ECSA Graduate Attributes Project

Stephan Heyns

23725044

Stellenbosch University

2022

Table of Contents

Introduction

The sales data of a business is analysed by using graphs, correlations and statistics to identify trends and possibly predict trends. These results are used to optimize their sales and other aspects such as delivery times. Proposals are made to improve the business model by referring to past data of customers and their trends.

Before solutions can be found the data needs to be neat and the report will clearly explain the structure of the data. RStudio is used as a tool to calculate probabilities, construct graphs, calculate correlations, and to do iterations with the use of for-loops.

This report will make suggestions on possible changes that will assist the company to maximise sales and customer satisfaction, through delivering not only a quality product but also a quality service.

Part 1: Data wrangling

Before data can be analysed it has to be pre-processed. Pre-processing of data can include sorting, cleaning, adding or removing features, and removing missing or invalid instances. In this section the data will be imported, assessed and split into a valid and invalid dataset.

The Data

The original dataset consists of 180 000 instances and 10 features with 8 of the features being numerical and 2 categorical. The structure of the dataset can be seen in figure 1:

```
'data.frame':   180000 obs. of  10 variables:
 $ X             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ ID            : int  19966 34006 62566 70731 92178 50586 73419 32624 51401 96430 ...
 $ AGE           : int  54 36 41 48 76 78 35 58 82 24 ...
 $ Class         : chr  "Sweets" "Household" "Gifts" "Technology" ...
 $ Price         : num  246 1708 4051 41843 19215 ...
 $ Year          : int  2021 2026 2027 2029 2027 2027 2029 2025 2025 2027 ...
 $ Month         : int  7 4 8 10 11 4 11 7 12 11 ...
 $ Day           : int  3 1 10 22 26 24 13 2 18 4 ...
 $ Delivery.time: num  1.5 58.5 15.5 27 61.5 14.5 4 2 12 3 ...
 $ Why.Bought    : chr  "Recommended" "Website" "Recommended" "Recommended" ...
```

Figure 1

Invalid Data

The dataset contains missing and negative values which should be removed before using the data. One of the ways to remove missing and negative values is to remove the instances or columns that contain the invalid values, removing columns could result in removing important descriptive features in the dataset. Splitting the data into invalid and valid datasets will ensure a dataset without missing values.17 Instances were found with missing values and 5 negative values. Figure 2 displays all the invalid instances.

	ID_Invalid	ID_Invalid	ID_Invalid	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
12345	1	1	1	12345	18973	93	Gifts	*NA*	2026	6	11	15.5	Website
16320	2	2	2	16320	44142	82	Household	-588.8	2023	10	2	48.0	EMail
16321	3	3	3	16321	81959	43	Technology	*NA*	2029	9	6	22.0	Recommended
19540	4	4	4	19540	65689	96	Sweets	-588.8	2028	4	7	3.0	Random
19541	5	5	5	19541	71169	42	Technology	*NA*	2025	1	19	20.5	Recommended
19998	6	6	6	19998	68743	45	Household	-588.8	2024	7	16	45.5	Recommended
19999	7	7	7	19999	67228	89	Gifts	*NA*	2026	2	4	15.0	Recommended
23456	8	8	8	23456	88622	71	Food	*NA*	2027	4	18	2.5	Random
34567	9	9	9	34567	18748	48	Clothing	*NA*	2021	4	9	8.0	Recommended
45678	10	10	10	45678	89095	65	Sweets	*NA*	2029	11	6	2.0	Recommended
54321	11	11	11	54321	62209	34	Clothing	*NA*	2021	3	24	9.5	Recommended
56789	12	12	12	56789	63849	51	Gifts	*NA*	2024	5	3	10.5	Website
65432	13	13	13	65432	51904	31	Gifts	*NA*	2027	7	24	14.5	Recommended
76543	14	14	14	76543	79732	71	Food	*NA*	2028	9	24	2.5	Recommended
87654	15	15	15	87654	40983	33	Food	*NA*	2024	8	27	2.0	Recommended
98765	16	16	16	98765	64288	25	Clothing	*NA*	2021	1	24	8.5	Browsing
144443	17	17	17	144443	37737	81	Food	-588.8	2022	12	10	2.5	Recommended
144444	18	18	18	144444	70761	70	Food	*NA*	2027	9	28	2.5	Recommended
155554	19	19	19	155554	36599	29	Luxury	-588.8	2026	4	14	3.5	Recommended
155555	20	20	20	155555	33583	56	Gifts	*NA*	2022	12	9	10.0	Recommended
166666	21	21	21	166666	60188	37	Technology	*NA*	2024	10	9	21.5	Website
177777	22	22	22	177777	68698	30	Food	*NA*	2023	8	14	2.5	Recommended

Figure 2

Valid Data

199978 Instances contains no missing or negative values. This dataset (figure 3) will be used for analysis in the remaining part of the report.

