QUALITY ASSURANCE 344

ECSA PROJECT: DATA ANALYSIS AND MANIPULATION

October 21, 2022

MPHO MABOEE
21807779

Contents

List of Figures

List of Tables

Introduction

Many organizations approach data analysis in numerous ways. However, data analysis can be defined as the process of cleaning, changing, and processing raw data, and extracting actionable, relevant information that help the business make informed decisions. This report will analyse sales data for an online business. The provided data consists of invalid data that may need to be cleaned or removed. Moreover, the data consists of NA values that need to be removed. The analysis process will be conducted in six parts; the first process is data wrangling where the data will be cleaned and categorised as either valid or invalid for analysis. The second process is descriptive statistics, this part will further study the valid data by constructing distribution graphs. The third process is statistical process control (R programming), the x-bar and s-charts will be computed for a sample of each class. The forth process is optimizing the data delivery processes which will study the sample means, sample standard deviations and deduce cost estimations. The fifth process is DOE and MANOVA where a hypothesis test will be conduct on the data to support the results in the previous parts. The last process is reliability of the service and products, in this part binomial probabilities will be computed for the data.

Part 1: Data Wrangling

Data wrangling is the process of transforming raw data into formats that are easier to use. The process consists of the following steps: data discovery, data structuring, data cleaning, data enriching, data validating and data publishing. This part of the report will focus on data cleaning and invalid data anomaly.

Data cleaning

In this process, an overall observation of the data is made. When looking at figure 1, it can be evaluated that the data consists of missing values (NAs); most of these values are found under the price attribute. Moreover, the price attribute constitutes of negative entries, these need to be removed in order to validate the data.

A subset called “validData” was created, within this subset, all missing values and negative values are removed to make the data reliable. The missing values and negatives values are stored in a separate subset named “invalidData”.

In the original dataset, the “primary key” is the “X” column. The new subsets have “X” as the secondary as it shows which entry it is in the original dataset. Moreover, the primary key shows the entry number in the new subset for instance, in the invalid subset, the observation

with primary key equal to 1 has a secondary key of X equal to 12345. The above process was done in a way to ensure that no data was removed from the original dataset.

Table 1: Summary of Sales Data

X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
Min. : 1	Min. :11126	Min. : 18.00	Length:180000	Min. : -588.8	Min. :2021	Min. : 1.000	Min. : 1.00	Min. : 0.5	Length:180000
1st Qu.: 45001	1st Qu.:32700	1st Qu.: 38.00	Class :character	1st Qu.: 482.3	1st Qu.:2022	1st Qu.: 4.000	1st Qu.: 8.00	1st Qu.: 3.0	Class :character
Median : 90001	Median :55081	Median : 53.00	Mode :character	Median : 2259.6	Median :2025	Median : 7.000	Median :16.00	Median :10.0	Mode :character
Mean : 90001	Mean :55235	Mean : 54.57	NA	Mean : 12293.7	Mean :2025	Mean : 6.521	Mean :15.54	Mean :14.5	NA
3rd Qu.:135000	3rd Qu.:77637	3rd Qu.: 70.00	NA	3rd Qu.: 15270.7	3rd Qu.:2027	3rd Qu.:10.000	3rd Qu.:23.00	3rd Qu.:18.5	NA
Max. :180000	Max. :99992	Max. :108.00	NA	Max. :116619.0	Max. :2029	Max. :12.000	Max. :30.00	Max. :75.0	NA
NA	NA	NA	NA	NA's :17	NA	NA	NA	NA	NA

Anomalies invalid data

An anomaly is the occurrence of an unexpected change within data patterns or, an event that does not conform to the expected data pattern. The abnormalities in the dataset are named “invalidData”. The negative and missing values are found under the price column. These inputs could have been incorrectly captured.

Part 2: Descriptive Statistics

For part 2 of the analysis process, emphasis will be put on descriptive statistics. The investigation will assist in describing or summarizing data in a meaningful way by observing patterns that may arise from the data.

Descriptive Statistics

In order to make remarks regarding data, one needs to study the patterns/trends that the dataset produces. Descriptive statistics can be visualised by measuring central tendencies and the spread of data. The figures below facilitate data visualisation and can be interpreted based on the distributions depicted.

Table 2: Summary of Valid Sales Data

primary.key	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
Min. : 1	Min. : 1	Min. :11126	Min. : 18.00	Length:179978	Min. : 35.65	Min. :2021	Min. : 1.000	Min. : 1.00	Min. : 0.5	Length:179978
1st Qu.: 44995	1st Qu.: 45004	1st Qu.:32700	1st Qu.: 38.00	Class :character	1st Qu.: 482.31	1st Qu.:2022	1st Qu.: 4.000	1st Qu.: 8.00	1st Qu.: 3.0	Class :character
Median : 89990	Median : 90005	Median :55081	Median : 53.00	Mode :character	Median : 2259.63	Median :2025	Median : 7.000	Median :16.00	Median :10.0	Mode :character
Mean : 89990	Mean : 90003	Mean :55235	Mean : 54.57	NA	Mean : 12294.10	Mean :2025	Mean : 6.521	Mean :15.54	Mean :14.5	NA
3rd Qu.:134984	3rd Qu.:135000	3rd Qu.:77637	3rd Qu.: 70.00	NA	3rd Qu.: 15270.97	3rd Qu.:2027	3rd Qu.:10.000	3rd Qu.:23.00	3rd Qu.:18.5	NA
Max. :179978	Max. :180000	Max. :99992	Max. :108.00	NA	Max. :116618.97	Max. :2029	Max. :12.000	Max. :30.00	Max. :75.0	NA

Figure 1: Distribution of Product Class by Age

Figure 1 shows the product distribution relative to the class. The overall dataset is skewed positively, the right skewness results in a mean greater than the median. This also means that the data has a higher number of data points having low values (i.e. there are more customers that fall within a lower age group).

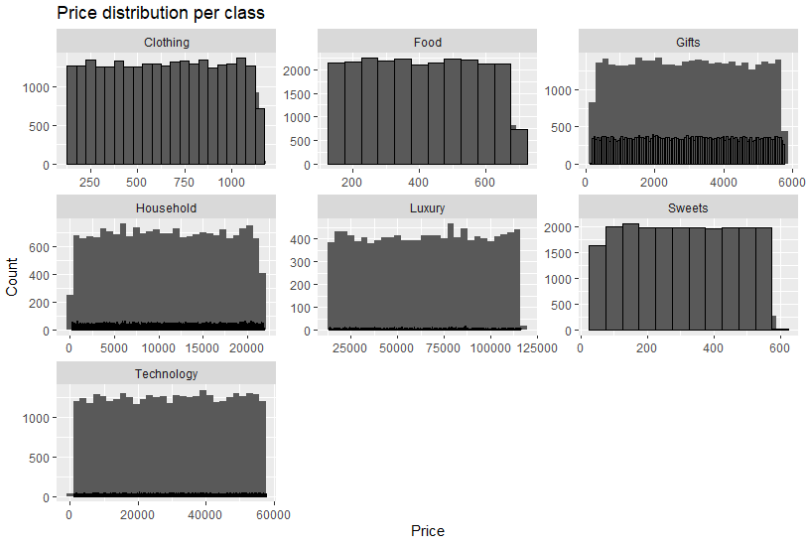


Figure 2: Distribution of Density by Price of each Class

In figure 2, we are presented with the density distribution of the price per class. Looking at the overall results, the data seems to be uniform distributed with a few variations. It can further be deduced that the dataset is continuous meaning there is an infinite number of equally likely measurable values. This also highlights an idealized random number generator in the prices and that the population is large.

