UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

Quality Assurance ECSA Report

Jessica Masefield (23596679)

October 2022

INDUSTRIAL
ENGINEERING
Stellenbosch University

Contents

List of Tables

List of Figures

1 Introduction

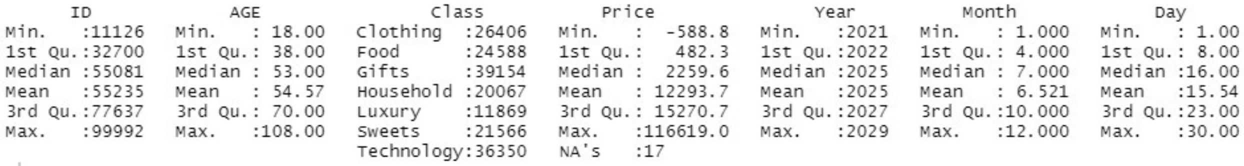
This report was compiled in partial fulfilment of the requirements of the Stellenbosch University, Faculty of Industrial Engineering's Quality Assurance 344 course. Data for analysis was provided in the Salestable2022 file comprising 180 000 sales records from the Client's online business. Variables include a client ID, age, class, price, year, month, day, delivery and why bought.

The data was first cleaned and arranged in an organised manner to allow analysis. Data was separated into valid and invalid data and process control limits were calculated to determine whether the process is in statistical control. The data was used to assist with optimising delivery processes and to determine and reduce the cost of late deliveries.

2 Data Wrangling

To analyse the data it was first arranged and cleaned into a useful form. A data quality plan was then created to determine any irregularities in the data (Das, Datta, and Chaudhuri, 2018). The numeric and categorical data that was was provided was sorted and ordered and is summarised in table 1 below.

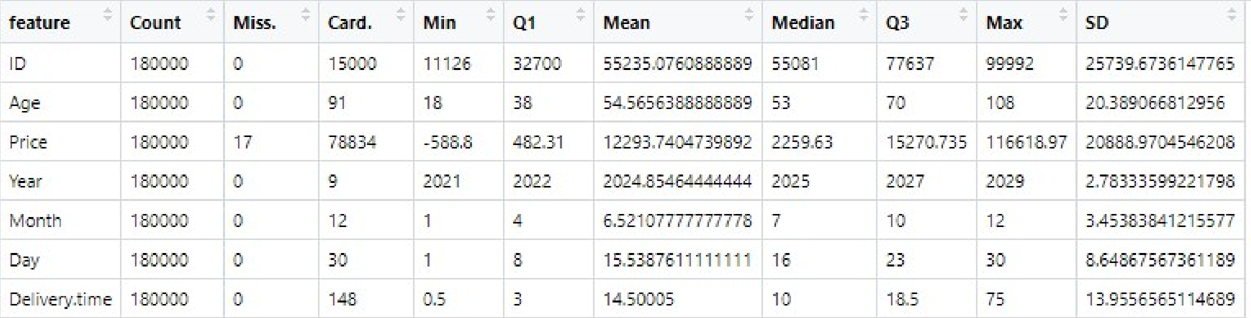
Table 1: Data Information: Date ID, Age, Class, Price and Date



```
       ID              AGE                  Class            Price                Year          Month             Day       
 Min.   :11126   Min.   : 18.00   Clothing  :26406   Min.   :   -588.8   Min.   :2021   Min.   : 1.000   Min.   : 1.00  
 1st Qu.:32700   1st Qu.: 38.00   Food      :24588   1st Qu.:    482.3   1st Qu.:2022   1st Qu.: 4.000   1st Qu.: 8.00  
 Median :55081   Median : 53.00   Gifts     :39154   Median :   2259.6   Median :2025   Median : 7.000   Median :16.00  
 Mean   :55235   Mean   : 54.57   Household :20067   Mean   : 12293.7   Mean   :2025   Mean   : 6.521   Mean   :15.54  
 3rd Qu.:77637   3rd Qu.: 70.00   Luxury    :11869   3rd Qu.: 15270.7   3rd Qu.:2027   3rd Qu.:10.000   3rd Qu.:23.00  
 Max.   :99992   Max.   :108.00   Sweets    :21566   Max.   :116619.0   Max.   :2029   Max.   :12.000   Max.   :30.00  
                                  Technology:36350   NA's   :17                                                          
```

Table 1 shows the minimum, maximum, mean, median and first and third quarters for each numeric feature and shows the number of instances for categorical features. Two main irregularities are evident which both occur in price; the minimum shows that there are negative values of price which must be removed as it is not possible to have a negative price, and it is also shows that price has 17 not applicable (NA) values.

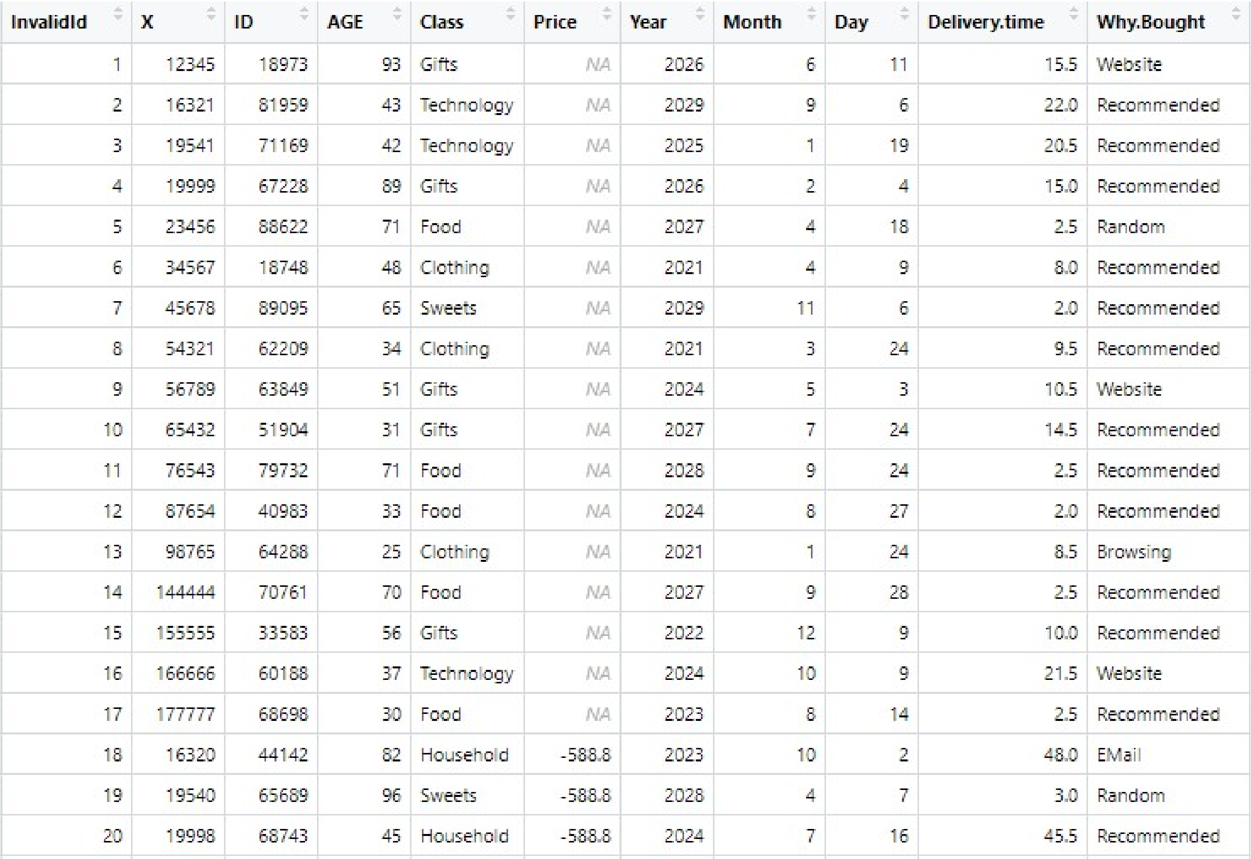
Table 2: Data Quality Plan



feature	Count	Miss.	Card.	Min	Q1	Mean	Median	Q3	Max	SD
ID	180000	0	15000	11126	32700	55235.0760888889	55081	77637	99992	25739.6736147765
Age	180000	0	91	18	38	54.5656388888889	53	70	108	20.389066812956
Price	180000	17	78834	-588.8	482.31	12293.7404739892	2259.63	15270.735	116618.97	20888.9704546208
Year	180000	0	9	2021	2022	2024.85464444444	2025	2027	2029	2.78333599221798
Month	180000	0	12	1	4	6.52107777777778	7	10	12	3.45383841215577
Day	180000	0	30	1	8	15.5387611111111	16	23	30	8.64867567361189
Delivery.time	180000	0	148	0.5	3	14.50005	10	18.5	75	13.9556565114689

Any data that has NA values or negative price values must be removed from the data set since these are considered compromised or incomplete (Stellenbosch University, 2022). The incomplete data was stored in a different data set keeping the original index of the data intact for use if required by the Client. Table 3 shows the first 20 invalid data records and table 4 below shows the first 20 of the valid data records. Both tables have the new key and the original key in the first and second columns.

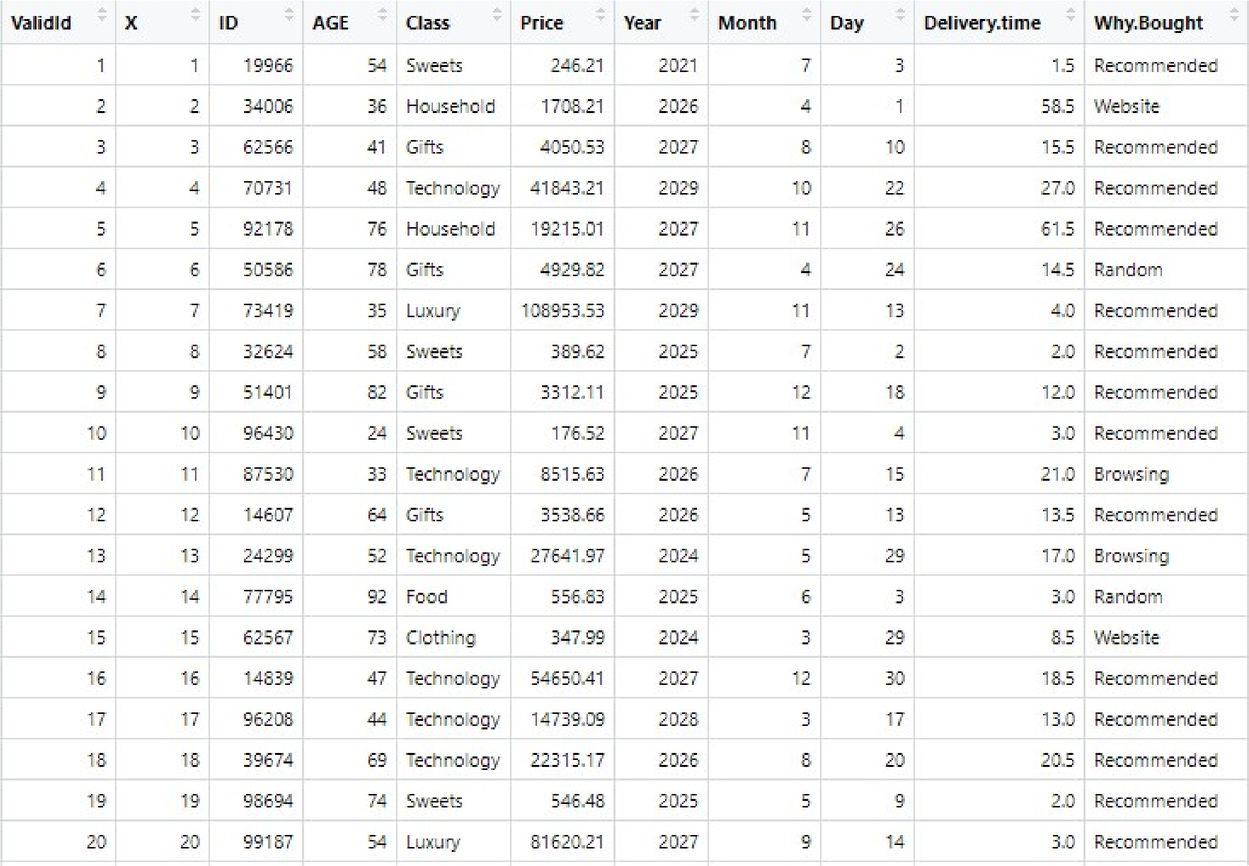
Table 3: Invalid Data



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

Price is the only feature with invalid data, i.e. NA and negative price values which were removed.

Table 4: Valid Data



ValidId	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
20	20	99187	54	Luxury	81620.21	2027	9	14	3.0	Recommended

3 Descriptive Statistics

Data-set characteristics were determined with the use of histograms, bar plots and statistical calculations using R. For each feature the central tendency and variation is examined to understand the types of values within a feature. The graphs are also used to determine distribution patterns and trends within the data for both categorical and numerical features.

