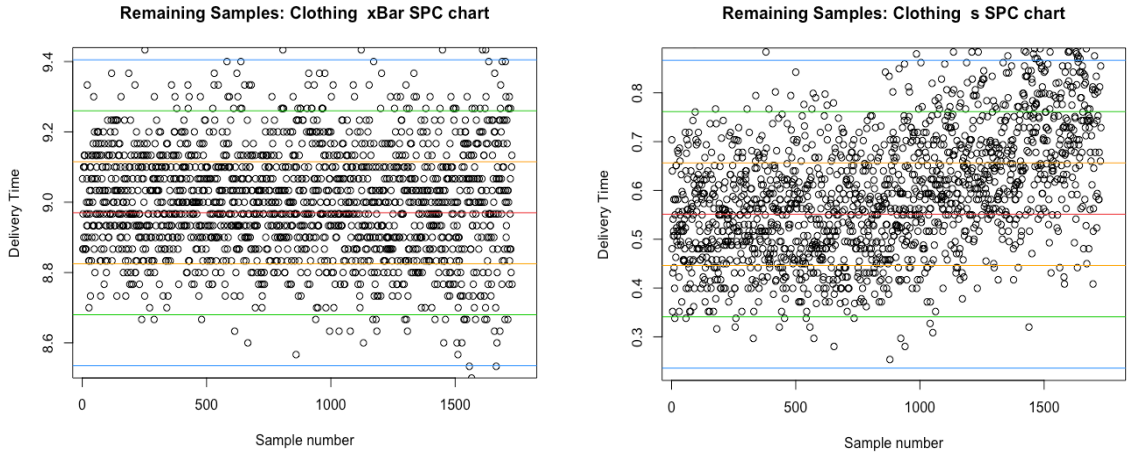

Figure 31: Remaining Clothing SPC Charts

From the figure above it is seen that majority of the samples fall within the control limits for the xBar SPC Chart. The samples in the xBar appear to be distributed along the control limits and different sigma lines. The s SPC chart is following an upward trend towards the right side of the chart (the later samples), which leads to more of the later samples being above the control limits.

Household

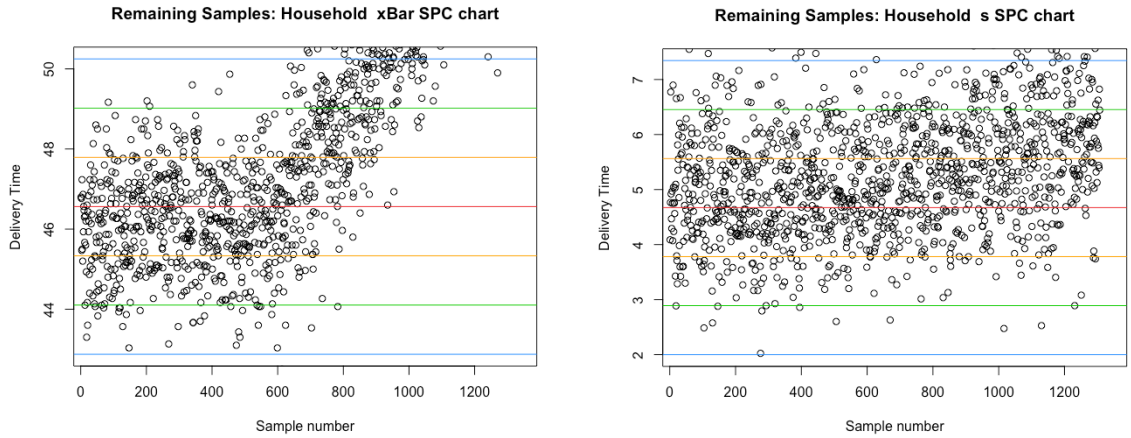

Figure 32: Remaining Household SPC Charts

Figure 32 shows that the xBar samples are following an exponential upward trend. As the sample indices increase, the samples move out control thus leading to many samples being above the upper control limit. The s SPC Chart follows a similar upward trend however the gradient is not as steep as seen in the xBar SPC Chart. These trends are important as it indicates that there may be problems within the delivery process.