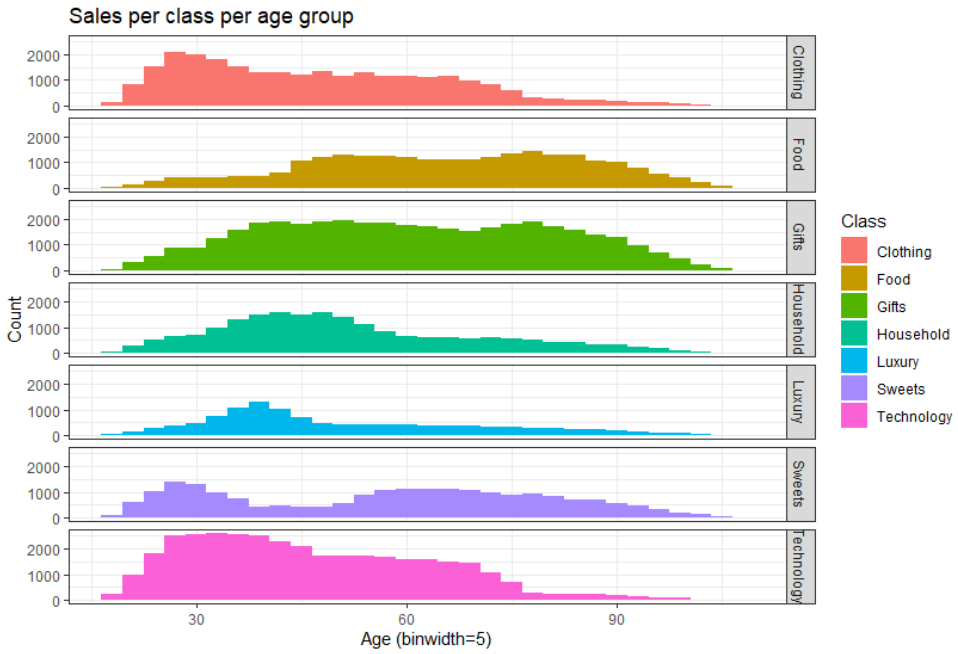


Figure 3: Distribution of Sales per Class per Age Group

Figure 3 depicts the distribution of sales within the various classes relative to the age groups. When analysing the classes specifically clothing, technology, luxury and household; we see that the data is positively skewed. When training a model on this data, it will perform better at predicting the sales produced by customers within a lower age group within the mentioned classes. The food and gifts classes seem normally distributed, this means the sales produced from the lower and higher age group are equal. The sweets class seems to consist of 2 normal distributions meaning there are two average points. This analysis can also be regarded for figure 4 as the distributions are almost similar.

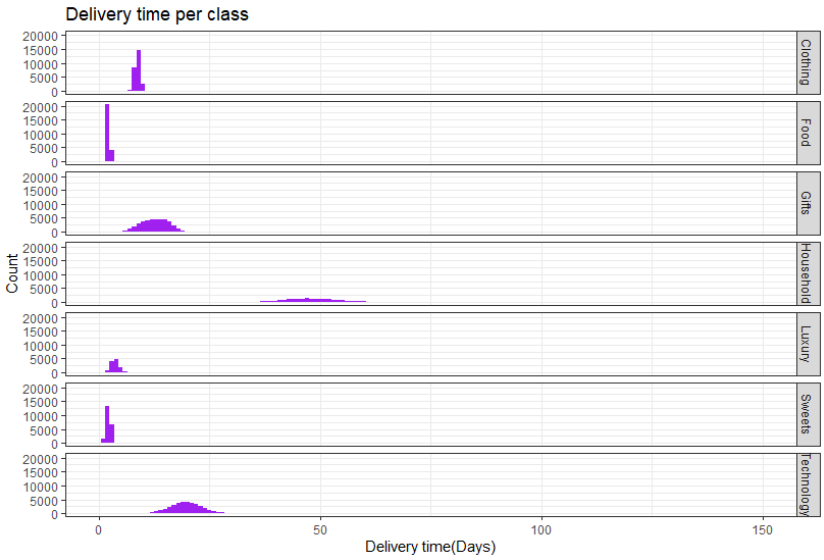


Figure 4: Distribution of Delivery Time per Class

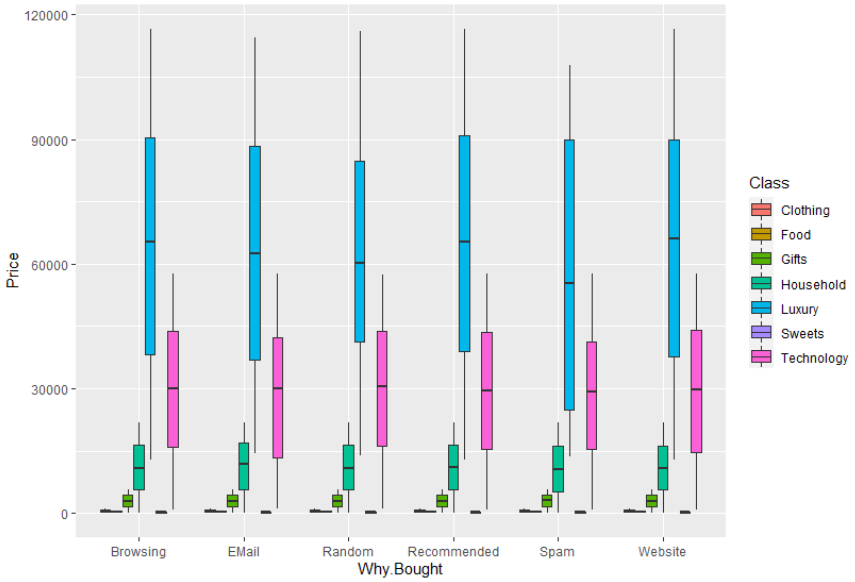


Figure 5: Box Plot Showing Price, Class & Purchase reason

In figure 4, we see the main reasons items where purchased in each class, as well as the prices. The dataset has no outliers and it can be concluded that the median per class in each reason is almost similar.

Process Capacity Indices

This section of the report looks at the process capability indices of process delivery times of data that falls under the technology class. The Cp, Cpu, Cpl, and Cpk were calculated. The aim of these calculations is to analyse and determine the relationship of the specifications of the delivery process times and variation that the data of this feature has. In order to conduct the said indices, an upper specification limit (USL) of 24 days and a lower specification limit (LSL) of 0 days. An LSL of 0 days is logical on the basis that items cannot be delivered in less than 0 days. Items can however be delivered on the same day as they are ordered.

Table 3: Capacity indices for delivery process of Technology

	Cp	Cpu	Cpl	Cpk
1	1.14224574364484	0.379695892565369	1.9047955947243	0.379695892565369

Part 3: Statistical Process Control

Part 3 will discuss and analyse the X-bar and S control charts of each class. In this part, the X-bar and S-chart were constructed using the first 30 samples of 15 sales in each class. The X-bar chart presents the sample means for the subgroup while the s chart shows the standard deviations.

Analysis of X-bar & S-chart for Clothing class

This sub-section will analyse the x-bar and s control charts of the data under the clothing class.