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

Figure 3

Part 2: Descriptive statistics

In this section graphs such as bar plots, histograms and scatter plots will be used to get a better understanding of the data. Descriptive statistics can be used as a tool to get a comprehensive understanding of not only the data, but also the business and its clients. Trends and correlation between features will be identified to make accurate prediction later in the report.

Feature overview

A brief description of each feature is given in table 1.

Feature	Description
ID_Valid	New index of the valid instances
X	The old primary key
ID	Unique values assigned to each customer
AGE	Age of customer
CLASS	Type of product bought
PRICE	Price paid for product
YEAR	Year of purchase
MONTH	Month of purchase
DAY	Day of purchase
DELIVERY.TIME	Days to make delivery
WHY.BOUGHT	Reason why customer bought the product

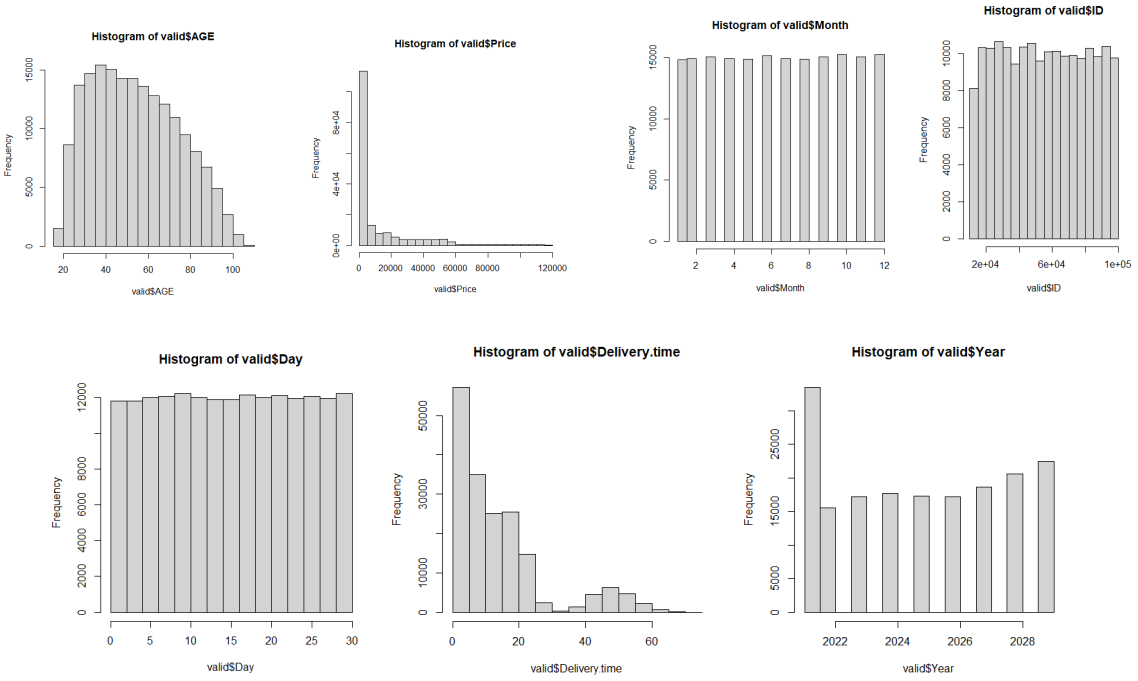
Table 1

Continuous features

There are 9 continuous features: ID_Valid, X, ID, AGE, Price, Year, Month, Day, and delivery time. Graphs will be constructed to visually represent the distributions of each feature. The features ID_Valid and X will not be analyzed since they will contribute no data, they are used as primary keys and no trend can be observed since they all are unique (uniform distribution).

A summary of the useful continous features can be seen below.

AGE	Price	Year	Month	Day	Delivery.time
Min. : 18.00	Min. : 35.65	Min. :2021	Min. : 1.000	Min. : 1.00	Min. : 0.5
1st Qu.: 38.00	1st Qu.: 482.31	1st Qu.:2022	1st Qu.: 4.000	1st Qu.: 8.00	1st Qu.: 3.0
Median : 53.00	Median : 2259.63	Median :2025	Median : 7.000	Median :16.00	Median :10.0
Mean : 54.57	Mean : 12294.10	Mean :2025	Mean : 6.521	Mean :15.54	Mean :14.5
3rd Qu.: 70.00	3rd Qu.: 15270.97	3rd Qu.:2027	3rd Qu.:10.000	3rd Qu.:23.00	3rd Qu.:18.5
Max. :108.00	Max. :116618.97	Max. :2029	Max. :12.000	Max. :30.00	Max. :75.0



Age

A skewed to the right distribution means that younger customers buy more often than older customers. Age also has outliers.