Categorical data is a collection of information that is divided into groups where as numeric data is any real numeric value (Zstatistics, 2019). Our data-set has 6 numeric features and 2 categorical features and 2 features that will not be analyzed as they do not give any value to the problem namely, x and ID. Each feature will be analyzed below (FormPlus, 2022).

3.1 Price

Price is a numeric feature. Figure 1 shows the distribution for price and figure 2 shows the box plot for price distribution. The shape of the graph in figure 1 shows that price is skewed to the right showing there are outliers which is also shown in the long whisker in the higher price values in figure 2. This indicates that there is a tendency towards lower numbers with a few higher numbers as outliers.

The price values did not show any seasonality as they were the same irrespective of the day, month or year.

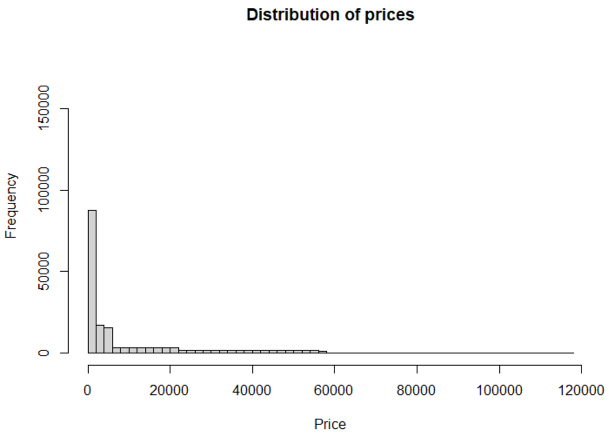


Figure 1: Price Distribution

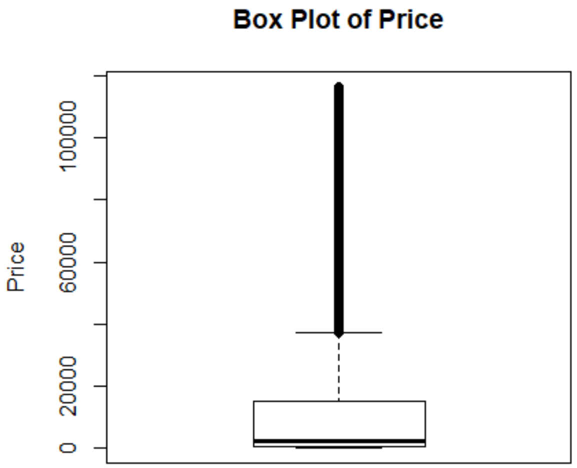


Figure 2: Price Box Plot

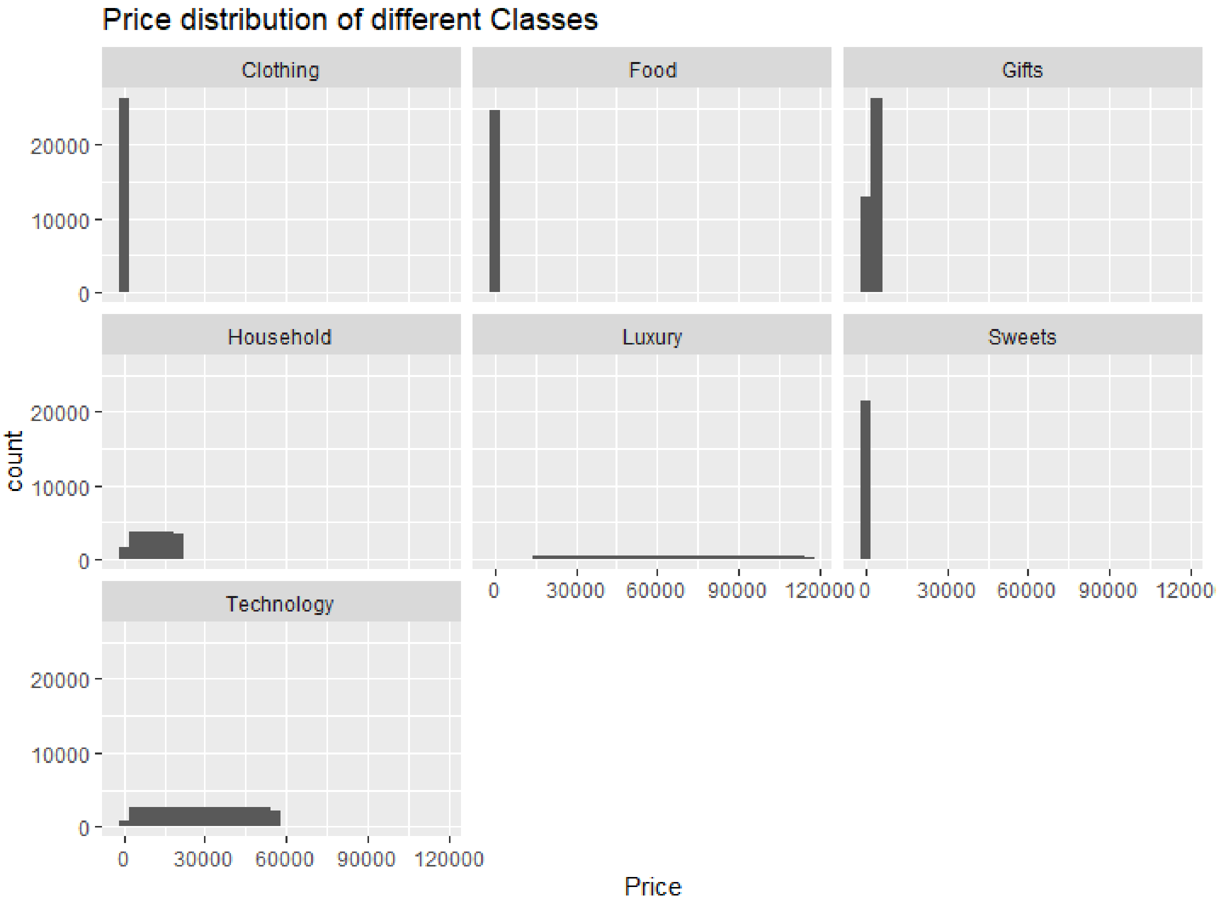


Figure 3: Price distribution of different Classes

The set of graphs above shows the price distribution of each class. Most classes have a high frequency towards lower values except for luxury items which have a low frequency for a large number of prices this shows that the luxury items is the class which is skewing the price data.

Luxury items and technology are the most expensive compared to the other classes with a low frequency of purchase. Clothing, food, gifts and sweets have a high frequency of purchase and lower costs.

3.2 Age

Age is a numeric feature with a minimum of 18 and maximum of 108 which aligns with the nature of the feature as that is what we would expect from the sample of people purchasing goods.

From the figure 3 below, it can be seen that the age feature is uni-modal with a slight skew right and higher than expected values from 2 age groups, which indicates that there is a tendency towards lower numbers with a slightly less higher numbers. Most ages fall within the 30-55 range which can be expected as this is the most active age economically. The mean age is 54.57 and the median is 53.00. The maximum age of 108 is unlikely and could be an outlier.

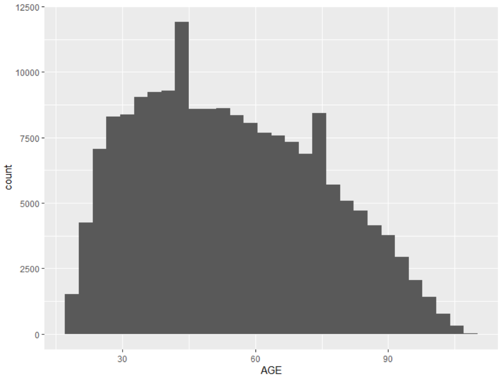


Figure 4: Age Distribution of Data

Figure 5 is a box plot showing the age of the customers purchasing the different class of goods. It can be seen that younger people buy technology, clothes and luxury goods whereas older people buy food, gifts and sweets.

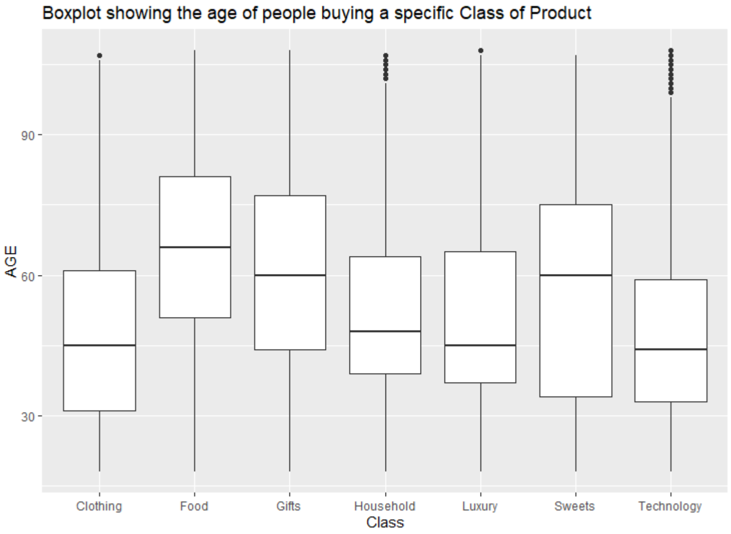


Figure 5: Box Plot Age per Class

3.3 Class

Class is a categorical feature and the figure below shows the frequency of each class. There are 7 classes of data, namely clothing, food, gifts, household, luxury, sweets and technology, with the mode being Gifts. This distribution shows that gifts were purchased the most frequently followed by technology. There is a reasonably even distribution of class of goods purchased with luxury goods being purchased the least frequently.

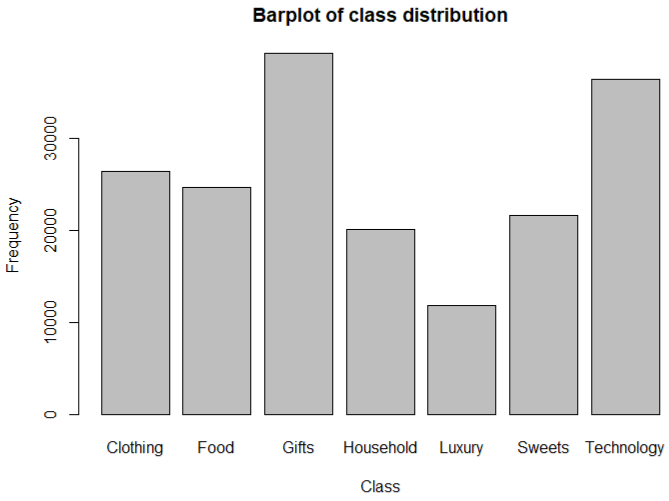


Figure 6: Distribution of classes

3.4 Why Bought

Why items where bought is a categorical feature the graph below shows the number of instances that each class occurred, the classes are as follows, browsing, email, random, recommended, spam and website. With recommended as the mode, meaning that the most common reason for items being bought is due to recommendations.

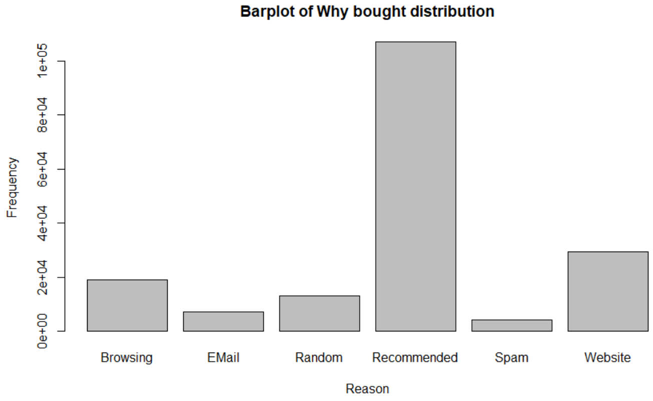


Figure 7: Why bought

3.5 Time

There are three different features of time; years, months and days for the times at which people made purchases. All the data for time is uni-modal and uniformly distributed and is equally likely to be any day or month. This is seen by the box shape the distribution of data creates, shown in the figures 6 and 7 below.

The only exception is for the feature years which is close to uniformly distributed except for 2021. The figure 8 below shows that more data was collected in the 2021 year with almost half the number of records in the following years. For the feature years the data spans the years 2021 to 2019.