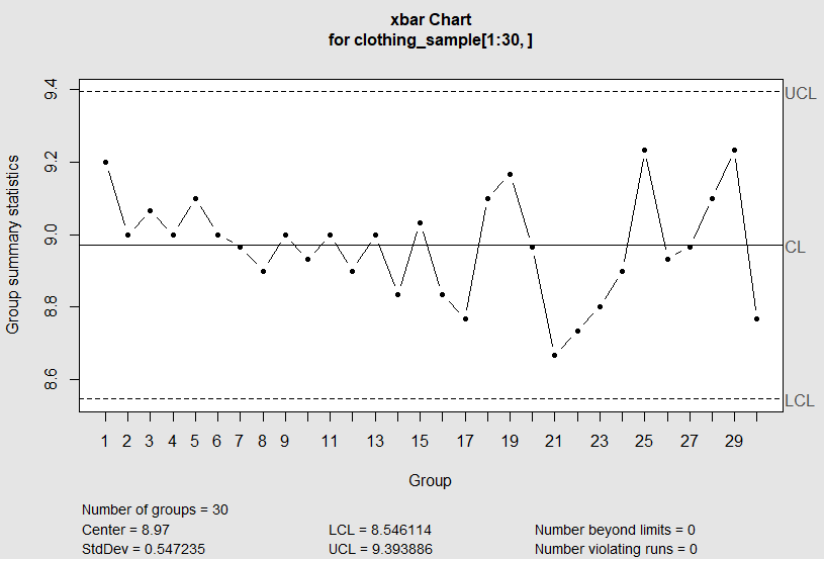


Figure 6: Xbar chart of Clothing sample

Figure 6 shows the sample mean, as well as the central line and control limits for the process, based on the 30 subgroups for the Clothing sample. There are no outliers, so the process appears to be in control. Moreover, there seems to be an equal number of samples on each side of the centre line with a few samples being extremely close to the centre. In addition, there is no specific trend/pattern in the layout of the samples.

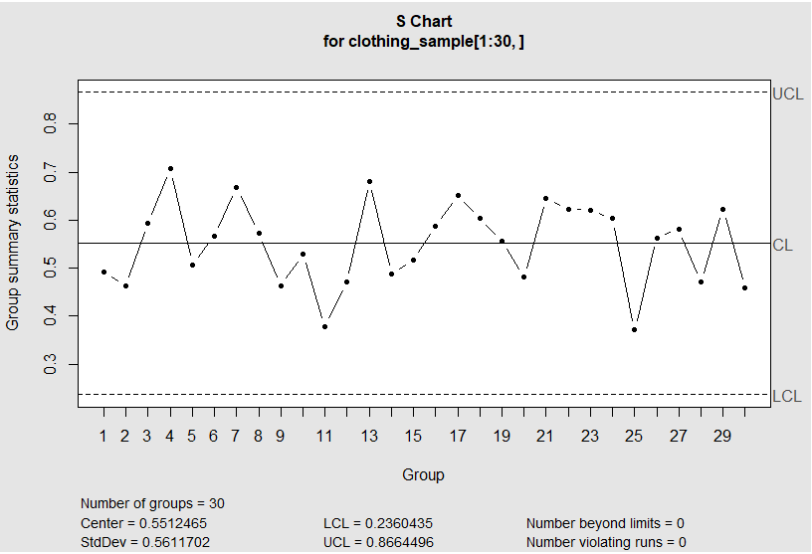


Figure 7: Schart of Clothing sample

Figure 7 shows the sample standard deviations as well as the corresponding central lines and control limits for each subgroup in the Clothing sample. There is no apparent trend/pattern in the layout of the samples. The number of samples is evenly distributed between both sides of the centre line and most importantly the s chart seems to indicate the variation is also in control.

Analysis of X-bar & S-chart for Food class

This sub-section will analyse the x-bar and s control charts of the data under the food class.

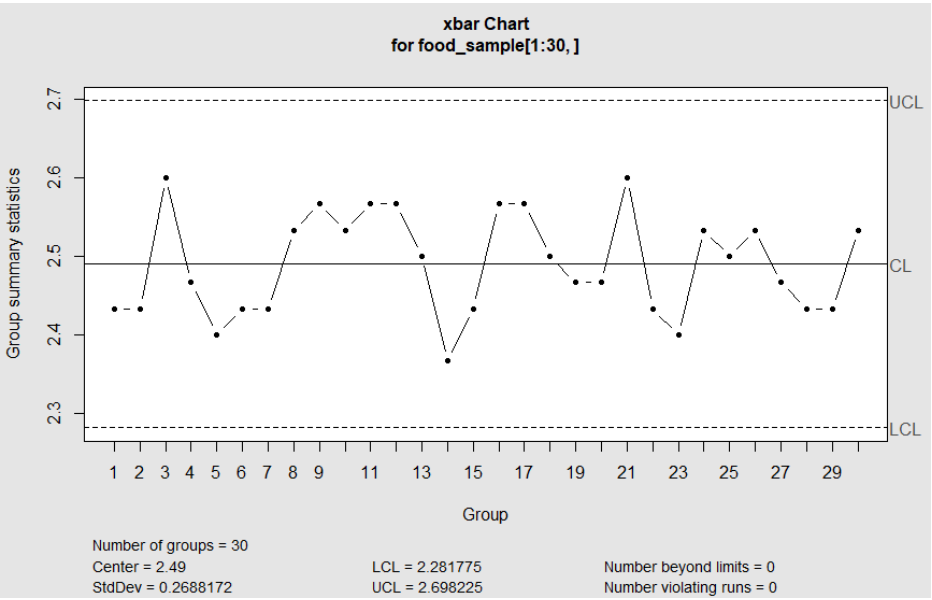


Figure 8: Xbar chart of Food sample

Figure 8 depicts the sample mean, as well as the central line and control limits for the process, based on the 30 subgroups of the food sample. When looking at the sample, there seems to be

no trend/pattern in the layout of the data. The process is within the control limits therefore, the process is statistically in control.

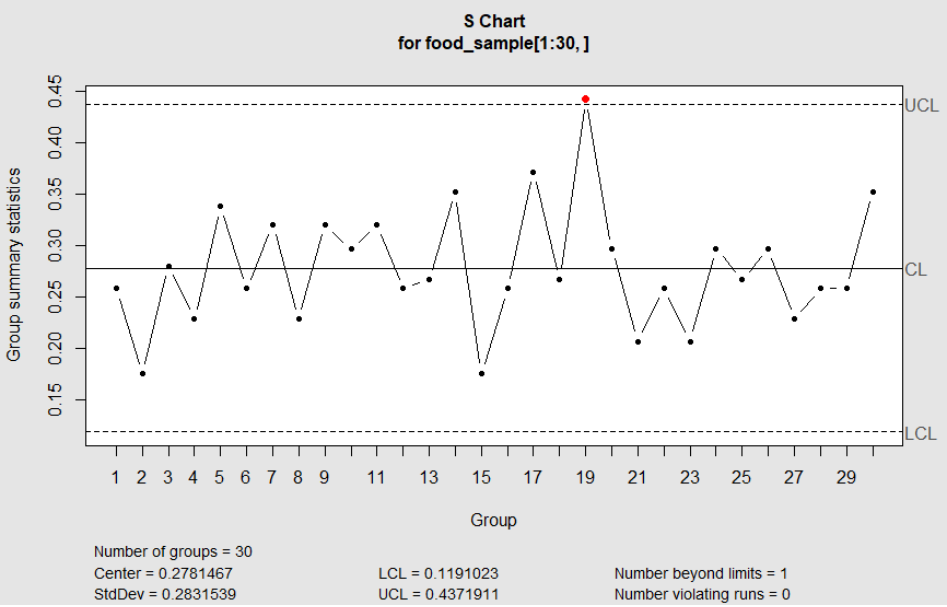


Figure 9: Schart of Food sample

Figure 9 shows the sample standard deviations as well as the corresponding central lines and control limits for each subgroup in the food sample. The sample has an outlier at group 19. This can be caused by a drastic difference in data within the specific samples which can be a result of incorrect data interpretation or calculation. Therefore, because not all samples lie within the control limits, it can be suggested that the process is not in control.

Analysis of X-bar & S-chart for Gifts class

This sub-section will analyse the x-bar and s control charts of the data under the gifts class.