Price

Skewed to the right distribution of price means that cheaper priced products are more often bought than more expensive products. Some outliers are present.

Month

An uniform distibution of feature month means that the compny has sales throughout the year constantly and that there is no major peak season.

ID

An uniform distribution of the ID feature shows that most customers keep on buying several times, this can be a good way to analyze customer satisfaction.

Day

An uniform distibution of feature day means that the compny has sales throughout the month constantly.

Delivery time

This feature shows a multimodal distribution with two peaks.

Year

Most sales took place at the opening (2021) and then dropped and is now increasing at a almost constant rate every year.

Categorical features

In this dataset there is two categorical features class, which refers to the category which the product falls in and why bought, which refers to the reason the customer bought the product. Both these features will be analysed.

Class

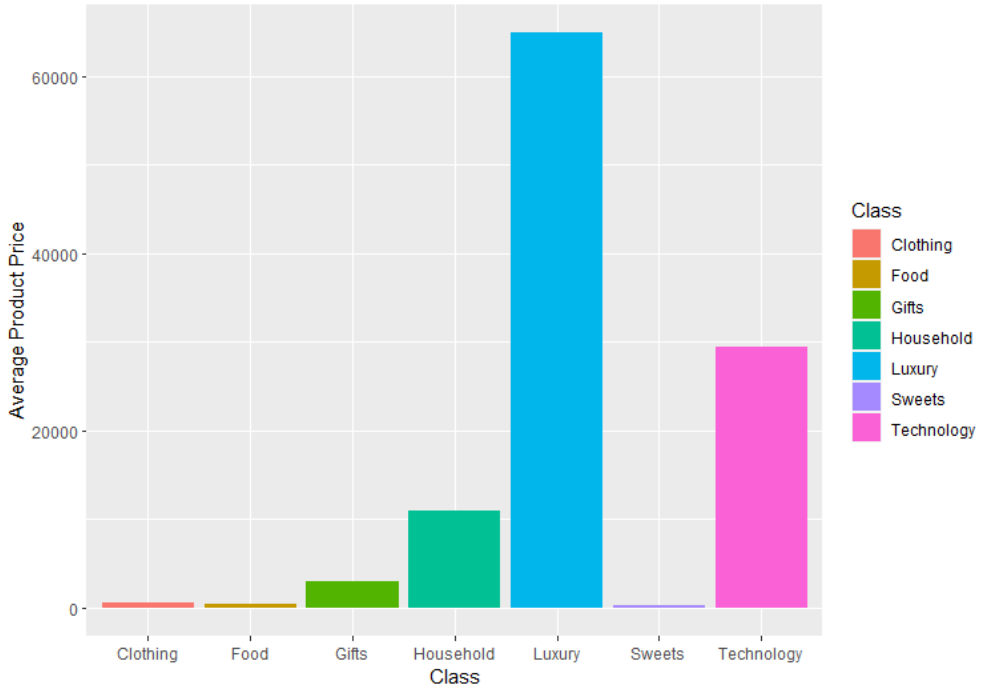


Figure 4

The average price of each class was compared with each other, this can be used to estimate which class has the heighest weighted contribution to income. This graph shows that luxury and secondly technology has the highest average price.