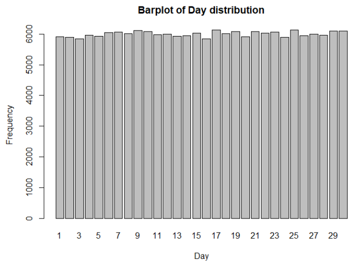


Figure 8: Day Distribution of Data

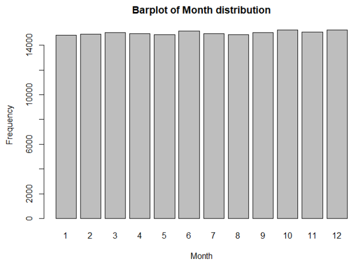


Figure 9: Month Distribution of Data

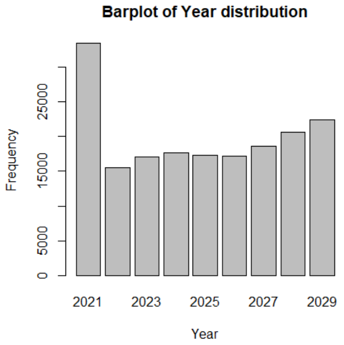


Figure 10: Year Distribution of data

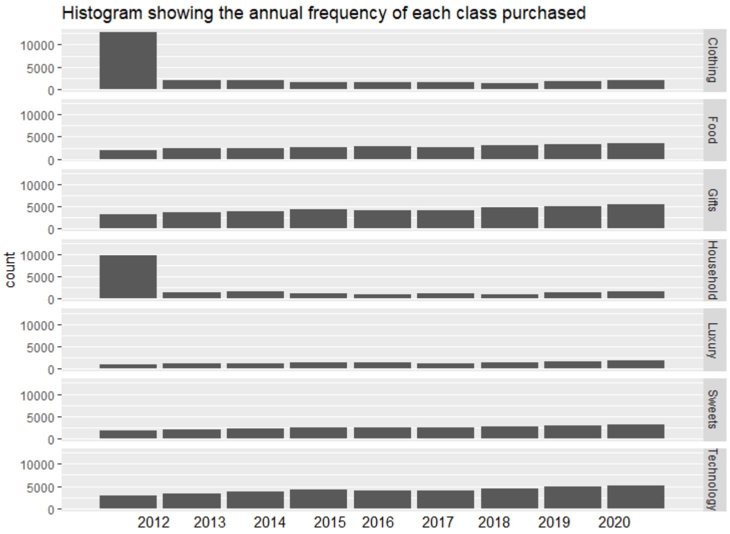


Figure 11: Histogram showing annual frequency of purchase

Figure 11 shows the frequency of purchase per class from between 2012 and 2020. Clothing and household goods were purchased far more frequently in 2012 than subsequent years. There is a slight upward trend in the purchase of technology, gifts and food.

3.6 Delivery Time

Delivery time is a numeric feature. From the graph below it can be seen that the delivery times have three regions around which they group, the first two between approximately 0 to 25 and the third between 40 to 60. This is shown in the scatter plot in figure 9 and when shown as distribution we can see the 3 peaks clearly i.e. a multi-modal distribution.

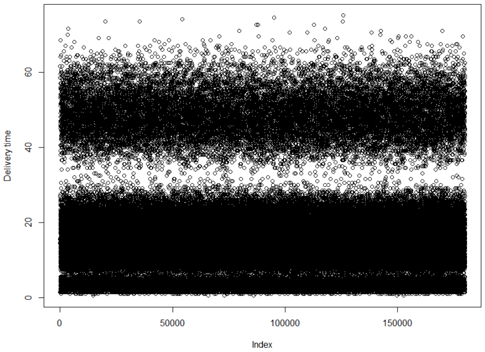


Figure 12: Delivery time scatter plot

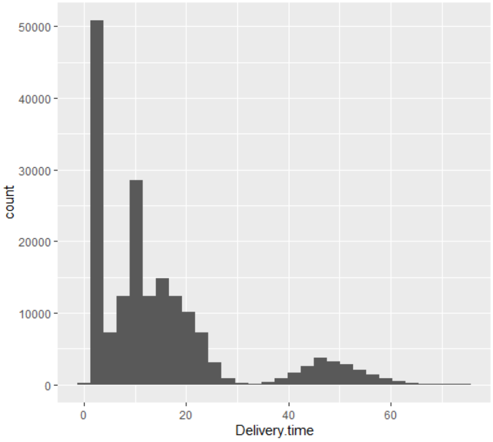


Figure 13: Delivery time distribution

Delivery time can be further split into its separate classes as shown in figure 14.

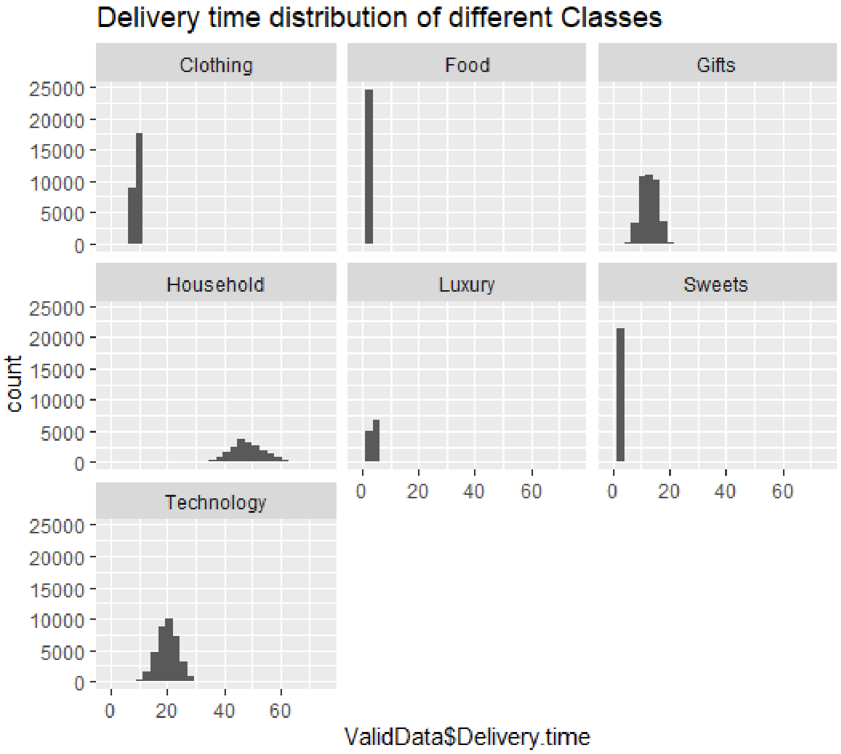


Figure 14: Delivery time of different classes

This shows a range of different types of distributions. Clothing, food, sweets, and luxury have a large number of instances in a small range and short delivery times, while technology and gifts show a normal distribution and household is triangular. Households items have the longest delivery times which is likely to result in consumer complaints.

3.7 Capacity calculations

Capacity calculations can be defined as the ability of a process to meet specifications (Stellenbosch University, 2022). From a designated point specified the process can deviate a certain amount. Although statistical control systems charts, to follow, indicate weather a process is stable they do not indicate whether the process is capable of producing acceptable output as per voice of customer (VOC) and whether the process is performing to its potential capacity.

The capability potential (C_P) and the capability performance (C_{PK}) illustrate a process's ability to meet set specifications (Stellenbosch University, 2022). The C_P ratio shows how well the process spread fits into the specification range and is expressed as a 6 standard deviation of the process (Stellenbosch University, 2022). The C_P is determined by dividing the specification limit i.e. the voice of the customer (VOC) by the process spread i.e. voice of process (VOP) (Stellenbosch University, 2022). The C_{PK} ratio also measures what C_P does and in addition it measures how close the process mean is to the target value of the specification. To do this, two intermediate ratios are used as shown below: Process Capability Indices include C_P, C_{PK}, C_{PU} and C_{PL} with the following formulas (Stellenbosch University, 2022):

$$C_P = (USL - LSL)/6\sigma$$

$$C_{PU} = (USL - \mu)/3\sigma$$

$$C_{PL} = (\mu - LSL)/3\sigma$$

$$C_{PK} = \min(C_{PL}, C_{PU})$$

In our case the USL = 24 days and the LSL = 0. A LSL is logical because these calculations are for the delivery time, thus you cannot have a time less than zero.

Using these values and the calculations above this results in:

$C_P = 1.142207$

The C_P value of the delivery times was calculated and compared to the target value 1 which is the value at which the process is deemed capable of meeting specifications (Stellenbosch University, 2022).

A C_{PK} value of less than 1 indicates that that the process is not capable of meeting the specifications. To improve C_{PK} we may either center the process on the target (by adjusting a dial or setting) or reduce the variation and spread of the process (by some fundamental redesign of the process or technologies).

We must now Verify that process variability is stable, i.e. no out-of-control patterns on the control charts (Stellenbosch University, 2022).

$C_{PU} = 0.3796933$

$C_{PL} = 1.90472$

$C_{PK} = 0.3796933$

4 Statistical process control

Statistical Process Control (SPC) is based on the central limit theorem (CTL) (Stellenbosch University, 2022). The central limit theorem for a sample means says that if you draw large samples and calculating their means, the sample means form their own normal distribution (the sampling distribution) even though the original individual outcomes are uniform (Stellenbosch University, 2022). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by, the sample size. The variable n=30 is the number of values that are averaged together, not the number of times the experiment is done. Since the sample means are normally distributed we may calculate its probabilities from the normal curve or use normal Z table.

4.1 Phases in control chart

SPC has two distinct phases; initialise and control.

Stage 1 : Initialize the chart

The manager must actively manage the process to an error-free period, while she collects enough data (typically 25-30 samples). The samples are then used to calculate the centre-line and control limits for phase 2. The σ_1 and σ_2 (upper and lower) limits divides the area between the CL and UCL in three equal parts. Using the formulas $U\sigma_1 = CL + (UCL - CL)/3$ and $U\sigma_2 = CL + (UCL - CL)/3 * 2$. All the samples used during initialisation must be between the LCL and UCL, there that ones not are removed and re-calculate the centre-line and control limits before starting phase 2 (Stellenbosch University, 2022).

Phase 2 : The control phase

Once the manager is certain she understands what the process must do, she steps away and lets the personnel run the process on their own; they have to use the control limits that were calculated during the initialise phase! The system gives feedback and workers or AI will flag any process that seems to be out of control. She may then step in to fix the process or motivate the personnel. When these patterns or flags or the rules occur in your SPC charts then investigation is required. Patterns in Control Charts or Rules for Interpreting Control Charts (Out-Of-Control Signals):

1. Any point (sample mean) beyond the control limits.

2. Seven (or more) consecutive points above or below the center-line.

3. Seven consecutive increasing or decreasing points

4. Two out of three beyond two sigma on the same side of the center-line.

5. Four out of five beyond one sigma on the same side of the center-line.

6. Ten out of 11 consecutive points on the same side of the center-line.

7. Twelve out of 14 consecutive points on the same side of the center-line

8. A series of points ”hugging” the center-line (may be that the process improved and that the manager may implement the improvements at other processes as well! This is a good signal, unless someone is fudging the data.)

9. A series of points ”hugging” the control limits

10. Fourteen consecutive points alternating up and down (saw-tooth pattern)

11. Any non random pattern (such as a cycle)

4.2 Initialize Control Charts

The control charts are being designed to control the delivery times for each class of product. The samples are collected by ordering the data in chronological order, and splitting the data into their respective classes. From this we can initialize our control charts from the first 450 data points from 30 samples of 15 instances each by calculating the $\mathrm{UCL}, \mathrm{U}\sigma_2, \mathrm{U}\sigma_1, \mathrm{CL}, \mathrm{L}\Sigma_1, \mathrm{L}\sigma_2, \mathrm{LCL}$, using the formulas below:

$$\sigma_1 = \mathrm{CL} + (\mathrm{UCL} - \mathrm{CL})/3$$

$$\mathrm{U}\sigma_2 = \mathrm{CL} + (\mathrm{UCL} - \mathrm{CL})/3 * 2$$

The graphs shown below are initialized x and s control chart values for each class. The first 30 samples are used to obtain the upper, lower and center lines. These are shown with red being the lower and upper limits and green being the center line. The samples are also plotted using circles to depict where the sample lies within the control chart.

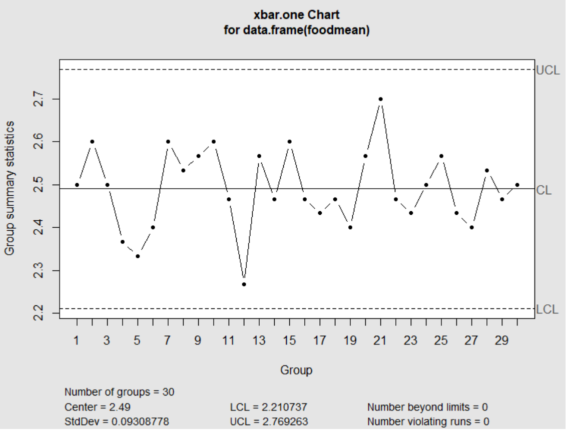


Figure 15: Food x Bar

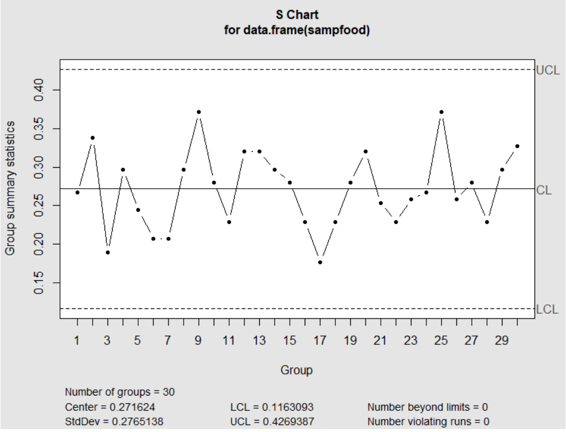


Figure 16: Food s Bar

From the class food shown in figures 15 and 16 we can see that the x and s values fall within the limits therefore the process should be monitored using these control limits.