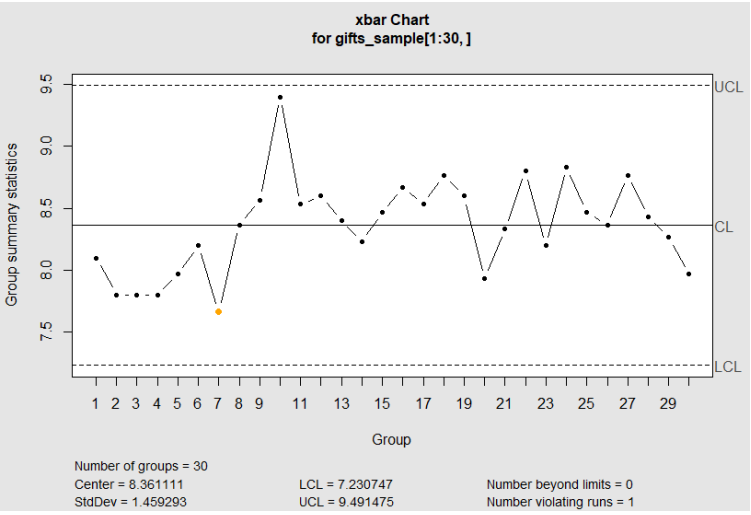


Figure 10: Xbar chart of Gifts sample

In figure 10, we study the sample mean, as well as the central line and control limits for the process, based on the 30 subgroups of the gifts sample. We see that group 7 has a violating run, this means that the data might the data might be invalid which has the potential to make the results unreliable in the sense that the control limits could be greater or lower. We also see that the sample at group 10 is approaching the upper control limit. Overall the process seems to be in control.

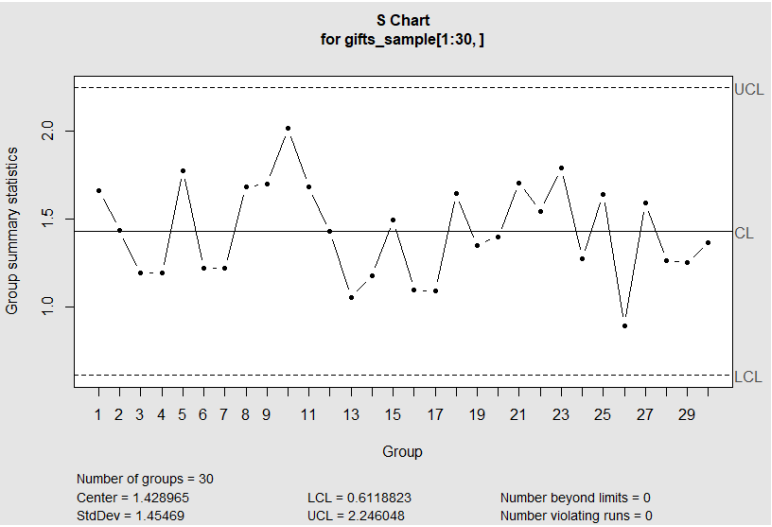


Figure 11: Schart of Gifts sample

Figure 11 shows the sample standard deviations as well as the corresponding central lines and control limits for each subgroup in the gifts sample. There is no apparent trend/pattern in the sample and the process is in control.

Analysis of X-bar & S-chart for Household class

This sub-section will analyse the x-bar and s control charts of the data under the household class.

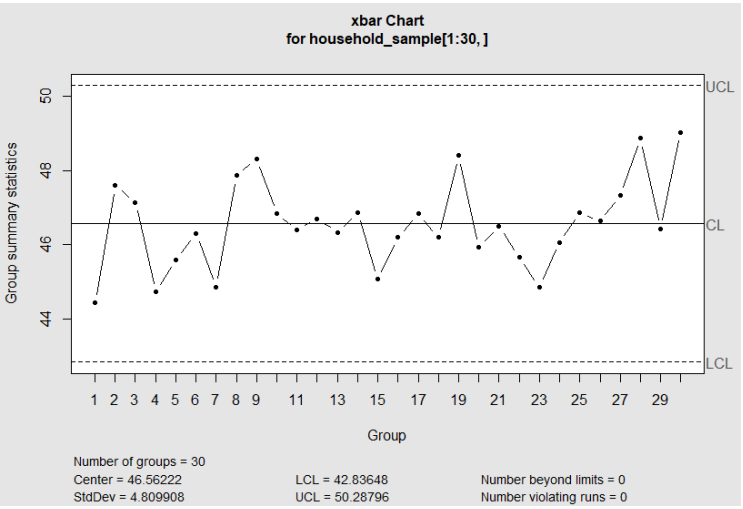


Figure 12: Xbar chart of Household sample

Figure 12 shows no noticeable trend. Furthermore, the number of samples on each side of the central lines is equal. When looking a group 9-13, the data points seem to be closer or even on the centre line which represents the actual process average. The data is within the control limits, the can be regarded as stable.

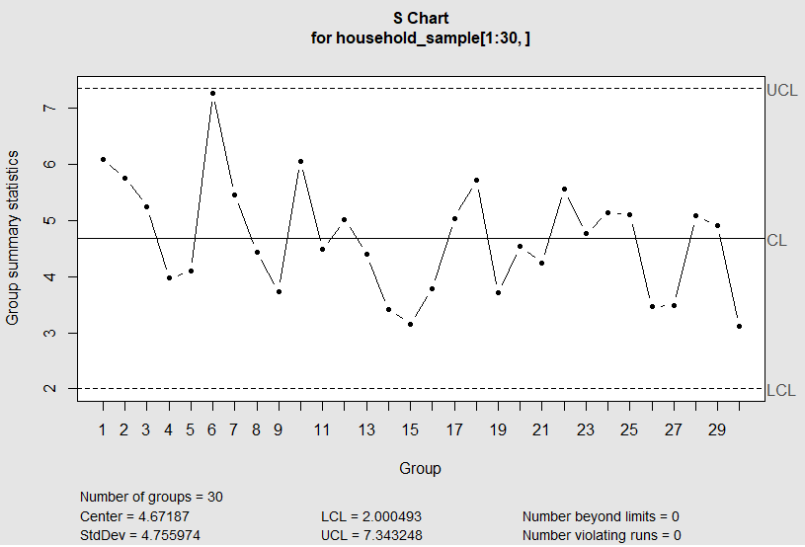


Figure 13: Schart of Household sample

Figure 13 shows the sample standard deviations as well as the corresponding central lines and control limits for each subgroup in the household sample. The graph shows one noticeable trend, at group 1-3, we see that there is a decrease in the summary statistics. It is also noticeable that group 6 has a spike with a sample almost reaching the upper central limit. The overall process is stable.

Analysis of X-bar & S-chart for Luxury class

This sub-section will analyse the x-bar and s control charts of the data under the luxury class.

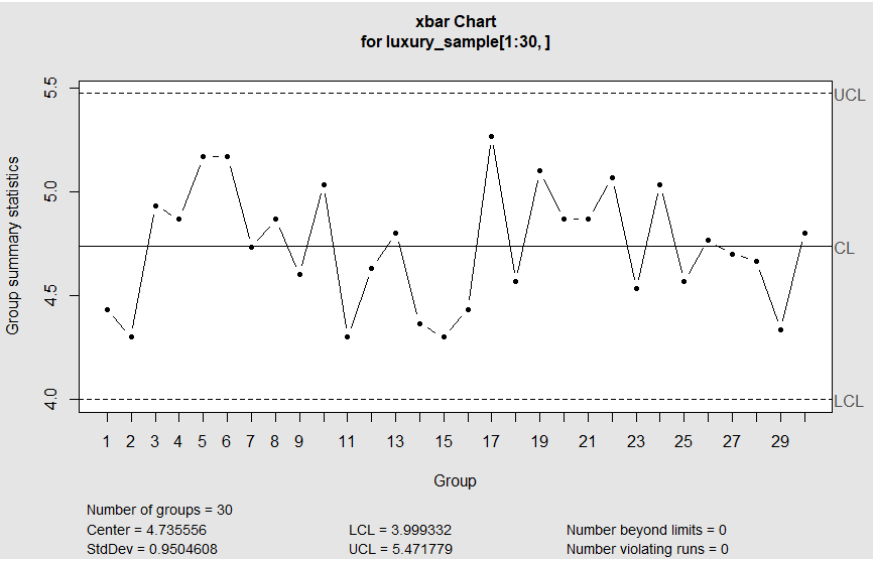


Figure 14: Xbar chart of Luxury sample

Figure 14 shows the sample means of the luxury sample including the central lines and control limits of each subgroup. There is one noticeable trend in group 25-28, we see that there is a decrease in the summary statistic. Moreover, the data is within the central limits therefore, the process is statistically in control.