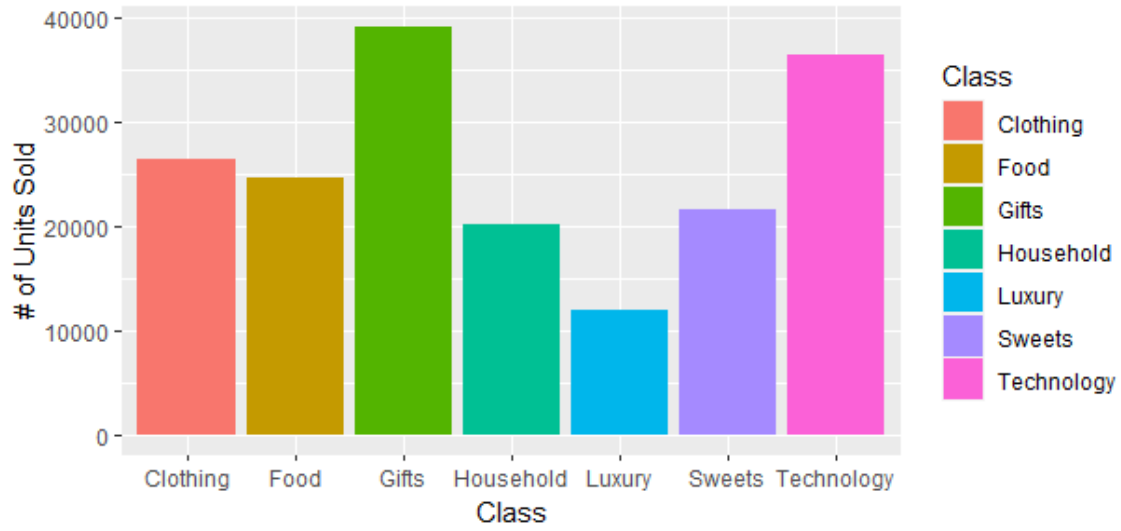


Figure 5

A graph displaying how many units per class was sold concludes that technology and gifts have the biggest demand by customers.

Why Bought

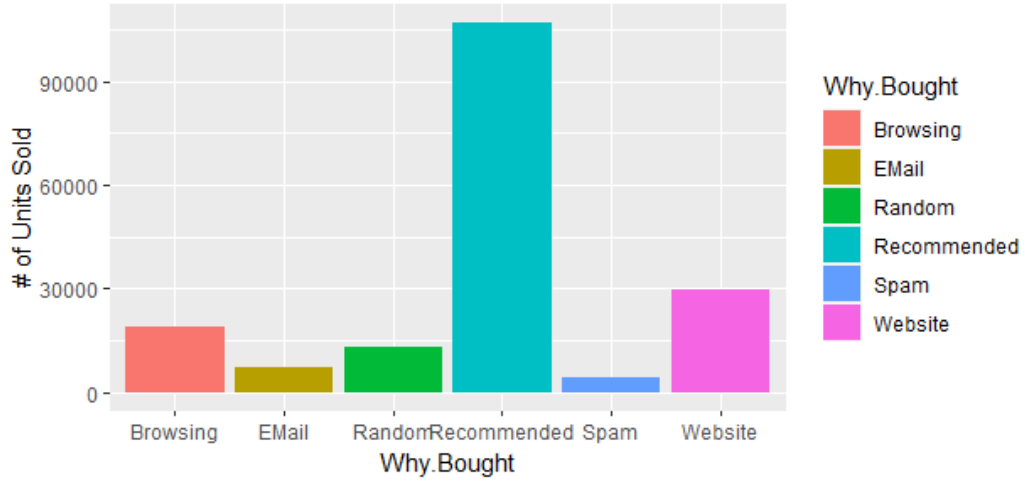


Figure 6

Using the histogram above the marketing team can establish in which sector to invest. Word of mouth (Recommended) has the highest number of units sold and then the website.