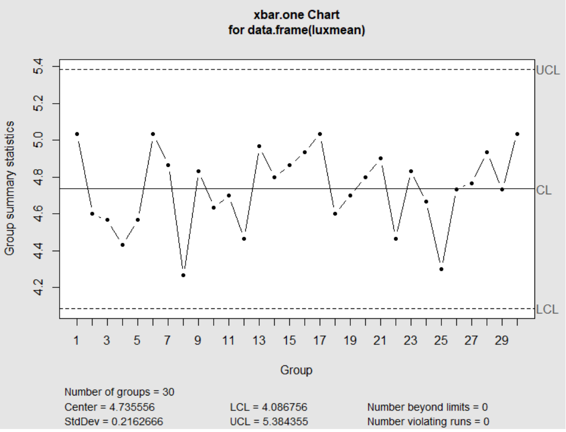


Figure 17: Luxury x Bar

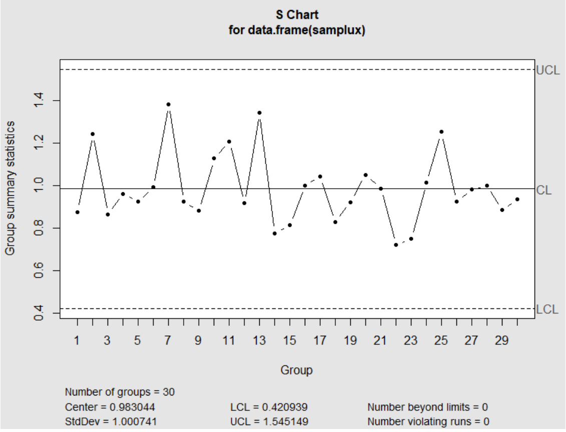


Figure 18: Luxuary s Bar

Luxury items x and s charts are plotted in figures 17 and 18. Both the x and s values fall within the limits and therefore there is no change to be made to the control limits.

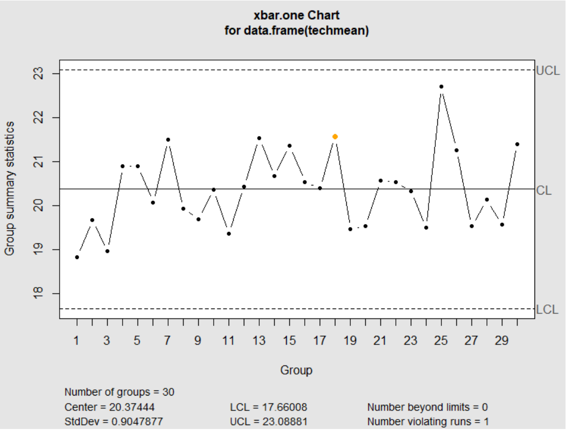


Figure 19: Technology x Bar

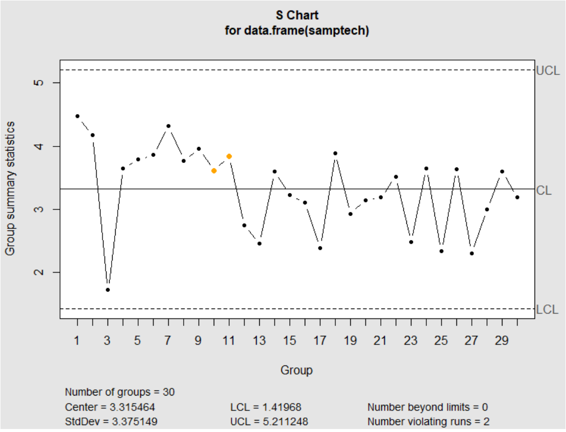


Figure 20: Technology s Bar

The class technology is plotted in figures 19 and 20. There is one point in the x bar plot that violates the rules and shows irregularities and is marked in yellow above this is because there are 7 points that fall on the same side of the CL and therefore the limits will need to recalculated without this sample (Stellenbosch University, 2022). The s bar graph has two values that have the same regularity and will need to be removed. Therefore there are two additional samples that need to be removed.

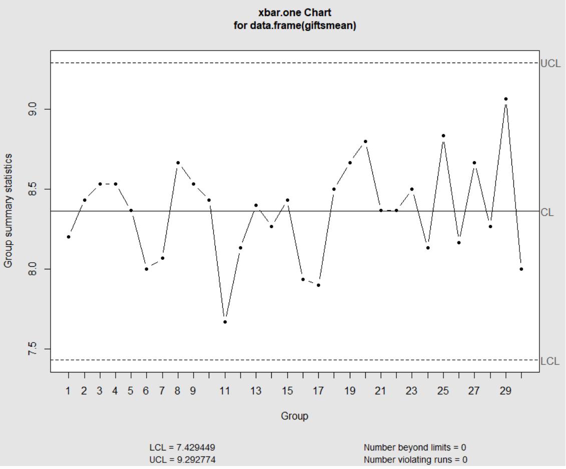


Figure 21: Gifts x Bar

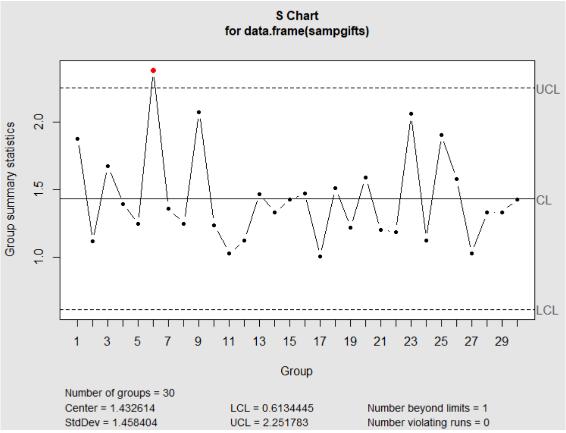


Figure 22: Gifts s Bar

The class gifts is shown in figures 21 and 22 from this we can see that the x values fall within the limits but there is one s value that falls outside these limits shown in red above. This sample will therefore have to be removed and the control limit values will be recalculated.

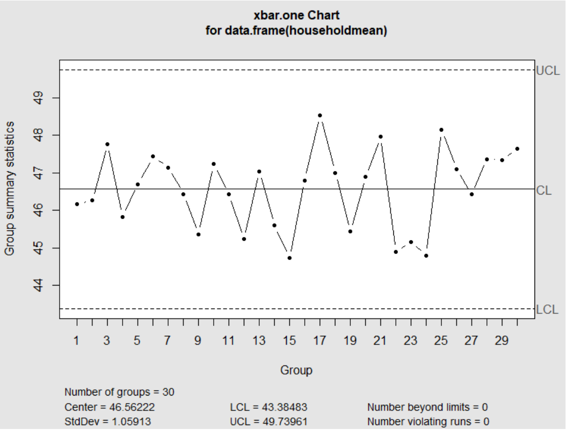


Figure 23: Household x Bar

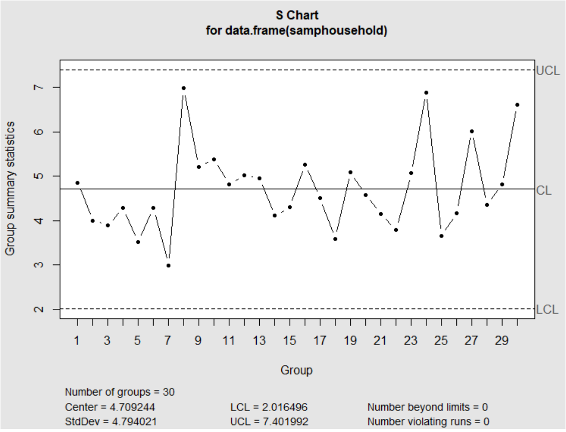


Figure 24: Household s Bar

The class household has x bar sample values and s bar sample value plotted above there are no irregularities therefore there is no need to adapt the control limits and the process is in control.

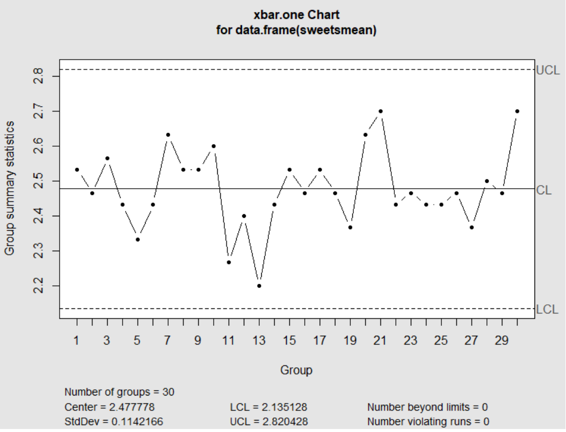


Figure 25: Sweats x Bar

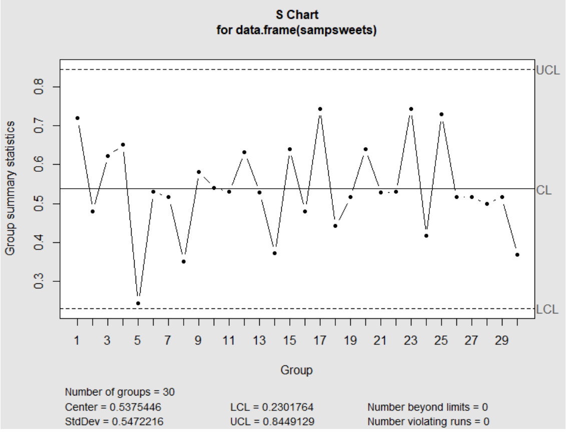


Figure 26: Sweats s Bar

The class sweets has no irregularities in the initialization chart and hence the values can be used for control charts.

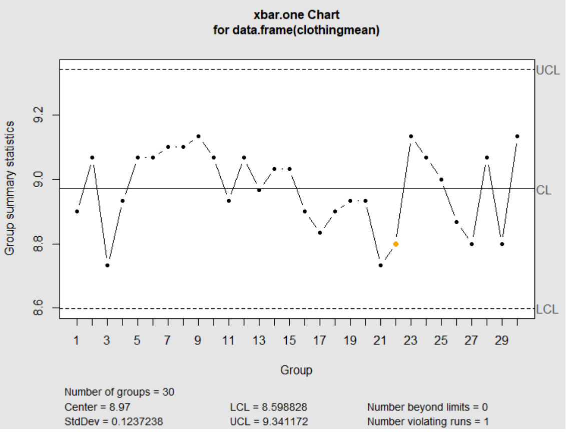


Figure 27: Clothing x Bar

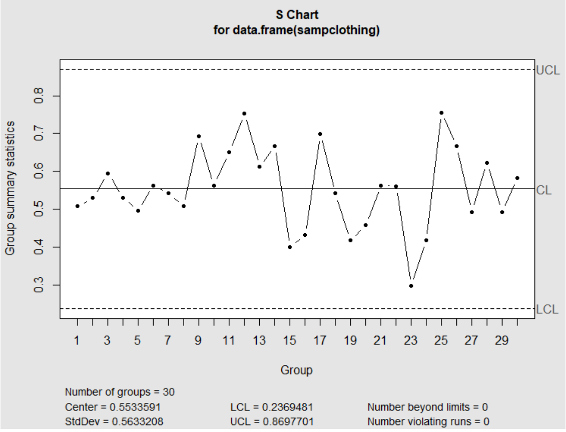
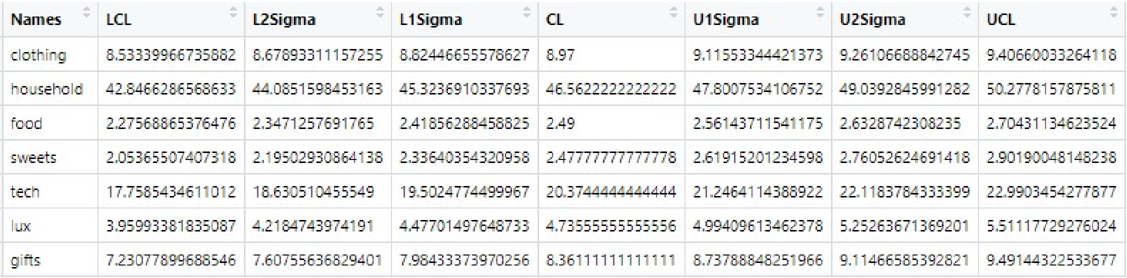


Figure 28: Clothing s Bar

All the samples for x bar and s bar fall within the limits but the x bar chart has one irregularity which is shown in yellow it violates the rule that no 7 points should consecutively be on the same side of the chart. Therefore the process should be analysed and the control limits adapted if needed.