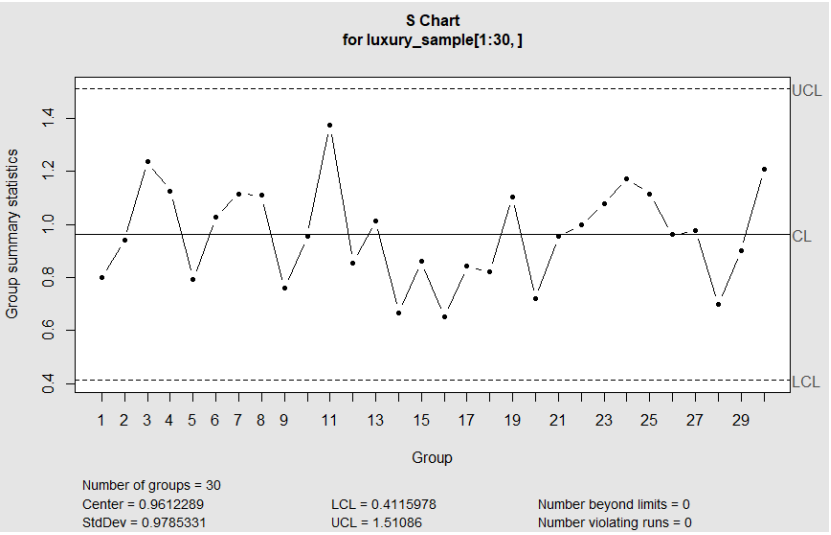


Figure 15: Schart of Luxury sample

In figure 15 we see the sample standard deviations as well as the corresponding central lines and control limits for each subgroup in the luxury sample. There is a pattern in the sample layout at group 21-24, we see that there is a gradual increase in the summary statistic. The sample lies within the control limits meaning the process is in control.

Analysis of X-bar & S-chart for Sweets class

This sub-section will analyse the x-bar and s control charts of the data under the sweets class.

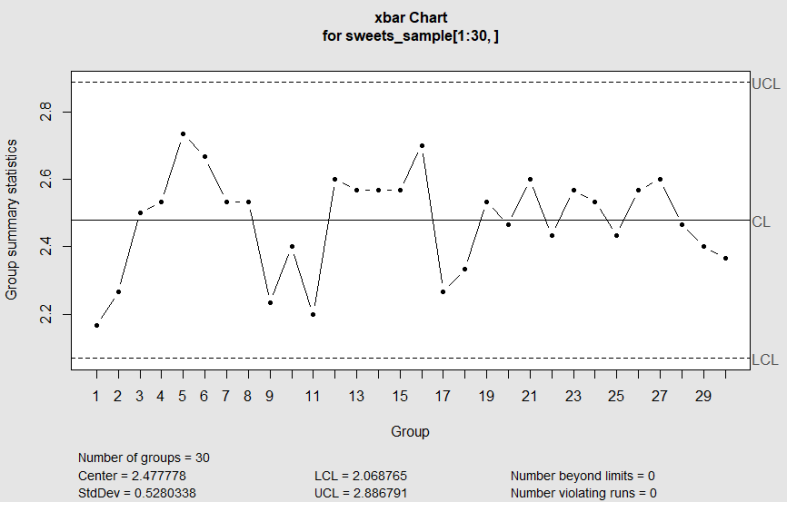


Figure 16: Xbar chart of Sweets sample

Looking at figure 16 we see the sample means of the sweets sample including the central lines and control limits of each subgroup. There are two noticeable trends; at group 12-15 there is gradual decrease in the statistics and eventually they become constant, and at group 25-30, there is a gradual decrease in the statistics. Lastly, the samples range within the control limits meaning the process is table.

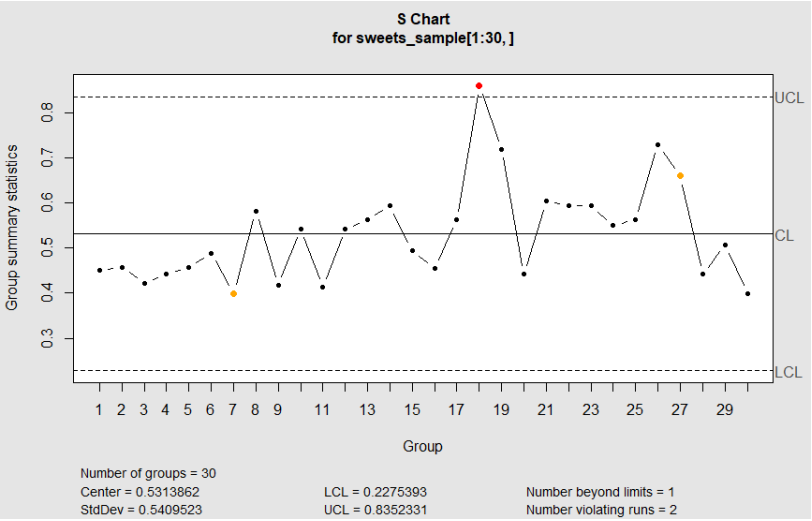


Figure 17: Schart of Sweets sample

In figure 17 we see the sample standard deviations as well as the corresponding central lines and control limits for each subgroup in the sweets sample. The graph shows that at group 18 there is an outlier furthermore, we see that there is violation at group 7 and 27. The violations could be a result of error due to the incorrect capturing of data or calculation errors, which makes the results unreliable for usage. With some values lying outside the control limits, the process is considered unstable.

Analysis of X-bar & S-chart for Technology class

This sub-section will analyse the x-bar and s control charts of the data under the sweets class.

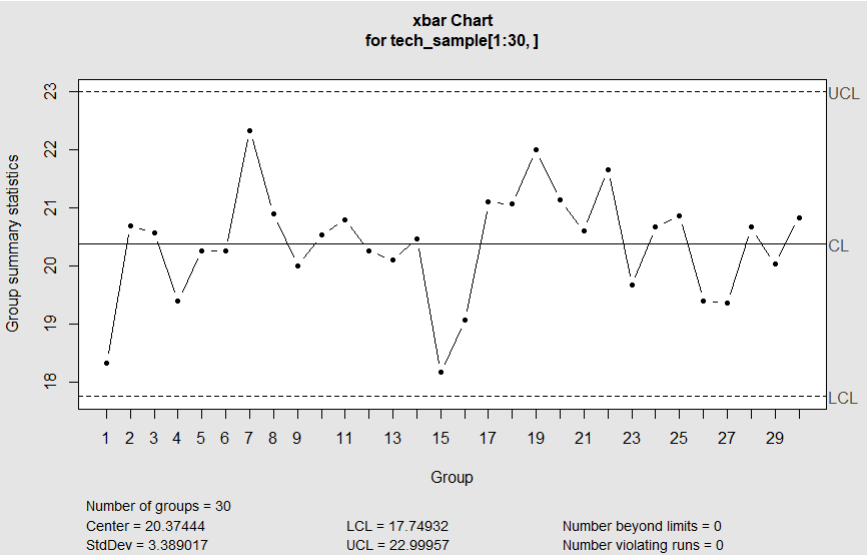


Figure 18: Xbar chart of Technology sample

Looking at figure 18 we see the sample means of the tech sample including the central lines and control limits of each subgroup. There is no noticeable trend in the sample. Lastly, the samples are within the control limits meaning the process is stable