Feature correlation

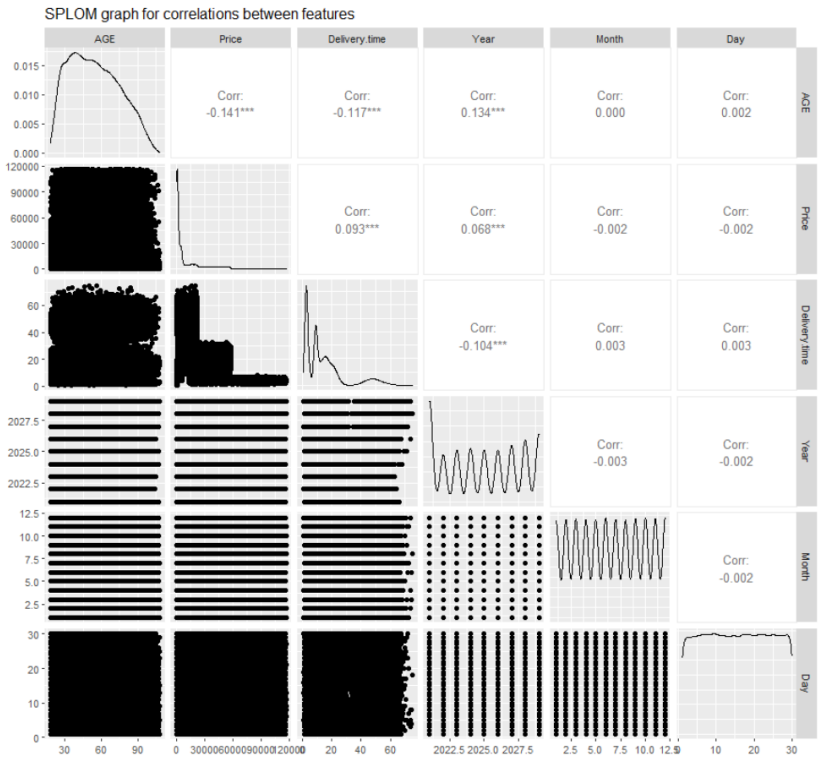


Figure 7

The SPLOM graph can be used to visualize correlations between numerical features. A correlation close to +/- is a considered a strong corelation (positive and negative respectively). The features have no clear strong correlations, but price and age have the strongest negative correlation, this

could be because older customers have more money to spend on more expensive products than younger customers

Process capability indices

By focusing on the delivery times for the class “Technology”, process capability indices can be calculated assuming that the Upper specification limit is 24 and the lower specification limit is 0. An LSL of 0 makes sense because if the delivery time is lower than 0 it means the product is delivered before it was ordered.

The mean of the data was calculated to be 20.01095 days and the standard deviation to be 3.50199 days. Using the mean and standard deviation the following was calculated:

CP	*CPU*	*CPL*	*CPK*
1.142207	0.3796933	1.90472	0.37969

Table 2

The Cp value indicates whether a process can produce a product within the specified limits. The minimum value for Cp is 1, but generally a larger value is preferred. This process is barely capable with a value of 1.14. (PQsystems, n.d.)

Cpk shows how well cantered the overall average is. A low Cpk value (<1) means that the process is not capable, in this process the Cpk is far below 1 and therefor performs poorly and is not capable. By shifting the process this the Cpk can be improved which will lead to a better product performance. (1factory, n.d.)

Part 3: Statistical Process Control (SPC)

Initialize Charts

The data of the valid dataset was ordered chronologically, Year, month, day, and then according to original indexes (X). It was then split into subsets according to classes. 30 samples of 15 instances each was used to calculate the values for the X-Charts and S-charts.

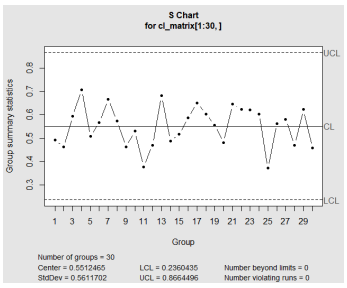
S-Charts

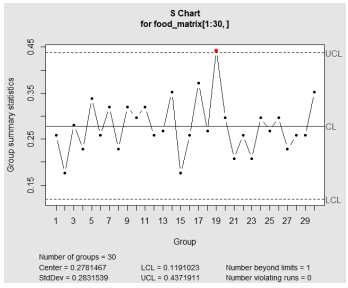
The centre line, outer control limits and the one and two Sigma control limits were calculated for the S-Charts and tabled.

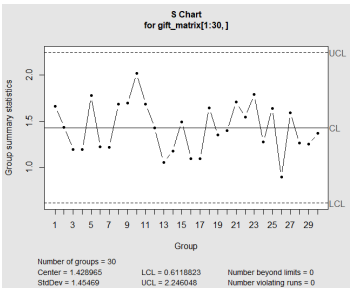
	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	0.86645	1.6736	1.1124	0.5512	-0.0099	-0.5711	0.2360
Food	0.43719	0.8445	0.5613	0.2781	-0.005	-0.2882	0.1191
Gifts	2.24605	4.3383	2.8837	1.4290	-0.0257	-1.4804	0.6119
Household	7.34325	14.1838	9.4278	4.6719	-0.0841	-4.8401	2.0005
Luxury	1.51086	2.9183	1.9398	0.9612	-0.0173	-0.9958	0.4116
Sweets	0.83523	1.6133	1.0723	0.5314	-0.0096	-0.5505	0.2275
Technology	5.17991	10.0052	6.6504	3.2955	-0.0593	-3.4142	1.4111

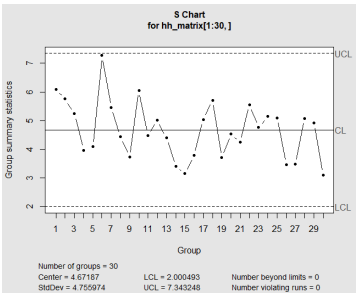
Table 3

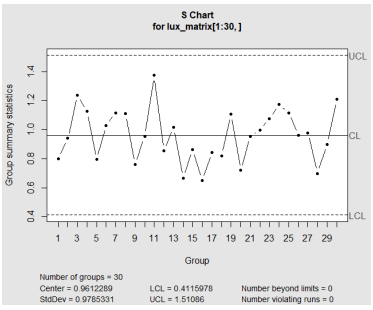
Using these values, s-charts are constructed to evaluate whether the sample averages fall within the limits.