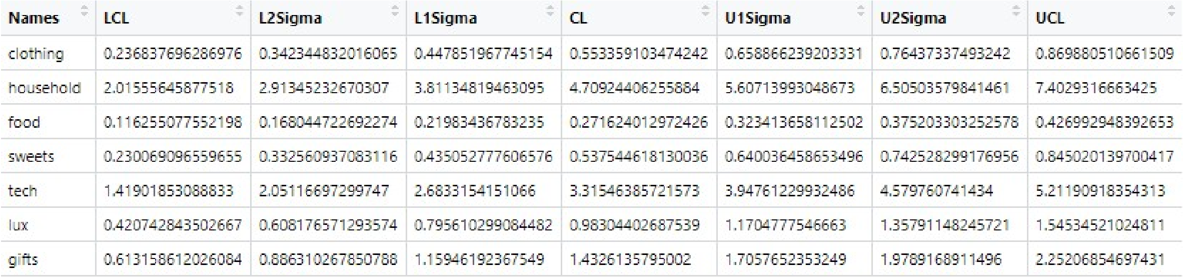
The following tables show the limits of all the x bar and s charts that were initialized above and will be the values shown in the control charts.

Table 5: X Control Limits



Names	LCL	L2Sigma	L1Sigma	CL	U1Sigma	U2Sigma	UCL
clothing	8.53339966735882	8.67893311157255	8.82446655578627	8.97	9.11553344421373	9.26106688842745	9.40660033264118
household	42.8466286568633	44.0851598453163	45.3236910337693	46.5622222222222	47.8007534106752	49.0392845991282	50.2778157875811
food	2.27568865376476	2.3471257691765	2.41856288458825	2.49	2.56143711541175	2.6328742308235	2.70431134623524
sweets	2.05365507407318	2.19502930864138	2.33640354320958	2.47777777777778	2.61915201234598	2.76052624691418	2.90190048148238
tech	17.7585434611012	18.630510455549	19.5024774499967	20.3744444444444	21.2464114388922	22.1183784333399	22.9903454277877
lux	3.95993381835087	4.2184743974191	4.47701497648733	4.73555555555556	4.99409613462378	5.25263671369201	5.51117729276024
gifts	7.23077899688546	7.60755636829401	7.98433373970256	8.36111111111111	8.73788848251966	9.11466585392821	9.49144322533677

Table 6: S Control Limits



Names	LCL	L2Sigma	L1Sigma	CL	U1Sigma	U2Sigma	UCL
clothing	0.236837696286976	0.342344832016065	0.447851967745154	0.553359103474242	0.658866239203331	0.76437337493242	0.869880510661509
household	2.01555645877518	2.91345232670307	3.81134819463095	4.70924406255884	5.60713993048673	6.50503579841461	7.4029316663425
food	0.116255077552198	0.168044722692274	0.21983436783235	0.271624012972426	0.323413658112502	0.375203303252578	0.426992948392653
sweets	0.230069096559655	0.332560937083116	0.435052777606576	0.537544618130036	0.640036458653496	0.742528299176956	0.845020139700417
tech	1.41901853088833	2.05116697299747	2.6833154151066	3.31546385721573	3.94761229932486	4.579760741434	5.21190918354313
lux	0.420742843502667	0.608176571293574	0.795610299084482	0.98304402687539	1.1704777546663	1.35791148245721	1.54534521024811
gifts	0.613158612026084	0.886310267850788	1.15946192367549	1.4326135795002	1.7057652353249	1.9789168911496	2.25206854697431

4.3 Control of process

After the control systems are initialized we can then sample the data as it flows into the process to check for any irregularities. This was done and obtained the following.

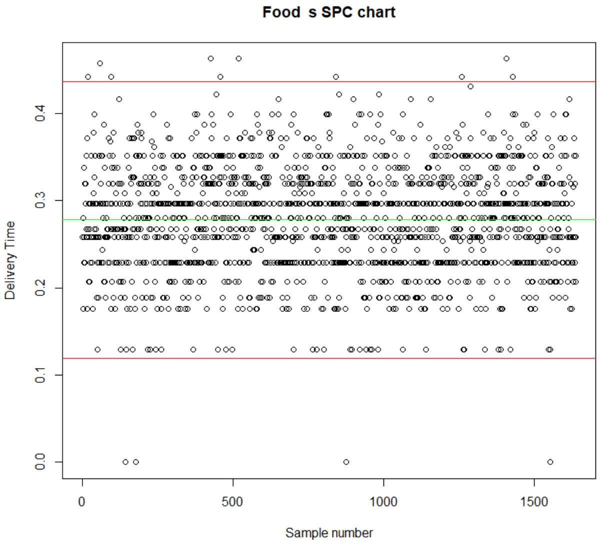


Figure 29: Food x Bar

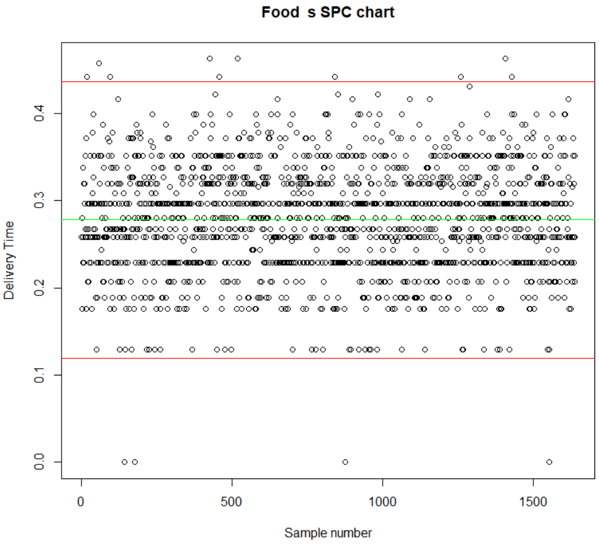


Figure 30: Food s Bar

Figure 29 above shows x bar control chart and figure 30 s control chart for the class food for all the instances. The process mostly falls within the control limits with a few outliers hence does not need to be adapted.

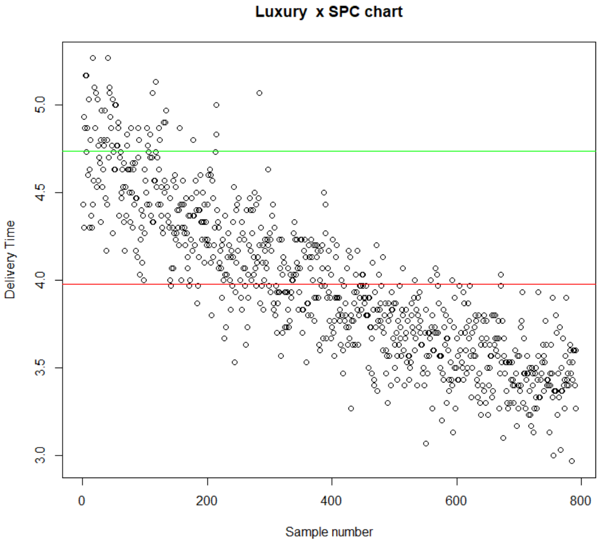


Figure 31: Lux x Bar

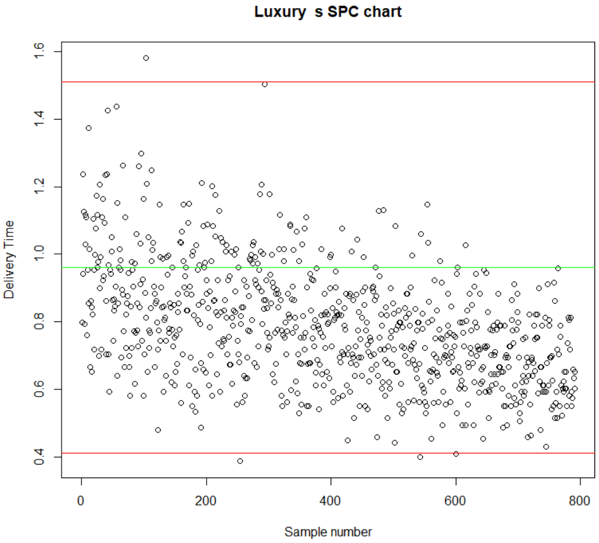


Figure 32: Lux s Bar

Both the x bar and s control charts for luxury items show that there is a downward trend. For the x bar chart the downward trend implies there is a lower delivery time and the lower s values show that there is less standard deviation. Therefore the process is becoming better. The process should be checked to see how improvement was made and if this improvement can be implemented in other categories.

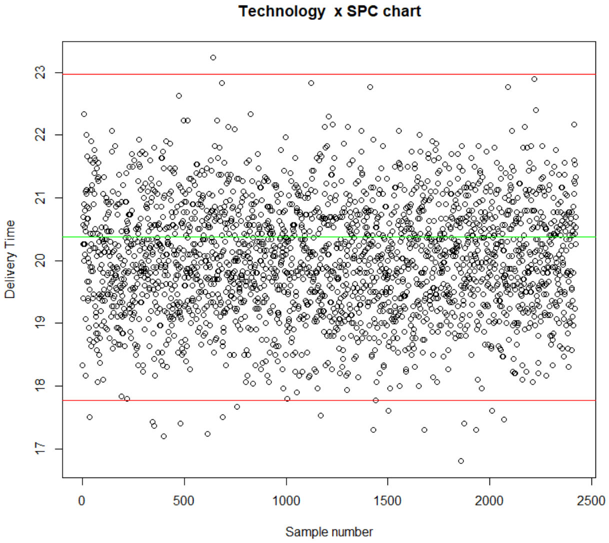


Figure 33: Tech x Bar

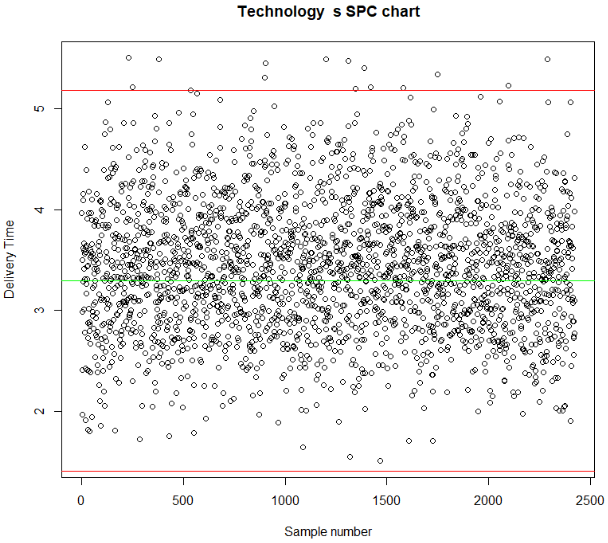


Figure 34: Tech s Bar

The figure above shows x bar and s control chart for the class technology the instances of the process mostly fall within the control limits with a few outliers hence does not need to be adapted.

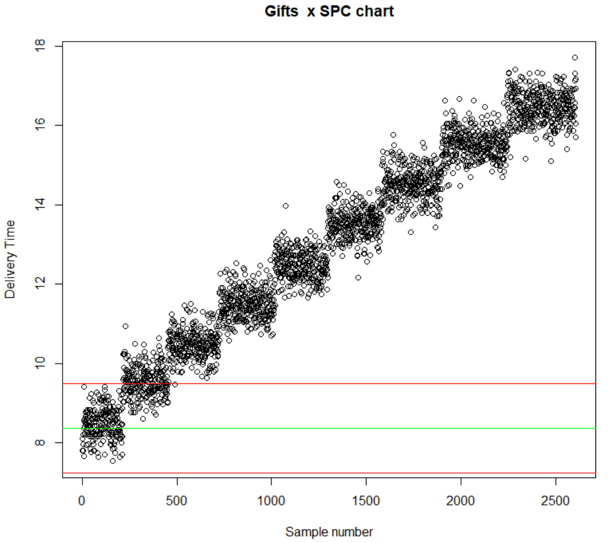


Figure 35: Gifts x Bar

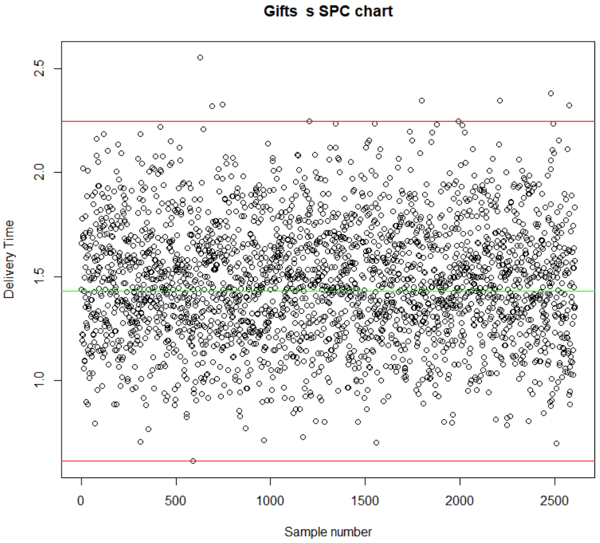


Figure 36: Gifts s Bar

The x bar graph for gifts has a very large upward trend showing that the process is out of control and should be analyzed to see why the delivery times are increasing so drastically. The s control chart is mostly stable as the values fall between the control limits.