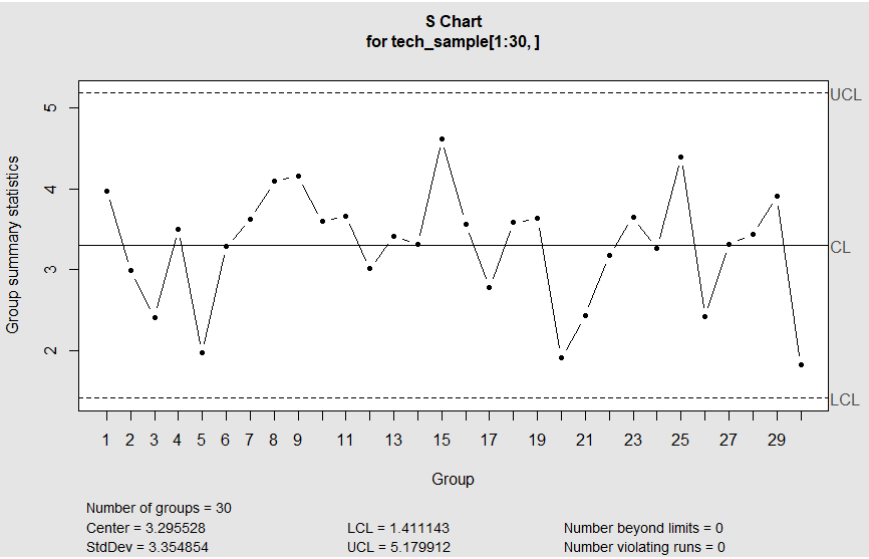


Figure 19: Schart of Technology sample

In figure 19 we see the sample standard deviations as well as the corresponding central lines and control limits for each subgroup in the tech sample. There are no noticeable trends/patterns in the sample layout. The sample values are within the control limits this means the process can be regarded as stable.

Control Charts Table

The following table tables (Table 4 & Table 5) contain the control charts for the delivery process times. A sample of 30 was extracted, each containing 15 sales to construct the centre lines, outer control limits, the 2-sigma-control limits and the 1-sigma-control limits for the X and S charts for the seven classes. The values depicted in the tables below is more or less similar to the values shown in the x-bar and s charts.

Table 4: Control Limit Table for sample 1:30 xbar chart

	classes	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
1	Technology	22.999566	22.124525	21.249485	20.374444	19.499404	18.624364	17.749323
2	Clothing	9.393886	9.252591	9.111295	8.970000	8.828705	8.687409	8.546114
3	Household	50.287961	49.046048	47.804135	46.562222	45.320309	44.078396	42.836484
4	Luxury	5.471779	5.226371	4.980963	4.735556	4.490148	4.244740	3.999332
5	Food	2.698225	2.628817	2.559408	2.490000	2.420592	2.351183	2.281775
6	Gifts	9.491475	9.114687	8.737899	8.361111	7.984323	7.607535	7.230747
7	Sweets	2.886791	2.750453	2.614116	2.477778	2.341440	2.205102	2.068765

Table 5: Control Limit Table for sample 1:30 s chart

	classes1	UCL1	U2Sigma1	U1Sigma1	CL1	L1Sigma1	L2Sigma1	LCL1
1	Technology	5.1799121	4.5517840	3.9236559	3.2955278	2.6673997	2.0392716	1.4111435
2	Clothing	0.8664496	0.7613819	0.6563142	0.5512465	0.4461789	0.3411112	0.2360435
3	Household	7.3432478	6.4527886	5.5623295	4.6718703	3.7814111	2.8909519	2.0004927
4	Luxury	1.5108600	1.3276496	1.1444392	0.9612289	0.7780185	0.5948081	0.4115978
5	Food	0.4371911	0.3841763	0.3311615	0.2781467	0.2251319	0.1721171	0.1191023
6	Gifts	2.2460482	1.9736872	1.7013262	1.4289652	1.1566042	0.8842432	0.6118823
7	Sweets	0.8352331	0.7339508	0.6326685	0.5313862	0.4301039	0.3288216	0.2275393

Part 4: Optimisation of Delivery Processes

Sample means

Table 6: Upper and lower control limits outliers

Class	Total found	First	Second	Third	Last3	Last2	Last1
Clothing	17	455	702	1152	1677	1723	1724
Household	400	252	387	629	1335	1336	1337
Food	5	75	633	1203	1467	1515	NA
Technology	17	37	398	483	1872	2009	2071
Sweets	5	942	1104	1243	1294	1403	NA
Gifts	2290	213	216	218	2607	2608	2609
Luxury	434	142	171	184	789	790	791

Sample standard deviations

Table 7: Summary of -0.3 & +0.4sigma consecutive values and ending sample numbers

CLASS	CONSECUTIVE SAMPLE BETWEEN -0.3 & +0.4 SIGMA	ENDING SAMPLE NUMBER
Clothing	16.2	600.9
Household	8.2	750.9
Food	12.2	1564.9
Technology	11	518.7
Sweets	14.2	698.9
Gifts	0	0
Luxury	1	2.7

Estimating Likelihood of Type I error

In order to estimate the likelihood of making a Type I (Manufacturer's) error for A and B, it is essential to understand that this is a theoretical value that holds for any process. It assumes the following; H_o: The process is in control and centred on the centreline calculated using the first 30 samples.

H_1: The process is not in control and has moved from the centreline or has increased or decreased in variation. This can be regarded as a Type I error.

Table 8: Type I and II

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

$A = pnorm(-3)^2 = 0.002699796$

Central Processing for best profit

When using the individual delivery times, it is imperative to use all the data that is available within the sample in order to calculate the cost reduction. The following considerations are made:

- R329/item-late/hour is lost in sales if technology items are delivered slower than 26 hours
- It costs R2.5/item/hour to reduce the average time by one hour.

This evaluation will analyse the number of hours to centre the delivery process for the best profit. An assumption is made that the process output distribution retains its shape when moving the centre and costs less (R-2.5/item/hour) if the delivery time is increased.

The starting point is to calculate the current additional cost using the supplied dataset. The current cost comparison purposes = 329*(item-late-hours). From the equation we can reduce the delivery time by x hours for every item by subtracting the variable deemed as 'x' from the 'current' time, which would give the new time.

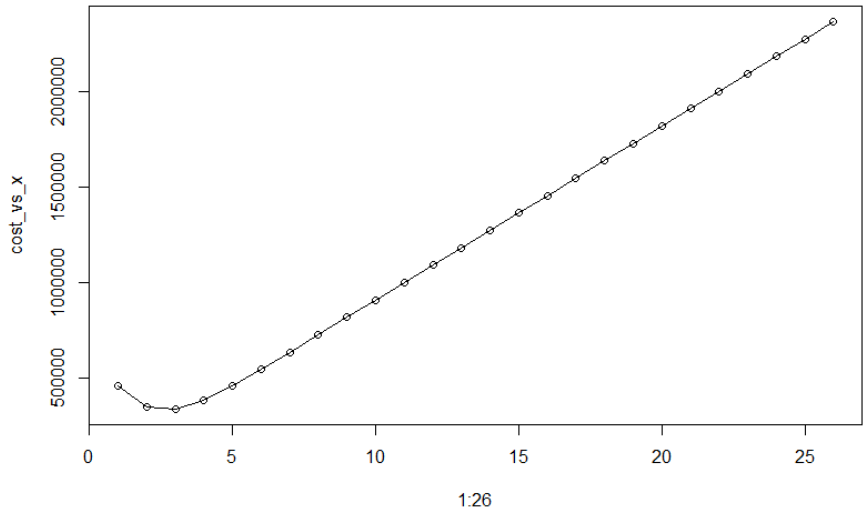


Figure 20: Delivery.Time graph for Technology Class

By looking at figure 20, we the lowest cost is R340892.5 and by shifting the process by 3 days, we reach an optimal. Therefore, the x value with the lowest cost is 3 days.