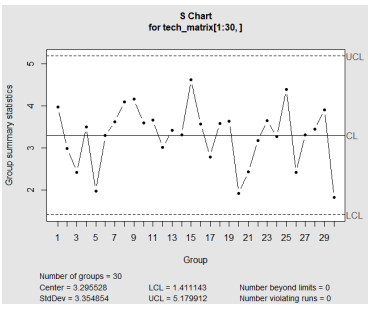


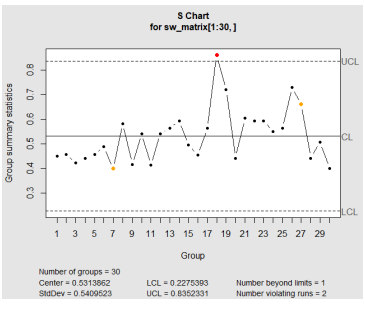












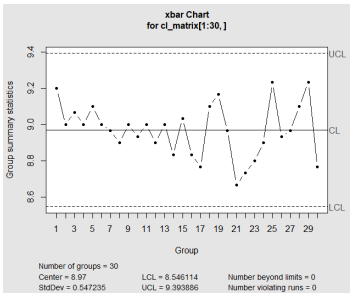
X-Charts

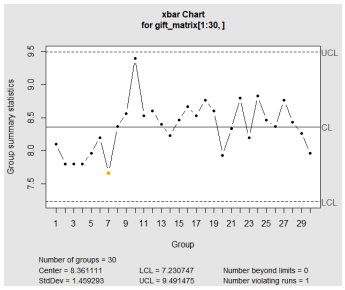
The centre line, outer control limits and the one and two Sigma control limits were calculated for the S-Charts and tabled.

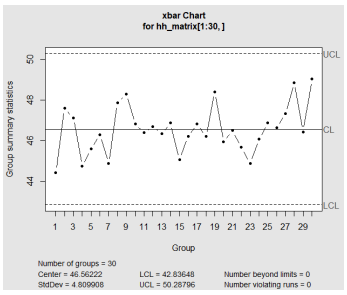
	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	9.394	10.064	9.517	8.97	8.423	7.876	8.546
Food	2.698	3.028	2.759	2.49	2.221	1.952	2.282
Gifts	9.492	11.280	9.820	8.361	6.902	5.443	7.231
Household	50.29	56.18	51.37	46.56	41.75	36.94	42.84
Luxury	5.47	6.64	5.69	4.74	3.79	2.83	3.999
Sweets	2.89	3.53	3.01	2.48	1.95	1.42	2.07
Technology	23.00	27.15	23.76	20.37	16.99	13.60	17.75

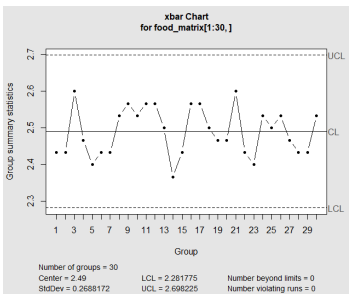
Table 4

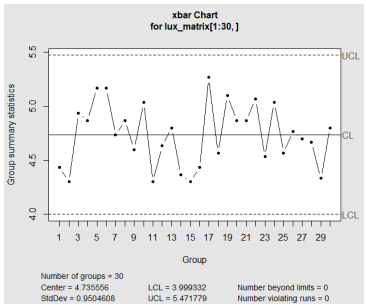
Using these values, X-charts are constructed to evaluate whether the sample averages fall within the limits.

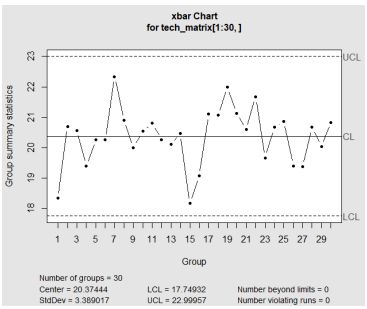


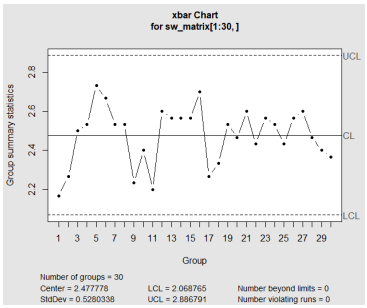












Remaining Samples

The rest of the samples in the dataset is plotted on S-Charts and X-Charts to visualize which samples fall outside of the calculated control limits.