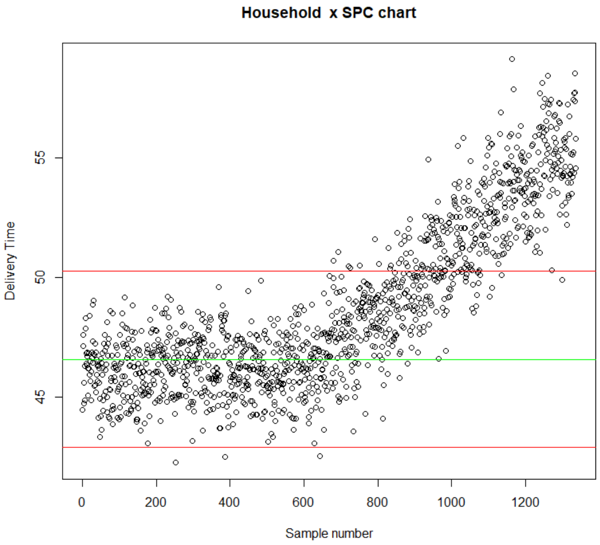


Figure 37: Household x Bar

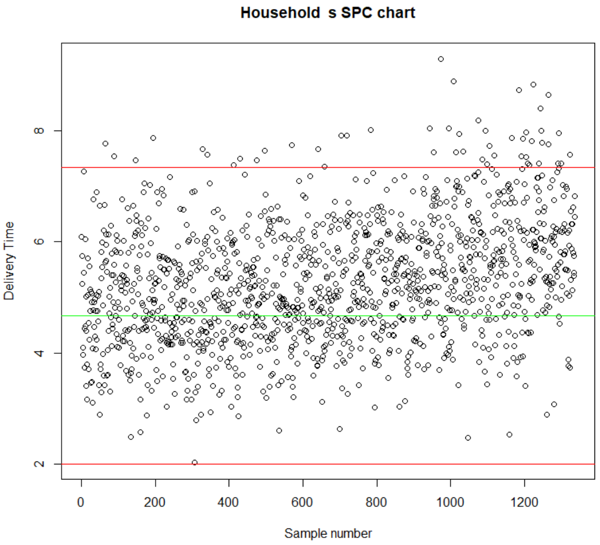


Figure 38: Household s Bar

The x and s graph show some irregularities in the data as after sample 600 in the x bar chart for household the x bar values start to increase rapidly showing that the process should be analyized and checked to bring it back within control limits. Most of the data is within the control limits for s household but there are a few too many above the control limits and the data occurs much more frequently above the center line than below it. Which adds to the need to analyse the process to be brought back down within control limits.

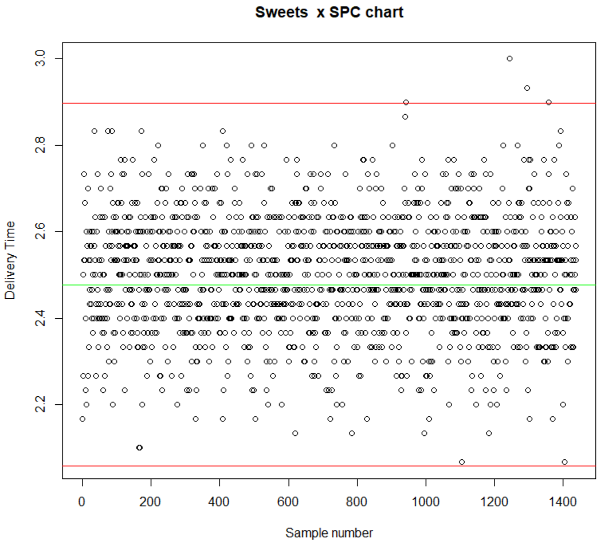


Figure 39: Sweats x Bar

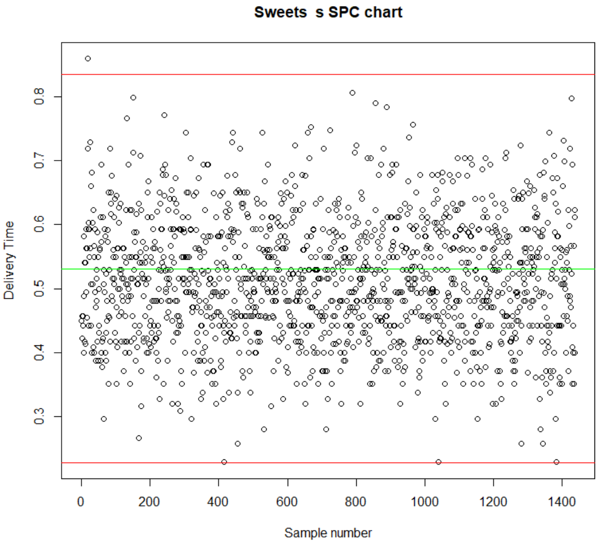


Figure 40: Sweats s Bar

The sweets x bar and s chart show that the process is in control the values fall within the control limits for almost all of the values.

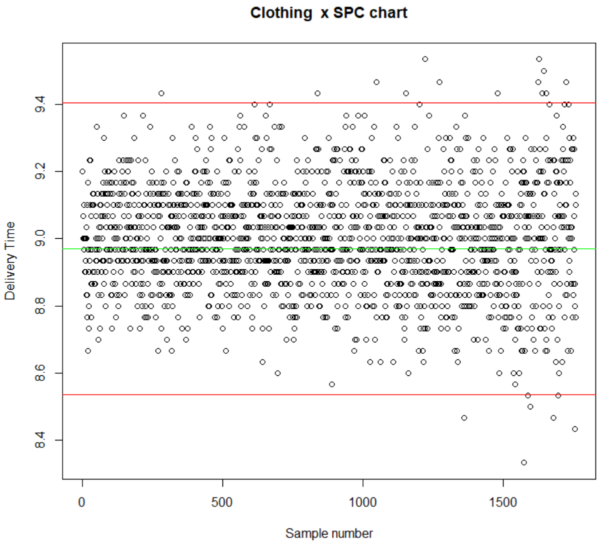


Figure 41: Clothing x Bar

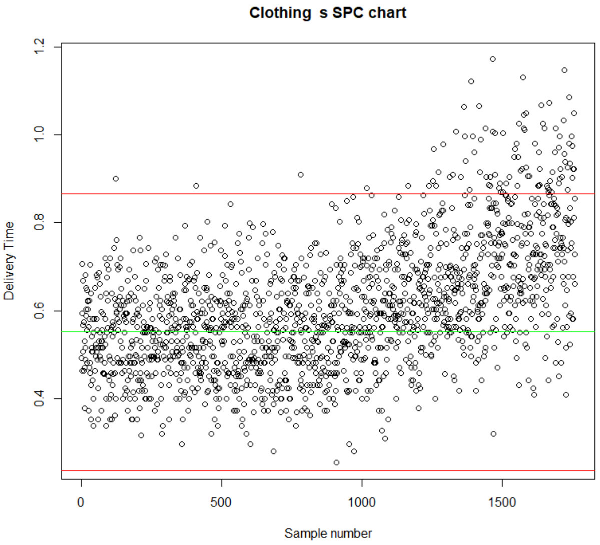


Figure 42: Clothing s Bar

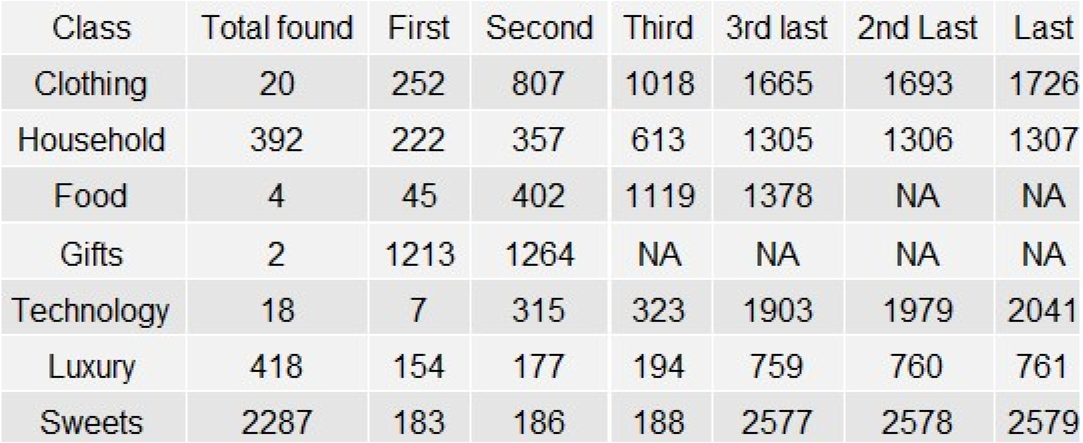
The clothing x bar and s charts show that the process should be analysed as the s chart shows a upward trend after sample 1000.

5 Optimizing delivery processes

For optimizing the delivery processes, we will be determining the number of x sample means that lie outside the control limits as well as the maximum number of samples that fall between the +0.4 and -0.3 sigma range of the center line of the s-chart. From the graphs in the control phase of control charts we can see the trends and where all the samples occur this part will be determining how many outliers and where exactly where they occur.

Table 7 below shows the number of sample mean that fall outside of the control limits and the index of those samples if there are many only the first 3 and last 3 indices are shown. With Gifts having the least outliers with only 2 and Sweets having the most with 2287 outliers

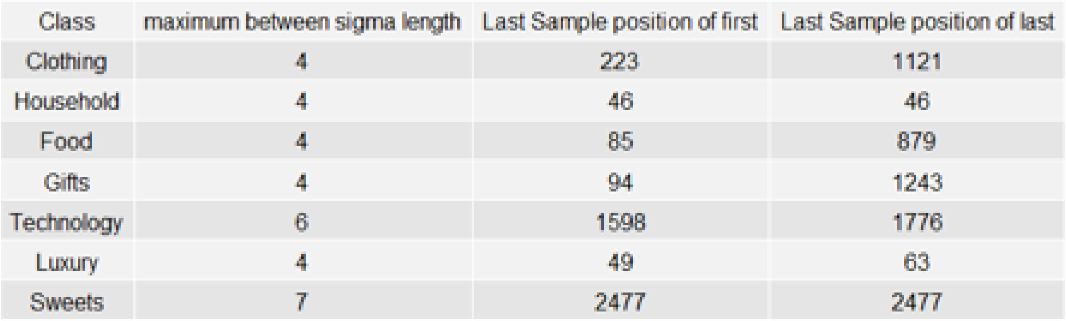
Table 7: Sample means outside of control limits



Class	Total found	First	Second	Third	3rd last	2nd Last	Last
Clothing	20	252	807	1018	1665	1693	1726
Household	392	222	357	613	1305	1306	1307
Food	4	45	402	1119	1378	NA	NA
Gifts	2	1213	1264	NA	NA	NA	NA
Technology	18	7	315	323	1903	1979	2041
Luxury	418	154	177	194	759	760	761
Sweets	2287	183	186	188	2577	2578	2579

The most consecutive samples of s-bar between -0.3 and +0.4 and the ending sample number are shown in 8 below. The class technology has the most number of consecutive s-bar samples between 0.3 and -0.4 with 7 and technology has 6 with all the other classes having a maximum of 4.

Table 8: Consecutive samples of s bar, last sample positions



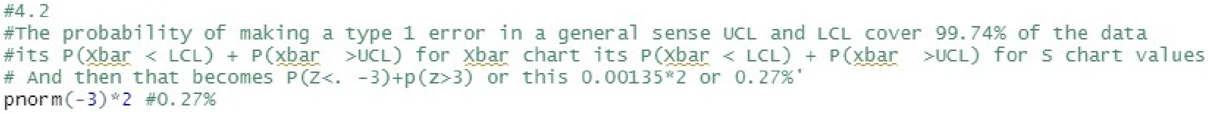
Class	maximum between sigma length	Last Sample position of first	Last Sample position of last
Clothing	4	223	1121
Household	4	46	46
Food	4	85	879
Gifts	4	94	1243
Technology	6	1598	1776
Luxury	4	49	63
Sweets	7	2477	2477

5.1 Errors

Keep in mind the patterns occur probabilistically, meaning they have a set or fixed probability (albeit small) to occur when there is nothing wrong with the process; then the signal was wrong about the process being unstable. We call this the Type I error. Type II errors are when the process is unstable, but our Sags, signals or rules give no indication from our graphs. Type I and Type II errors are two well-known concepts in quality engineering, which are related to hypothesis testing. Often engineers are confused by these two concepts simply because they have many different names. We list a few of them here (Stellenbosch University, 2022).

Type I errors are also called:

1. Producer's risk or manufacturer's risk (we investigate a process, but it was stable all along)
2. False alarm
3. Error



```
#4.2
#The probability of making a type 1 error in a general sense UCL and LCL cover 99.74% of the data
#its P(Xbar < LCL) + P(xbar  >UCL) for Xbar chart its P(Xbar < LCL) + P(xbar  >UCL) for S chart values
# And then that becomes P(Z<. -3)+p(z>3) or this 0.00135*2 or 0.27%'
pnorm(-3)*2 #0.27%
```

Figure 43: Type 1 Errors

Type II errors are also called:

1. Consumer's risk(we fail to investigate a process, but unknown to us the process is unstable)
2. Misdirection
3. Error

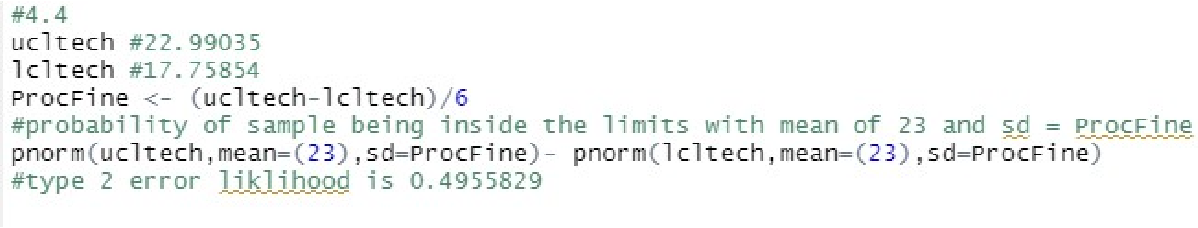
Type I and Type II errors can be defined in terms of hypothesis testing.