Estimating Likelihood of Type II error

In this section, the Type II error was calculated and determined, the probability of a type II error being made for A in Technology is 0.997. This means that the probability of assuming a process is running well when it isn't doing well is quite high. As such, from the data gives failures in the process are not easily identifiable.

Part 5: DOE and MANOVA

In part 5 of this report, we will analyse the difference between multiple groups in the independent variable by conducting a hypothesis test using the Multivariate Analysis of Variance (MANOVA) on the valid data. The test will specifically study the influence that 'Why.Bought' variable has on Age, Year, Price and Delivery.Time in the sales dataset.

We define the following parameters:

μ_1 = mean value of the browsing reason.

μ_2 = mean value of the email reason.

μ_3 = mean value of the random reason.

μ_4 = mean value of the recommended reason.

μ_5 = mean value of the spam reason.

μ_6 = mean value of the website reason.

H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ (the Why.Bought variable has no impact on the validData attribute)

H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5 \neq \mu_6$ (the Why.Bought variable has an impact on the validData attribute)

Result Analysis of MANOVA

In figure 21, we see that there are no outliers meaning there is no variation in the dataset. Upon analysis, it is also evident that the mean value for each Why.Bought is not equal. Therefore, the alternative hypothesis is applicable to the study. H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5 \neq \mu_6$, this substantiates that the sales relating to age are dependent on the Why.Bought variable.

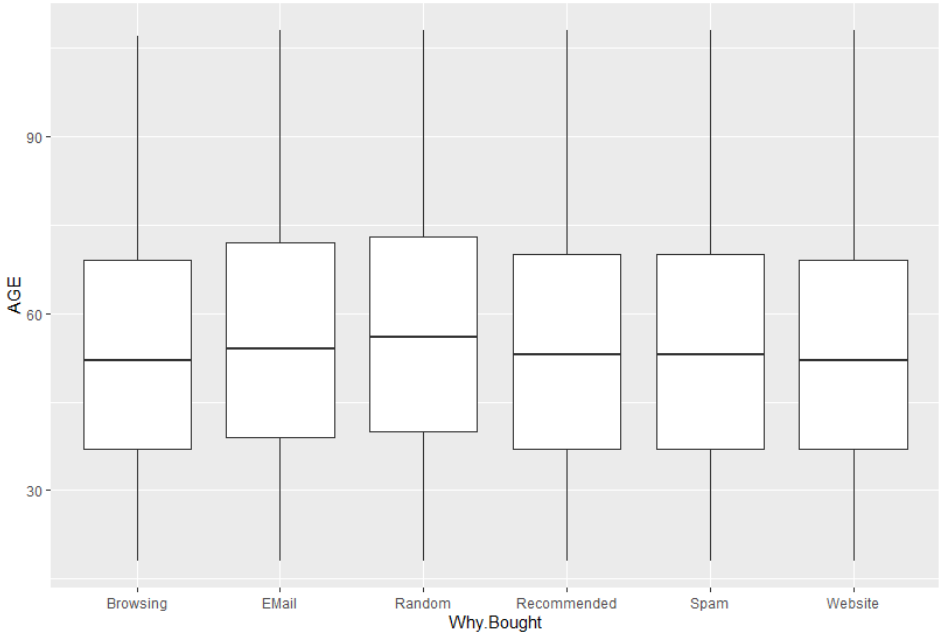


Figure 21: Boxplot showing Age relative to Why.Bought

In figure 22, we also observe that there are no outliers meaning there are no drastic variation in the dataset. When looking at the graph, it is evident that the mean value for each Why.Bought is equal. Therefore, the null hypothesis is applicable to the study. H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$, this proves that the sales relating to year are independent on the Why.Bought variable.

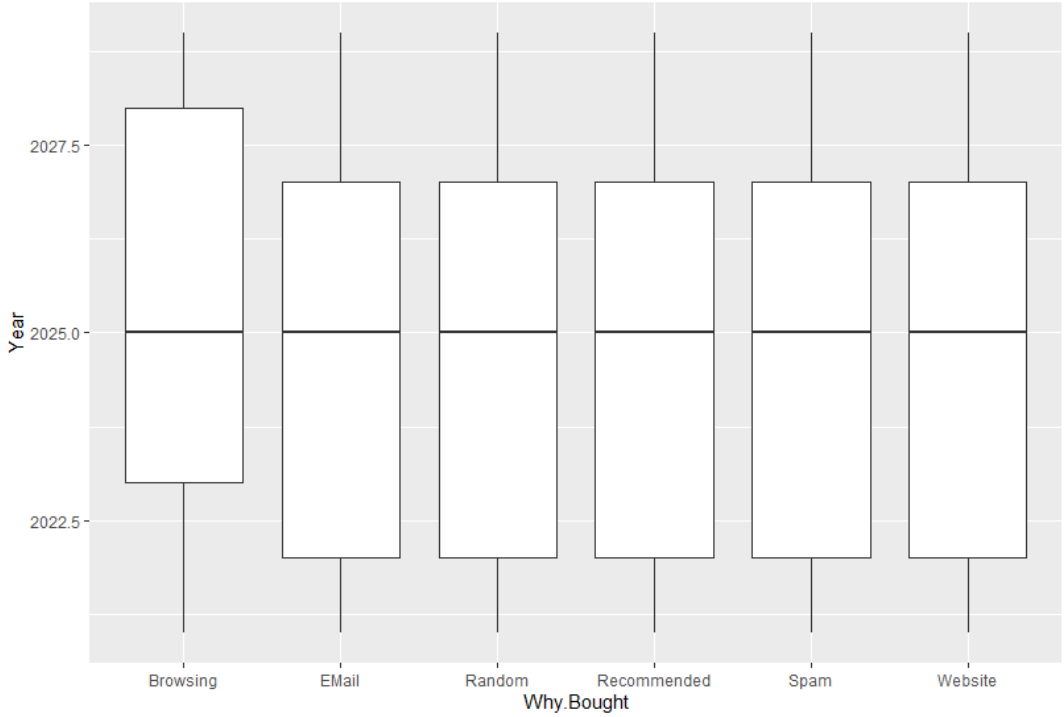


Figure 22: Boxplot showing Year relative to Why.Bought

When evaluating figure 23, we also observe that there are plenty of outliers meaning there are drastic variation in the dataset (there many small values and big values). This variation makes the data unreliable as the data may be captured incorrectly. When looking at the graph, it is evident that the mean value for each Why.Bought is not equal. Therefore, the alternative hypothesis is applicable to the study. H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5 \neq \mu_6$, this proves that the sales relating to price is dependent on the Why.Bought variable.