S-Charts

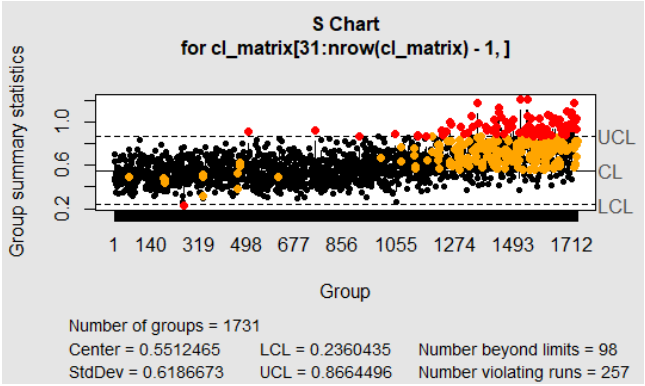


Figure 8

There is an upward trend in samples that fall outside the control limits, therefor there could be future problems with the process. The process starts out in control and then moves out of control.

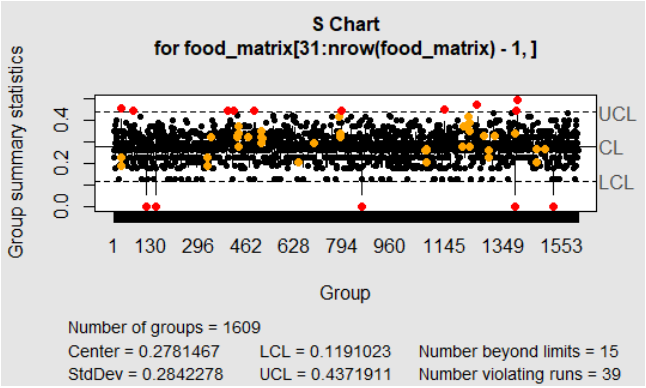


Figure 9

There are little samples out of control and therefor the process can be classified as under control.

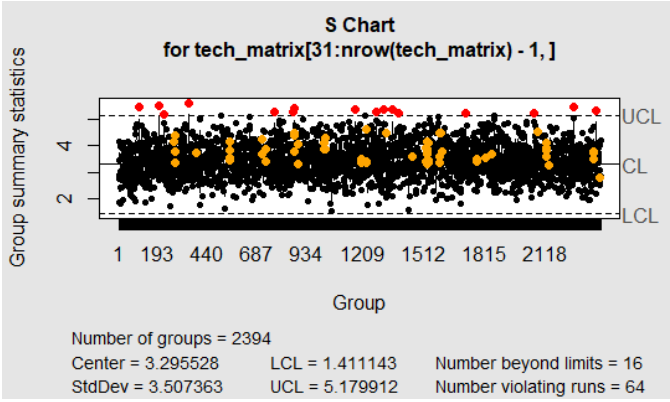


Figure 10

There are little samples out of control and therefor the process can be classified as under control.

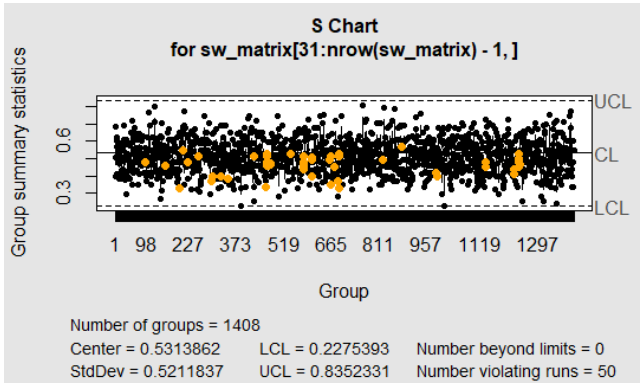


Figure 11

There are no samples outside the limits and a few consecutive runs, the process is under control.

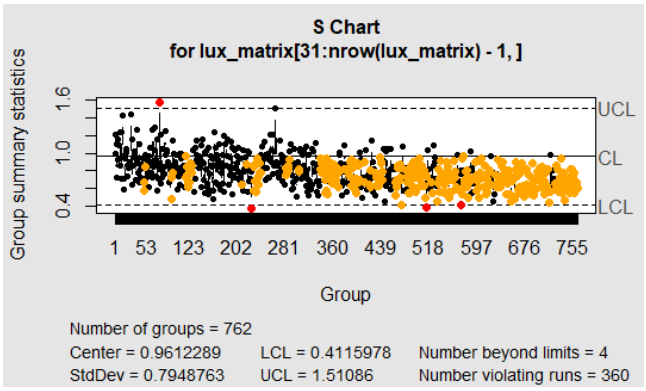


Figure 12

There are little samples out of control and therefor the process can be classified as under control.