Food

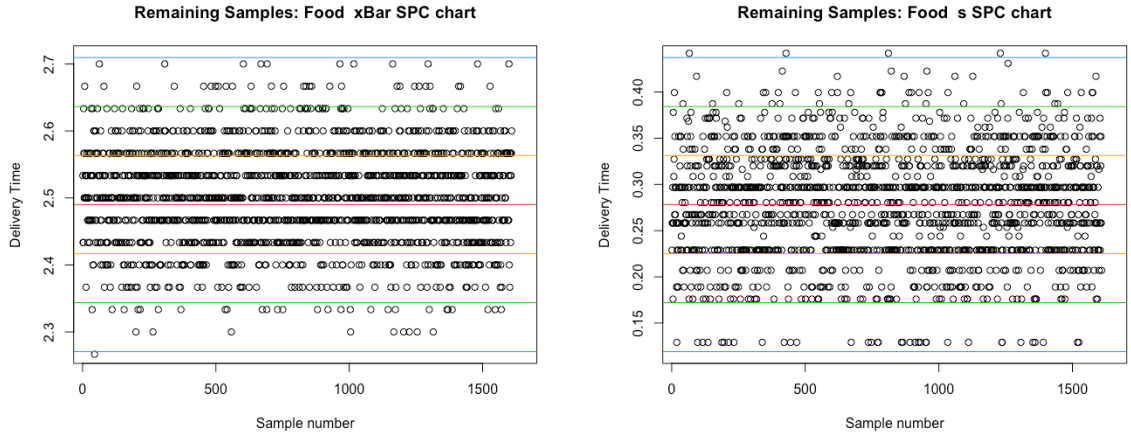

Figure 33: Remaining Food SPC Charts

The SPC charts for both xBar and s follow a similar linear trend. The samples are distributed relatively uniformly along the different control limits and sigma lines. The xBar is contained within the control limits, the few that are out are negligible in comparison to the amount of samples in control.

Technology

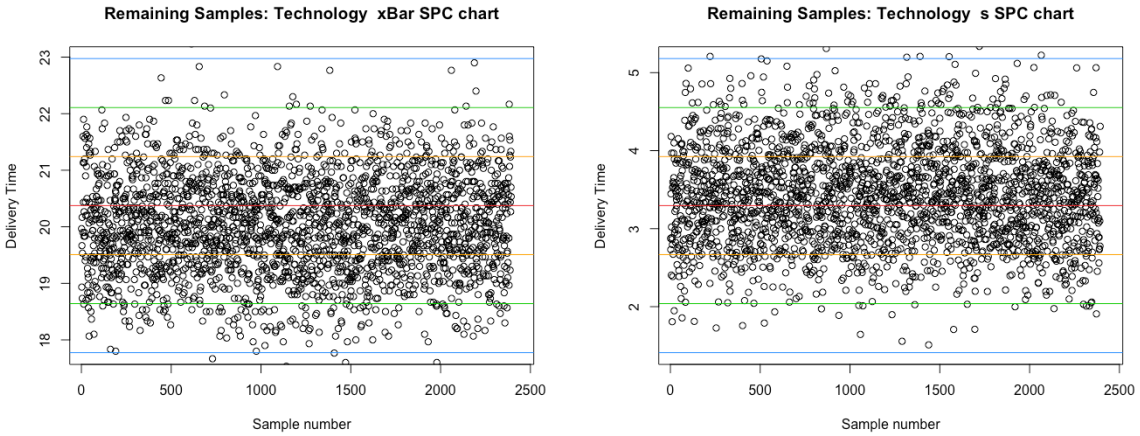

Figure 34: Remaining Technology SPC Charts

As seen in the figures above, the xBar and s SPC Chart are centred around the centre line. The xBar chat is more dense in the lower delivery time region, this is preferred as it means the process is working well whilst have very few samples out of the control limits.