A type I error is the probability of rejecting a true null hypothesis, i.e: a type I error is the probability of telling you things are wrong, when they are actually correct (Stellenbosch University, 2022). For these type I errors the following hypothesis is used: H0 = mean of the process is within the upper and lower control limits

A Type II error is the probability of failing to reject a false null hypothesis, i.e: a type II error is the probability of telling you things are correct, when they are actually not (Stellenbosch University, 2022). For these type II errors the following hypothesis is used: H1 = mean of the process is not within the upper and lower control limits

Estimating type I data can be done as follows:

Estimating the likelihood of making a type II error for the class technology when the delivery process average moves to 23 hours. This is calculated using the upper control limit of 22.99035 and a lower control limit of 17.75854. Using these two values we can calculate our range of being within the sample limit. Therefore the type II error likelihood is 0.4955829.



```
#4.4
ucltech #22.99035
lcltech #17.75854
ProcFine <- (ucltech-lcltech)/6
#probability of sample being inside the limits with mean of 23 and sd = ProcFine
pnorm(ucltech,mean=(23),sd=ProcFine)- pnorm(lcltech,mean=(23),sd=ProcFine)
#type 2 error liklihood is 0.4955829
```

Figure 44: Type II Errors

5.2 Cost of late deliveries for technology

In order to maximize profit, the center line for delivery time should be determined in order to maximize the profit.

For every hour over 26 hours it costs R329 and costs R2.5 per item per hour to reduce the average time by one hour. Therefore to optimize profit the company should reduce their average time lost by 3 hours as shown in the graph below.

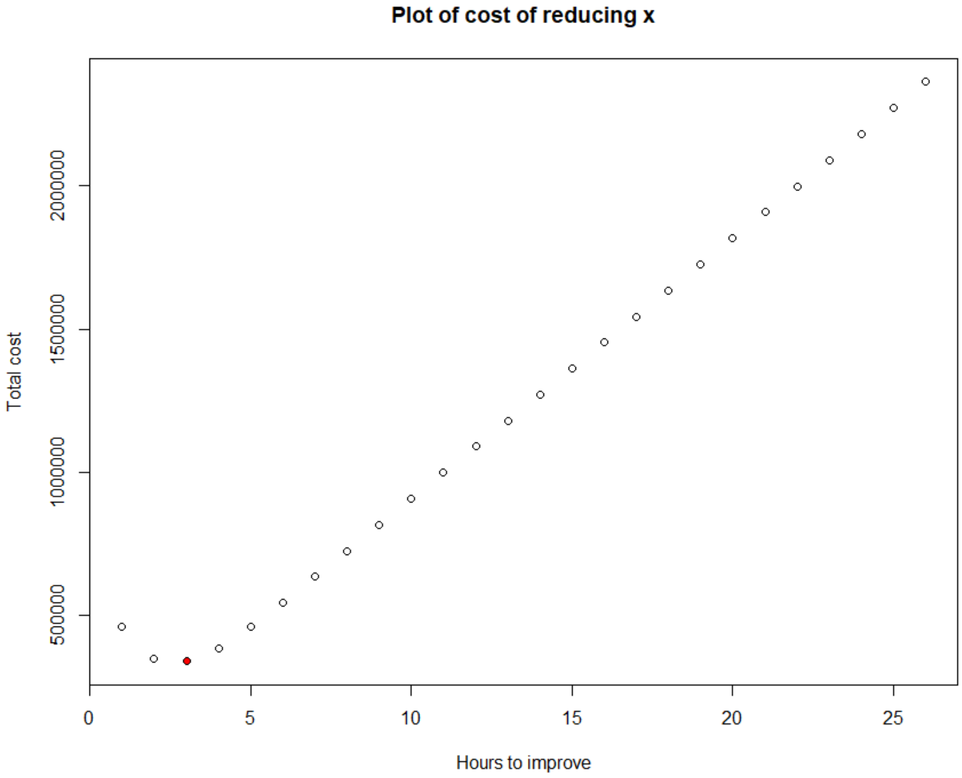


Figure 45: Plot of cost of reducing delivery time

6 Manova Results

MANOVA in R uses Pillai's Trace test for the calculations. An F-statistic is then used to check the significance of the group mean differences (Radečić, 2022). The Manova was carried out for 2 features; classes and why bought. These features were selected because they showed the most difference from the descriptive statistical analysis. The time of purchase showed little influence and was therefore not considered.

The null hypothesis for each is that the sample means of the different classes or why purchased are the same. The alternate hypothesis is that at least one of the group means are different from the rest.

6.1 Manova Class

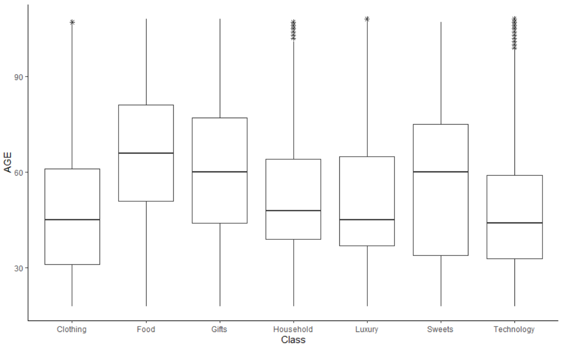


Figure 46: Manova Age Class Box Plot

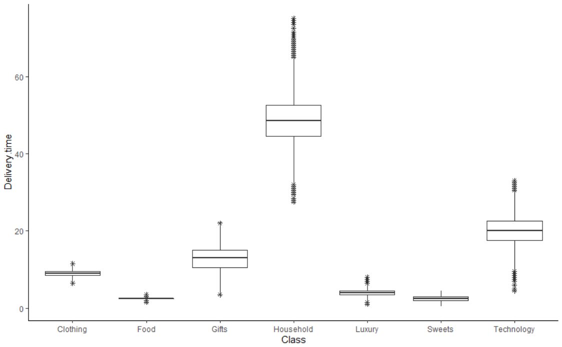


Figure 47: Manova Delivery Time Class Box Plot

The Class Manova showed the link between the age of buyers and the class of the items (see figure 46). This was also evident in the box-plots in the descriptive statistics. The link between technology and clothing and a younger age can be used to aim marketing of these products.

A link between delivery time and class is shown in figure 47 where household items have a significantly higher delivery time mean than the other classes. The longer delivery times for household items and technology could impact on the purchase of these items.

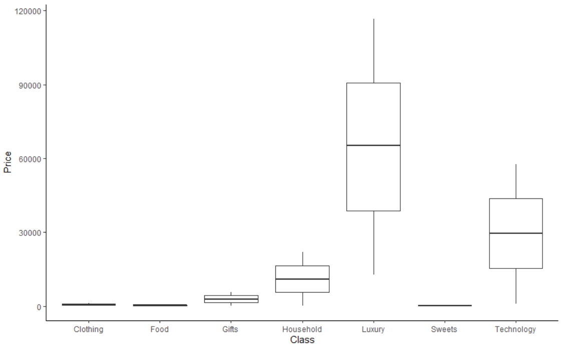
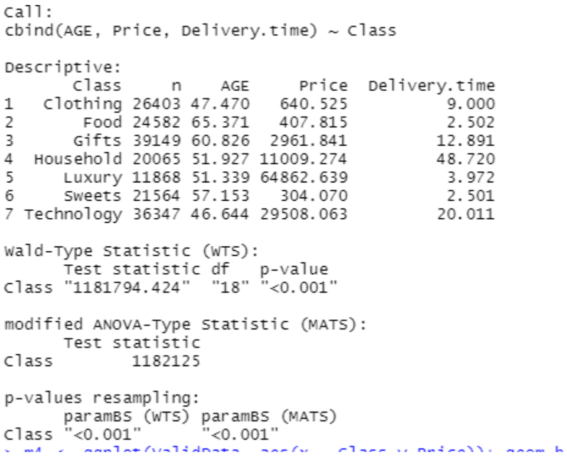


Figure 48: Manova Price Class Box Plot



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

Descriptive:
       Class     n    AGE      Price  Delivery.time
1   Clothing 26403 47.470    640.525          9.000
2       Food 24582 65.371    407.815          2.502
3      Gifts 39149 60.826   2961.841         12.891
4  Household 20065 51.927  11009.274         48.720
5     Luxury 11868 51.339  64862.639          3.972
6     Sweets 21564 57.153    304.070          2.501
7 Technology 36347 46.644  29508.063         20.011

Wald-Type Statistic (WTS):
      Test statistic df   p-value
Class "1181794.424"  "18" "<0.001"

modified ANOVA-Type Statistic (MATS):
      Test statistic
Class        1182125

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

Figure 49: Manova Summary

From the Class Manova a link was identified between class and price as shown in figure 48 with the luxury goods having far higher mean prices. Technology also showed higher mean prices.

From figure 49 it is shown that at least one of the sample means of class are different from the rest, and why bought is the same ie $p < 0.05$ (Radečić, 2022).

6.2 Manova Why Bought

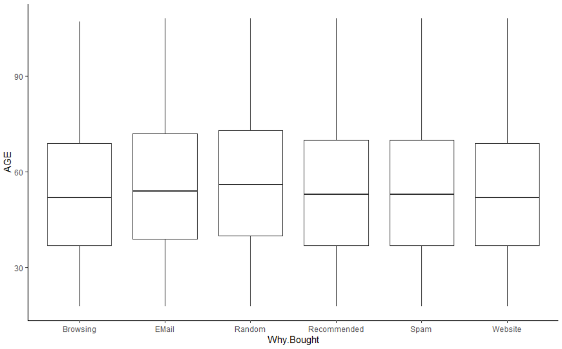


Figure 50: Manova Age Why Purchased Box Plot

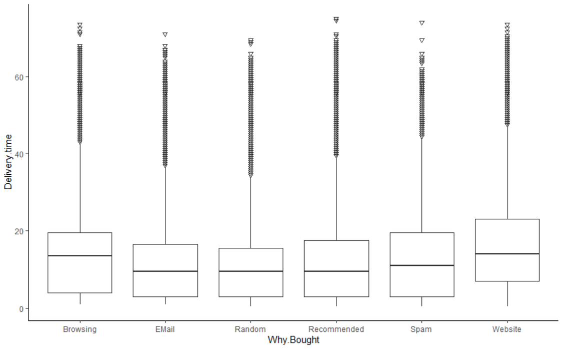


Figure 51: Manova Delivery Time Why Purchased Box Plot

The why bought manova shows the average age of each feature are fairly similar. The why bought manova in figure 51 shows that the delivery times are similar for all purchases except for website where the times for delivery could be reduced to ensure that sales are not impacted by consumers not wanting to wait for their goods.

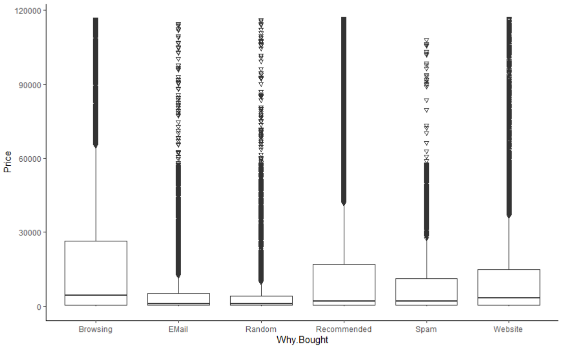
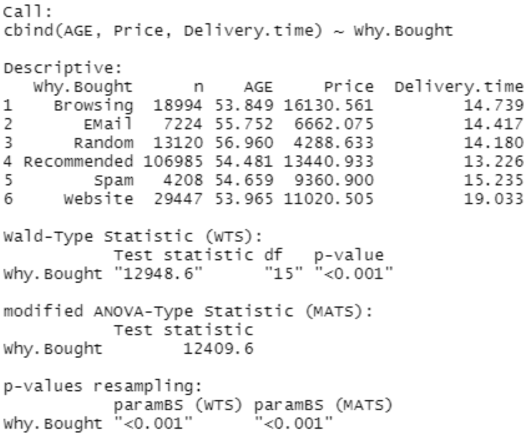


Figure 52: Manova Price Why Bought Box Plot



```
Call:
cbind(AGE, Price, Delivery.time) ~ Why.Bought

Descriptive:
   Why.Bought      n    AGE     Price  Delivery.time
1    Browsing  18994 53.849 16130.561         14.739
2       EMail   7224 55.752  6662.075         14.417
3      Random  13120 56.960  4288.633         14.180
4 Recommended 106985 54.481 13440.933         13.226
5        Spam   4208 54.659  9360.900         15.235
6     Website  29447 53.965 11020.505         19.033

Wald-Type Statistic (WTS):
           Test statistic df   p-value
Why.Bought "12948.6"      "15" "<0.001"

modified ANOVA-Type Statistic (MATS):
           Test statistic
Why.Bought        12409.6

p-values resampling:
           paramBS (WTS) paramBS (MATS)
Why.Bought "<0.001"      "<0.001"
```

Figure 53: Manova Why Bought Summary

The average price of of browsing purchases is the highest showing that customers still want to physically see goods especially when they are expensive.