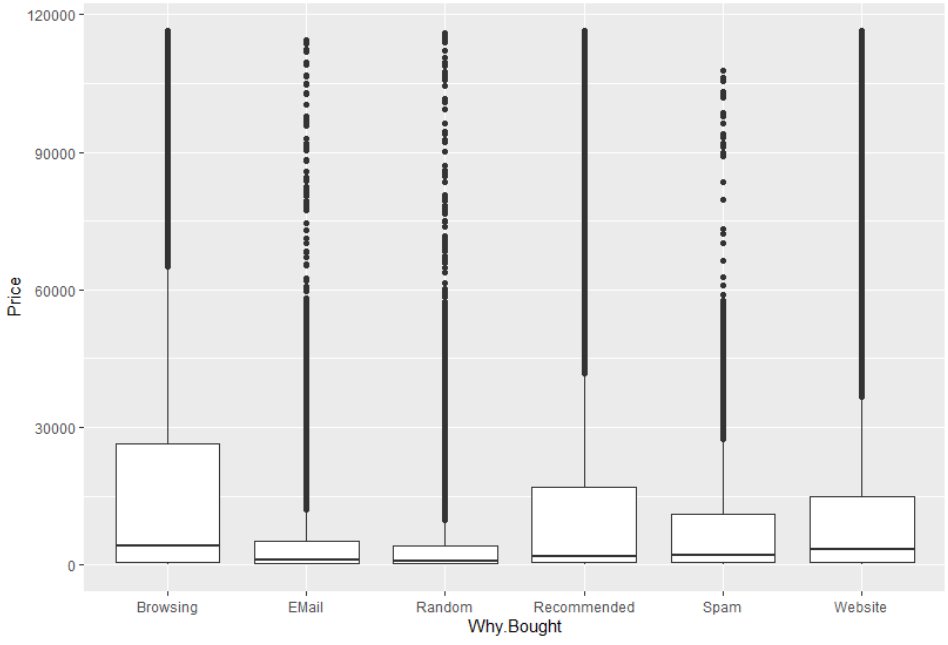


Figure 23: Boxplot showing Price relative to Why.Bought

Figure 24, we also observe that there are plenty of outliers meaning there are drastic variation in the dataset (there many small values and big values). This variation makes the data unreliable as the data may be captured incorrectly. The graph shows that the mean value for each Why.Bought is not equal. Therefore, the alternative hypothesis is applicable to the study. H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5 \neq \mu_6$, this proves that the sales relating to delivery.time is dependent on the Why.Bought variable.

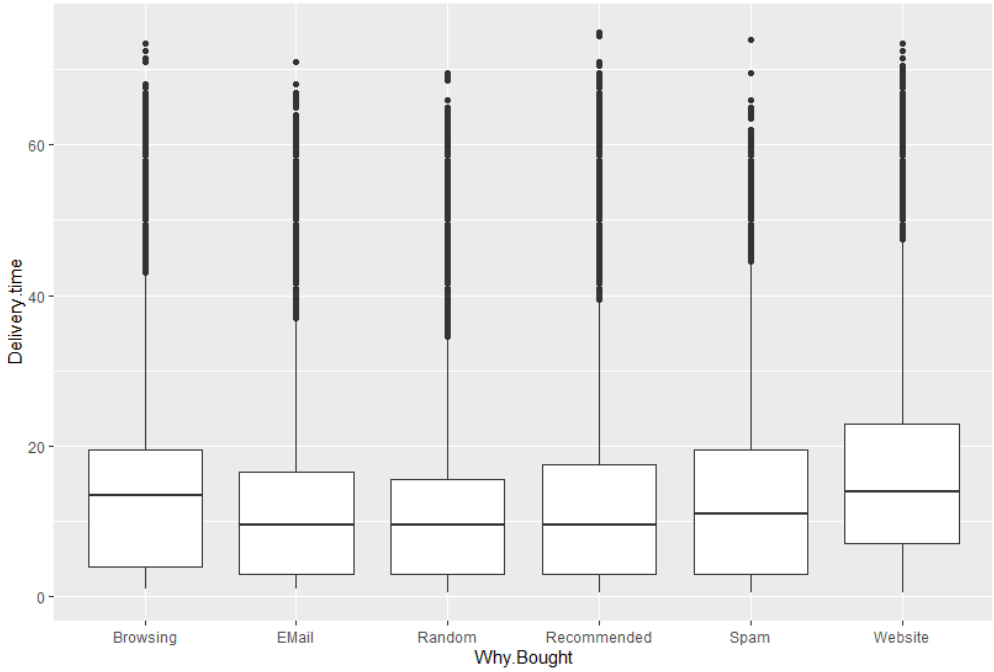


Figure 24: Boxplot showing Delivery.Time relative to Why.Bought

Part 6: Reliability of the Service and Products

In part 6, we will look at the concept of reliability particularly of services and products by answering the exercises in the textbook and further modelling the delivery process to estimate reliability.

Taguchi Loss Function

Problem 6

Tolerance = +- 0.04

Scrap cost = $45

Therefore, $k = \frac{45}{(0.04)^2}$

= $ 28125

The Taguchi Loss is: $L(x) = \$\ 28125(x-T)^2$

Problem 7

a. The approach is similar to that in problem 6, we need to find k constant. Therefore,

$$k = \frac{35}{(0.04)^2}$$
$$= \$\ 21875$$

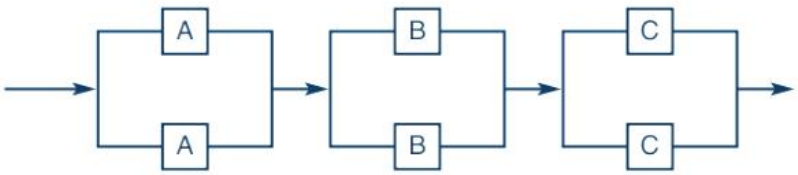
The Taguchi Loss is: $L(x) = \$\ 21875(x-T)^2$

b. In this problem the process deviation is reduced to 0.027 cm.
The Taguchi can be calculated as follows:

$$L(x) = \$\ 21875(0.027 - 0.04)^2$$
$$= \$\ 21875 \times 0.000169$$
$$= \$\ 3.697$$

Reliability

Problem 27



Production System

The reliabilities of the machines are as follows:

Machine	Reliability
A	0.85
B	0.92
C	0.90

Figure 25: Reliability problem

a. The reliability of the system can be calculated as follows:

System reliability (R_s) = $R_A \times R_B \times R_C$

= $0.85 \times 0.92 \times 0.90$

= 0.7038

b. The improvement of the system can be calculated as follows:

System reliability (R_s) = $R_A \times R_B \times R_C$

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

= $0.9775 \times 0.9936 \times 0.99$

= 0.961

Improvements = $(0.961- 0.7038) \times 100$

= 25.72

From the results obtained, we can deduce that the reliability of the system improves tremendously when both machines work at each stage.

Binomial Probabilities

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 were only 20 vehicles available was 190, only 19 vehicles were available for 22 days, only 18 vehicles for 3 days and 17 vehicles for 1 day.
There are also 20 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, only 19 drivers for 6 days and only 18 drivers on 1 day.

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

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

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

Using R programming, the p-value was determined to be 0.02.

Using the above formula on RStudio, it was concluded that when increasing the vehicle fleet to 21, the number of days in which reliable delivery time occurs will be 335 days.

Conclusion

The sales data that was provided consisted of missing values and unusable data (negative price values). The data needed to be cleaned in order to ensure that the analysis produced quality results that are reliable. The data analysis was done using R programming to compile central tendencies (control limits etc) that each feature within the data contained. The analysis also included the computation of graphs (x-bar charts, s charts, boxplots etc) to visualise the distribution of the data in order to make informed decisions.

When evaluating the Xbar-charts and the Sc control charts, it can be concluded that the business is generally capable of meeting the specifications. In instances where the business cannot (where the business had violations), verification and validation measures need to be implemented to ensure that the data is correct and dependable.

References

Dirkse-van Skhalkwyk TG, Course document on Sunlearn, 2022

Evans, J.R. & Lindsay W.M. Managing for Quality and Performance Excellence. 10th Edition. United States: Conveo