Sweets

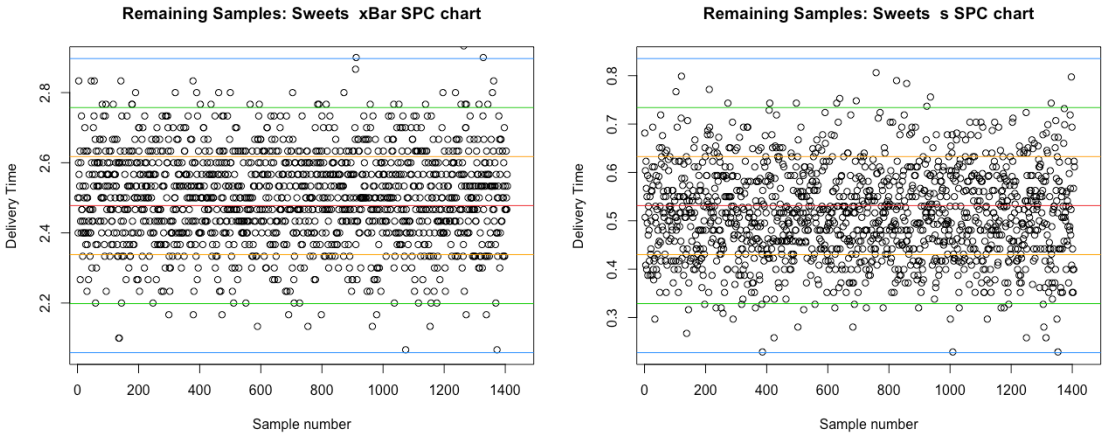

Figure 35: Remaining Sweets SPC Charts

The figures above illustrate that the most dense regions are around the centre line, with few instances falling outside of the control limits. These factors all lead to the conclusion that the process is in control.

Gifts

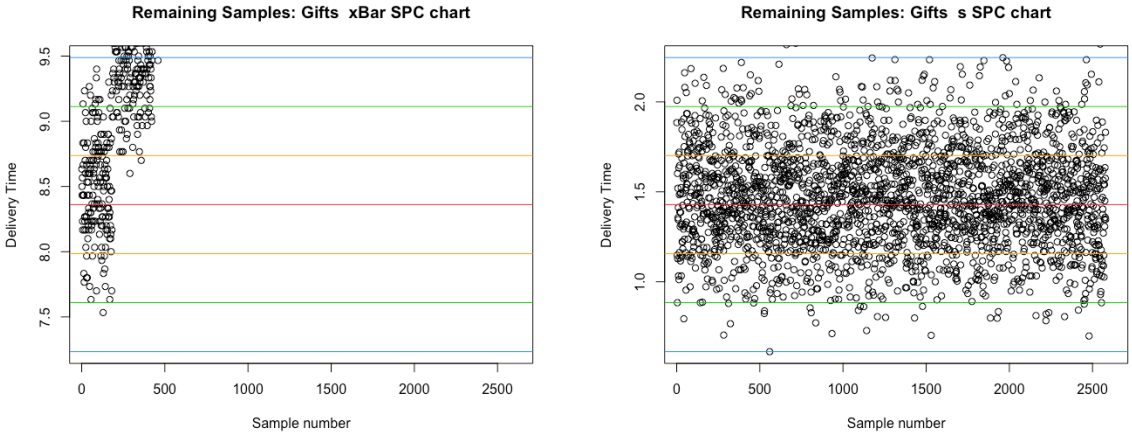

Figure 36: Remaining Gifts SPC Charts

The xBar SPC Chart for Gifts shown above illustrates that the process is completely out of control. It has a upward trend with a very steep gradient. This indicates that there is a problem within the process and measures should be put in place to resolve it. The s SPC Chart is condensed around the centre line, therefore the standard deviations of delivery times for Gifts is in control.