7 Reliability of the service and product

7.1 Problem 6

Taguchi measured the quality as the variation from the design specification value and placed a monetary value on the loss. The larger the deviation from the specifications, the larger the loss. The figure 54 below shows the shape of a Taguchi loss function with the quadratic form showing the increasing costs (loss) as the variation is further from the target.

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

$\mathrm{L(x) = k(x - T)}$

$\$45 = \mathrm{k}(0.04)^2$

$\mathrm{k} = 28125$

$\mathrm{L(x)} = 218125(\mathrm{x} - \mathrm{T})^2$

Figure 54 shows the costs follow a quadratic form with increased cost of loss as you move further from the specification of 0.06 cm. At 0.06 cm there is no loss and the cost is zero. The lines are ±0.04cm indicating the acceptable variation and the green line at $45 show the scrap.

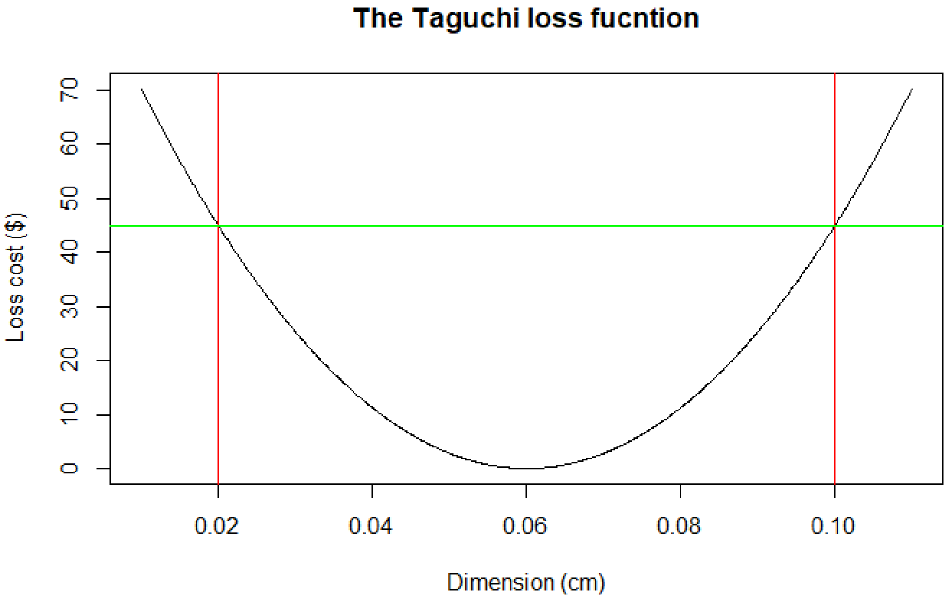


Figure 54: Taguchi Loss Function

7.2 Problem 7

A team was formed to study the refrigerator part at Cool Food, Inc. described in Problem 6. While continuing to work to find the root cause of scrap, they found a way to reduce the scrap costs to \$35 per part. Determine the Taguchi loss function for this situation.

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

$\mathrm{L(x) = k(x - T)}$

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

$\mathrm{k} = 21875$

$\mathrm{L(x)} = 21875(\mathrm{x - T})^2$

$\mathrm{L}(0.027) = 21875(0.027)^2$

$= \$15.95$

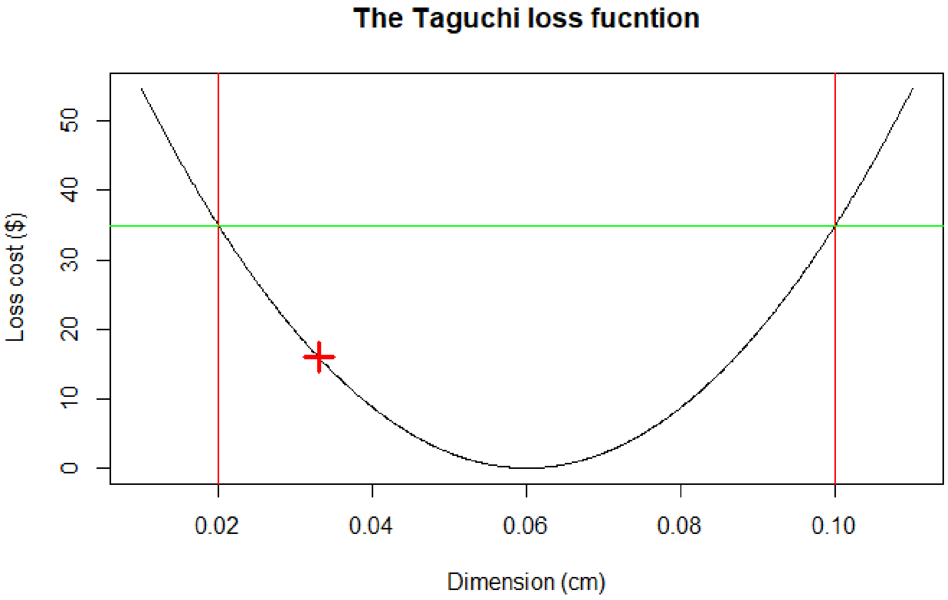
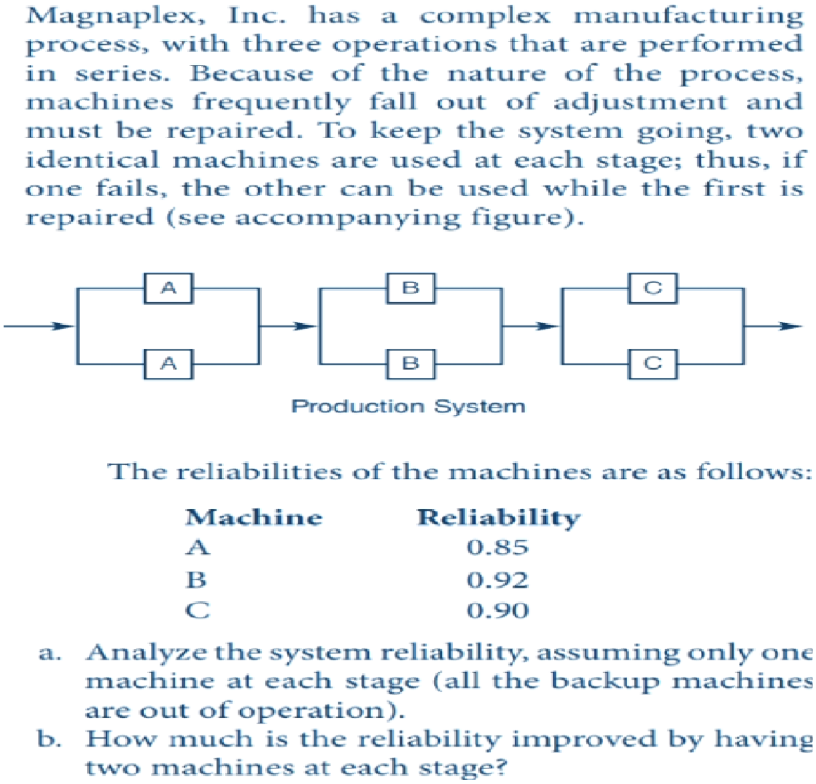


Figure 55: Taguchi Loss Function

7.3 Problem 27



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 (see accompanying figure).

Production System

The reliabilities of the machines are as follows:

Machine	Reliability
A	0.85
B	0.92
C	0.90

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

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

a) The reliability of the system when in series:

$\mathrm{R_aR_bR_c} = (0.85)(0.92)(0.90)$

$\mathrm{R_aR_bR_c} = 0.7038$

b) The reliability of the system with machines in parallel

$\mathrm{R_{aa}R_{bb}R_{cc}} = [1-(1-\mathrm{R_a})^2] * [1-(1-\mathrm{R_b})^2] * [1-(1-\mathrm{R_c})^2]$

$\mathrm{R_{aa}R_{bb}R_{cc}} = [1-(1-0.85)^2] * [1-(1-0.92)^2] * [1-(1-0.9)^2]$

$\mathrm{R_{aa}R_{bb}R_{cc}} = [1-0.0225] * [1-0.0064] * [1-0.01]$

$\mathrm{R_{aa}R_{bb}R_{cc}} = 0.9615$

When the system is in parallel there is a much higher reliability.

7.4 Part 6.3 : Binomial Problem

For the delivery process, there are 21 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 22, how many days per year we should expect reliable delivery times?

The binomial distribution describes the probability of obtaining exactly x 'successes' in a sequence of n trails. In this example a success is the number of vehicle failures. The probability of a success is a constant calue called P/ Therefore we define P as the probability of x number of vehicle failures or drivers unavailable.

The binomial distribution formula is

$\mathrm{P(x)} = \binom{\mathrm{n}}{\mathrm{x}}\mathrm{p^x q^{(n-x)}} = \frac{\mathrm{n!}}{\mathrm{(n-x)!x!}}\mathrm{p^x q^{(n-x)}}$

where

n = the number of trials (or the number being sampled)

x = the number of successes desired

p = probability of getting a success in one trial

q = 1 − p = probability of getting a failure in 1 trail

The drivers and vehicles are dealt with separately and then the probabilities are combined to find the overall reliability deliveries.

Vehicle calculations

In order for a successful delivery must have 19 vehicles therefore we can only have at the most 2 failure. Therefore, we consider the probabilities for 0, 1 and 2 failures:

P(0 failures) = 0.8616898

P(1 failures) = 0.1287268

P(2 failures) = 0.009157301

Vehicle reliability is therefore = P(0 failures) + P(1 failure) + p(2 failures) = 0.9995739

Total days reliable = 364.8445

Driver calculations

We consider probabilities of 0, 1 and 2 unavailable, meaning there are 21, 20 or 19 drivers available

Driver reliability probability

P(0 unavailable) =0.9343324

P(1 unavailable) = 0.06356541

P(2 unavailable) = 0.002059307

Total driver reliability = P(0 unavailable) + P(1 unavailable) + P(2 unavailable) =0.9999571

Days reliable per year for drivers = * 365 = 364.994

Overall reliability : Drivers and Vehicles

Overall reliability = 0.999531

Overall Days reliable per year = 0.999531 * 365 = 364.8288

Changing to 22 vehicles

In order for a successful delivery must have 19 vehicles therefore we can only have at the most 3 failure. Therefore, we consider the probabilities for 0, 1, 2 and 3 failures:

P(0 failures) = 0.8556033

P(1 failures) = 0.1339041

P(2 failures) = 0.01000188

P(3 failures) = 0.0004743392

Vehicle reliability is therefore = P(0 failures) + P(1 failure) + P(2 failures) + P(3 failures) = 0.9999836

Total days reliable of vehicles = 364.994

Therefore as drivers has not changed we can use the same values getting a total overall reliability of 0.9999407.

With a total days of 364.9783.

8 Conclusions

Analysis of the data-set showed the relationships between different features. Understanding these relationships and differences can assist management in their decision making as the consequences of changes on other parts of the business can be anticipated and managed.

From the data-set provided it was evident that most people bought products from recommendations, and that gifts and technology were the most frequently purchased items. The time, day and month had little impact on the purchases but a large decrease in clothing and household goods were seen after 2012. The majority of customers are between the ages of 40 to 60.

Improvements in delivery time should be considered for web purchases and opportunities should be provided for customers to browse products as the highest value items are purchased when browsing.

The upward trends in the delivery time for gifts and luxury goods should be investigated and deliveries optimised to save costs and realize higher sales.

References

Das, S., Datta, S., and Chaudhuri, B.B., 2018. Handling data irregularities in classification: foundations, trends, and future challenges. *Pattern recognition*, 81, pp.674–693.

FormPlus, 2022. *Categorical data: definitions* [Online]. Available from: `https://www.formpl.us/blog/categorical-data`.

Radečić, D., 2022. *Manova in r – how to implement and interpret one-way manova* [Online]. Available from: `https://www.r-bloggers.com/2022/01/manova-in-r-how-to-implement-and-interpret-one-way-manova/`.

Stellenbosch University, 2022. *Quality assurance 344 : statistical methods in quality assurance part 1 summary* [Online]. Stellenbosch: Stellenbosch University. Available from: `https://learn.sun.ac.za/pluginfile.php/3514418/mod_resource/content/1/QA344%20Statistics.pdf`.

Zstatistics, 2019. *How to learn statistics in half an hour* [Online]. Available from: `https://www.youtube.com/watch?v=kyjlxsLW1Is`.