Luxury

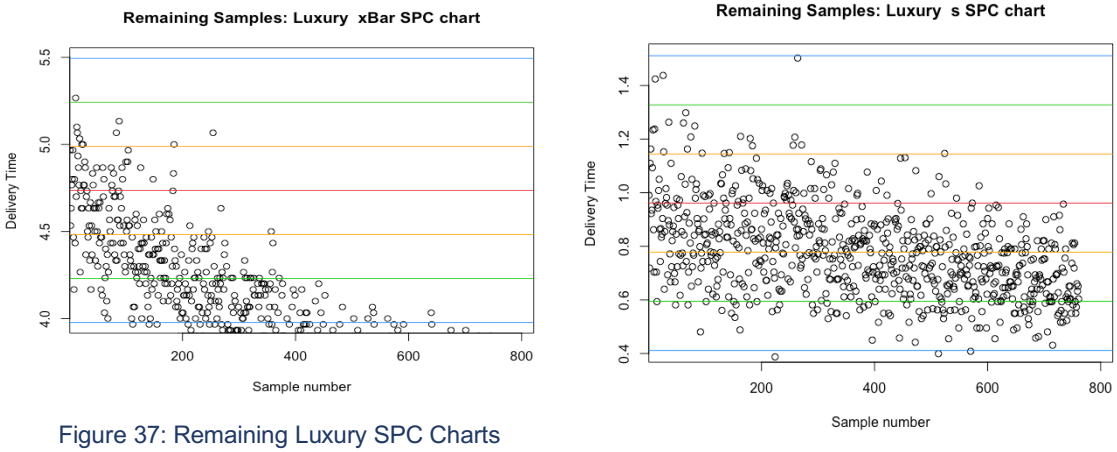

Figure 37: Remaining Luxury SPC Charts

Both the figures above follow a downward trend, with the xBar SPC Chart having a steeper gradient than the s SPC Chart. The downward trend leads to many samples being below the control limit for the xBar Chart. Although it is desirable to have lower delivery times, further analysis should be done to determine the reason for this problem. This process appears out of control.

4. Optimising the Delivery Processes

In order to perform an analysis of the delivery times for each respective class, the remaining sample data is used, as seen in section 3.2. These samples start from 31 and continue until the last sample. This analysis will provide thorough insight into the data allowing certain conclusions to be made about the delivery process.

4.1 Indication of Samples Out of Control Limits

4.1.1 Sample Means, xBar (A)

The table below represents the samples in each class that fall outside, either above or below, the upper and lower limits respectively. The indices of the first three, last three as well as the total number of samples identified are listed in the table below.

CLASS	1ST	2ND	3RD	3RD LAST	2ND LAST	LAST	TOTAL NO. OF INSTANCES
CLOTHING	455	702	1152	1677	1723	1724	17
HOUSEHOLD	252	387	629	1335	1336	1337	400
FOOD	75	633	1203	NA	1467	1515	5
TECHNOLOGY	37	398	483	1872	2009	2071	17
SWEETS	942	1104	1243	NA	1294	1403	5
GIFTS	213	216	218	2607	2608	2609	2290
LUXURY	142	171	184	789	790	791	434

Table 6: xBar Samples Out of Control

From the table above it is seen that the classes Household, Gifts, and Luxury contain the highest number of samples outside of the control limits. These numbers correspond to the graphical representation in figures 32,36 and 37. As a result of these findings, it can be concluded that these processes are not reliable and further, specific analysis should be performed to solve the problem. A way to do this could be through the analysis of the processes that contain few samples out of control. Through this a comparison can be made between the different processes.

4.1.2 Most Consecutive Sample Standard Deviations (B)

The table below illustrates the most consecutive sample standard deviations (s-bar samples) between -0.3 and 0.4 sigma-control limits as well as the index of the last sample in the given range.

CLASS	NO. OF CONSECUTIVE SAMPLES	LAST NO. IN SEQUENCE
CLOTHING	8	578
HOUSEHOLD	10	214
FOOD	8	746
TECHNOLOGY	9	199
SWEETS	6	284
GIFTS	9	313
LUXURY	4	26

Table 7: Consecutive s Samples Out of Sigma Limits

The class Gifts, Household and Technology have the highest number of consecutive samples between the sigma control limits. This indicates that those instances were delivered in very accurate and desired times. This could indicate that a few deliveries were done inadequately in order to appear correct and fast.

4.2 Type I Error

In this section an estimation of the likelihood of making a Type I error for 4.1.1 and 4.1.2 is made. A Type I error, also known as a manufacturer's error is a theoretical value that holds true for any process. A Type I error is essentially a false positive, as it indicates the process is out of control when in reality there is nothing with the process (Bhandari, 2022).

4.2.1 Likelihood of making a Type I error for A

The upper and lower control limits for A are +3 and -3 sigma values away from the centre line, respectively. Determining the likelihood of making a Type I error is done by calculating the probability between the Z values, -3 and +3.

$$P(Type\ I\ Error) = pnorm(-3) * 2$$

$$= 0.002699796$$

$$= 0.2699796\%$$

This probability is not very high, therefore it is not very likely that a Type I error will occur. The business should therefore accept that there is a risk without being too concerned about it.

4.2.2 Likelihood of making a Type I error for B

The upper and lower control limits for B are +0.4 and -0.3 sigma values away from the centre line, respectively. In order to determine the likelihood of making a Type I error, the samples for the class Household is chosen as the specific example. Household is chosen because it has the highest number of consecutive samples in these limits.

$$\begin{aligned} P(Type\ I\ Error) &= \big(pnorm(0.4) - pnorm(-0.3)\big)^{10} \\ &= 0.00000232768 \\ &= 0.000232768\%. \end{aligned}$$

As seen in part A, this probability is very low, therefore it is not very likely that a Type I error will occur. The business should therefore accept that there is a risk without being too concerned about it.

4.3 Optimising Delivery Times

The data in this section uses the individual delivery times for the Technology class. The aim of this section is to determine the optimal reduction in all delivery times for Technology, whilst taking into account the relevant costs and penalties associated with this reduction. This is done to obtain the number of hours the delivery process should be centred around to make the best profit. There is a penalty cost charge when the delivery of products is slower than 26 hours. Below are the penalty costs used to calculate the optimal reduction in delivery times.

Cost per hour that a product is late = R329 /item-late-hour

Cost to reduce delivery time by one hour = R2.5 /item/hour

After careful calculations the following graph was plotted.

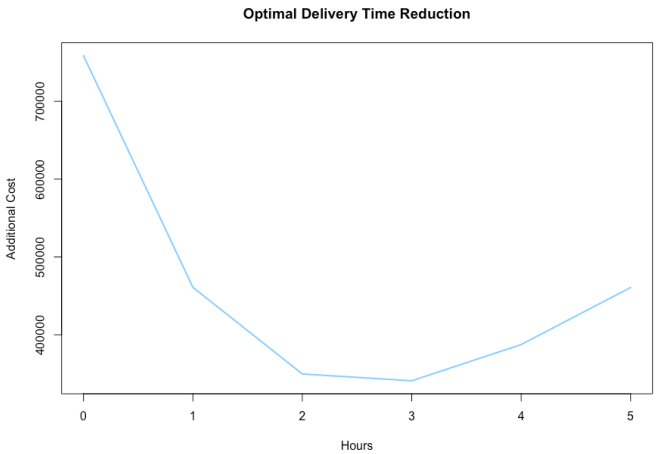

Figure 38: Optimal Delivery Time Reduction

From this graph, it is seen that the maximum reduction in delivery time is 3 hours. After this point the additional no longer decreases, indicating that reducing the delivery time will incur a penalty. With all the delivery times reduced by 3 hours, the new average delivery time is 17 hours. The trend seen in this graph follows that of the Taguchi Loss curve.

4.4 Estimate the likelihood of making a Type II Error

This section will focus on the Technology class, specifically the samples used in 4.1.1 (A). An estimation of the likelihood of making a Type II error is performed. A Type II error, also referred to as Consumer's error is essentially a false negative. This means the process indicates there is no error, when in reality the process is out of control (McLeod, 2019). This can lead missing important information regarding the process. Given the delivery process average moves to 23 hours, the likelihood of making a Type II error between the upper and lower limits is calculated below.

$$\begin{aligned} P(Type\ II\ Error) &= pnorm(UCL, \mu, \sigma) - pnorm(LCL, \mu, \sigma) \\ &= 0.4955224 - 0.01042696 \\ &= 0.4850954 \\ &= 48.5\% \end{aligned}$$

This is a high probability. The risk of obtaining a Type II error can be reduced by ensuring the sample size is large enough to detect whether there is actually a practical difference.

5. DOE and MANOVA

The MANOVA is performed on the categorical feature Class. Class is the category the items bought belong to. The features “Delivery.time”, “Age” and “Price” are used to determine the MANOVA.

Null Hypothesis: The factors of Delivery time, Age and Price do not significantly affect the number of items purchased from certain classes.

Alternative Hypothesis: At least one factor has a significant influence on the number of items purchased from certain classes.

Alpha is chosen to be 0.05. This indicates that a p-value of smaller than 0.05 will mean the tests is significant. From the test performed in R, a p-value of 2.2^{-16} is obtained. This value is considerably smaller than value of alpha indicating at least one factor has an effect on the number of items purchased from certain classes.

After setting up MANOVA’s, boxplots are generated, and the results are discussed below:

Figure 39: Box Plot of Age per Class

The medians of the seven classes differ. Food has the highest median whereas clothing has the lowest. This is consistent with the expected items those respective age groups would purchase. From this, it is seen that age does affect the number of items purchased from certain classes.

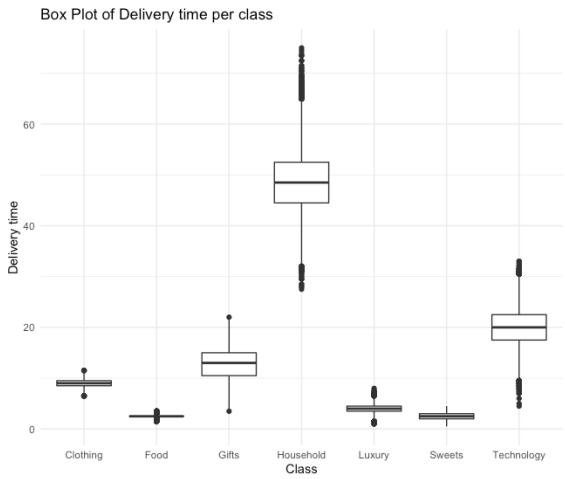

Figure 40: Box Plot of Delivery Time per Class

As illustrated in Figure 40, delivery time has a large effect on the number of items purchased from certain classes. The delivery time for the Household class has the slowest delivery time median which is consistent as household items usually assembly or are stored in warehouses that could increase delivery time. The Food class has the fastest delivery time, which could indicate this process is efficient and the other classes could replicate those processes to improve its delivery times.

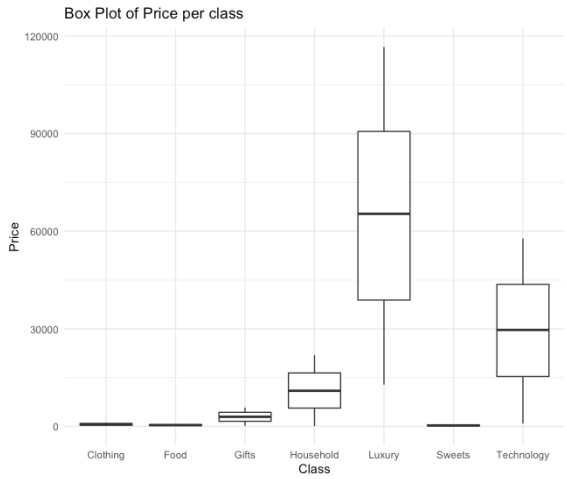

Figure 41: Box Plot of Price per Class

As seen in Figure 41, Price has a significant effect on the number of items purchased from certain classes. As expected the Luxury class has the highest price, and Sweets has the lowest. From figure it is seen that Technology and Luxury generate the highest income for the business.

6. Reliability of the service and products

6.1 Problem 6 and 7

Problem 6 description:

A blueprint specification for the thickness of a refrigerator part at Lafrigeradora, Inc. is 0.06 ± 0.04 centimetres (cm). It costs $45 to scrap a part that is outside the specifications. Determine the Taguchi loss function for this situation.

Thickness = 0.06 ± 0.04 cm

Cost= $45

Taguchi loss function: $L(x) = k(x - T)^2$

Where, $L(x) = 45$

T = 0.06

$(x - T) = 0.04$

Solve for k by substituting the known values:

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

$$k = 28125$$

Thus, $L(x) = 28125(x - 0.06)^2$

Problem 7 description:

A team was formed to study the refrigerator part at Lafrigeradora Inc. described in Problem 6. While continuing to work to find the root cause of scrap, they found a way to reduce the scrap cost to $35 per part.

a) Determine the Taguchi loss function for this situation.

$$k = \frac{35}{(0.04)^2} = 21875$$
$$L(x) = 21875(x - 0.06)^2$$

b) If the process deviation from target can be reduced to 0.027 cm, what is the Taguchi loss?

$$L(x) = 21875(0.027)^2 = 15.95$$

6.2 Problem 27

Problem description:

Magnaplex, Inc. has a complex manufacturing process, with three operations that are performed in series. Because of the nature of the process, machines frequently fall out of adjustment and must be repaired. To keep the system going, two identical machines are used at each stage; thus, if one fails, the other can be used while the first is repaired.

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

$$\begin{aligned} Reliability &= R_A \times R_B \times R_C \\ &= 0.85 \times 0.92 \times 0.90 \\ &= 0.7038 \\ &= 70.38\%. \end{aligned}$$

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

$$\begin{aligned} Reliability &= [1-(1-RA)\times(1-RA)]\times[1-(1-RB)\times(1-RB)]\times[1-(1-RC)\times(1-RC)] \\ &= [1-(1-0.85)\times(1-0.85)]\times[1-(1-0.92)\times(1-0.92)]\times[1-(1-0.90)\times(1-0.90)] \\ &= (0.9775)(0.9936)(0.99) \\ &= 96.15\% \end{aligned}$$

From this it can be concluded that the system improved by 25.77% when having 2 machines instead of one.

6.3 Binomial Probability

Problem description:

For the delivery process, there are 20 delivery vehicles available, of which 19 is required to be operating at any time to give reliable service. During the past 1560 days, the number of days that there was only 20 vehicles available was 190 days, only 19 vehicles available was 22 days, only 18 vehicles available was 3 days and 17 vehicles available only once. There are also 21 drivers, who each work an 8 hour shift per day. During the past 1560 days, the number of days that there were only 20 drivers available was 95 days, only 19 drivers available was 6 days and only 18 drivers available, once only. Estimate on how many days per year we should

expect reliable delivery times, given the information above. If we increased our number of vehicles by one to 21, how many days per year we should expect reliable delivery times?

After thorough calculations and iterations using the dbinom() function in R, the following conclusions are drawn:

$$Probability\ of\ reliable\ number\ of\ vehicles\ = 0.9278192$$

$$Probability\ of\ reliable\ number\ of\ drivers\ = 0.9471258$$

$$Total\ Probability\ of\ reliability\ = P(Vehicles) * P(Drivers)$$

$$= 0.8787615.$$

$$= 87.8\%$$

Thus the expected number of days reliable is:

$$Expected\ days = 0.8787615021 * 365$$

$$= 319.8692$$

$$= 319\ days$$

Therefore 319 out of 365 days will have a reliable delivery.

After repeating the calculations as before, with the change of adding one extra vehicle (from 21 to 22) , the following conclusions are drawn:

$$Probability\ of\ reliable\ number\ of\ vehicles\ = 0.9601772$$

$$Probability\ of\ reliable\ number\ of\ drivers\ = 0.9471258$$

$$Total\ Probability\ of\ reliability\ = P(Vehicles) * P(Drivers)$$

$$= 0.9094086$$

$$= 90.94\%$$

If an extra vehicle is added the expected number of days is:

$$Expected\ days = 0.9094086 * 365$$

$$= 331.0247$$

$$= 331\ days$$

This is a 12 day increase from the first set of calculations. Thus adding a vehicle will prove to be beneficial to the business as they will have more reliable days per year.

Conclusion

Through this report, a data set containing information regarding the sales of an online business is analysed. This is first achieved by cleaning the data. Through descriptive statistics methods, many conclusions are drawn about the data. It is found that the class of items, Technology, generated the highest income and Gifts sold the most number of units. The primary reason for purchase of items is Recommended, indicating that this is a useful method for the business and it should work to improve the sales numbers for the remaining reasons for purchase. The Delivery Time x-bar and s Statistical Process Control charts are configured. Through assessment of these graphs, it is found that the delivery process of the class Food is in control whereas the processes for Gifts and Luxury follow a trend that rapidly grows out of control. The analysis of the sales data allows the business to identify the processes that are working well and those that need improvement and adjustments.

References

Bhandari, P., 2022. *Type I & Type II Errors | Differences, Examples, Visualizations.* [Online]
Available at: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/
[Accessed 9 October 2022].

McLeod, S., 2019. *What are Type I and Type II Errors?.* [Online]
Available at: https://www.simplypsychology.org/type_I_and_type_II_errors.html
[Accessed 9 October 2022].