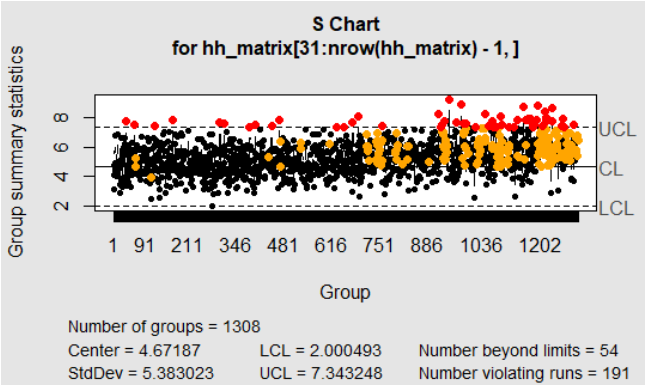


Figure 13

There are a few samples outside of the control limits and a lot of samples violating runs, therefor the process is out of control.

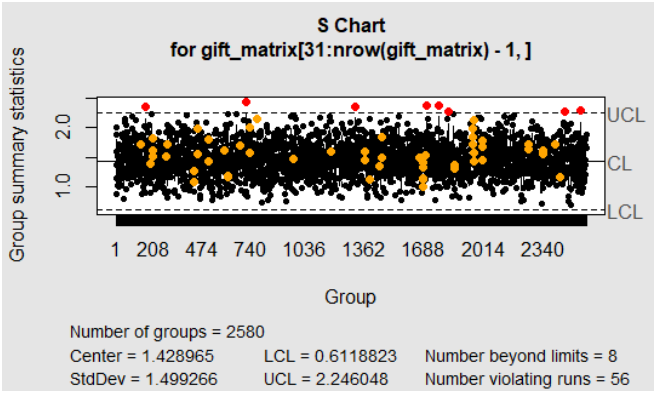


Figure 14

There are little samples out of control and therefor the process can be classified as under control.

X-Charts

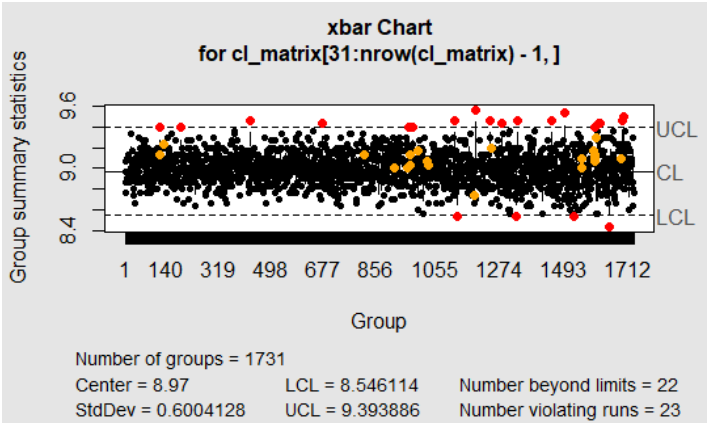


Figure 16

Not a lot of out-of-control samples, therefore the process is under control.

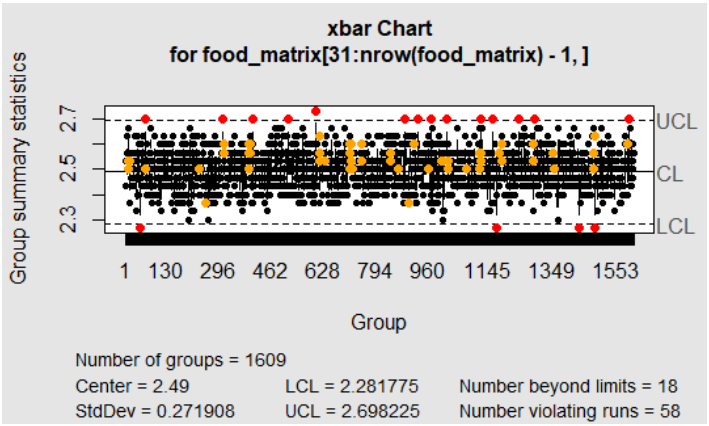


Figure 15

A lot of samples outside of control. The process is outside of control.

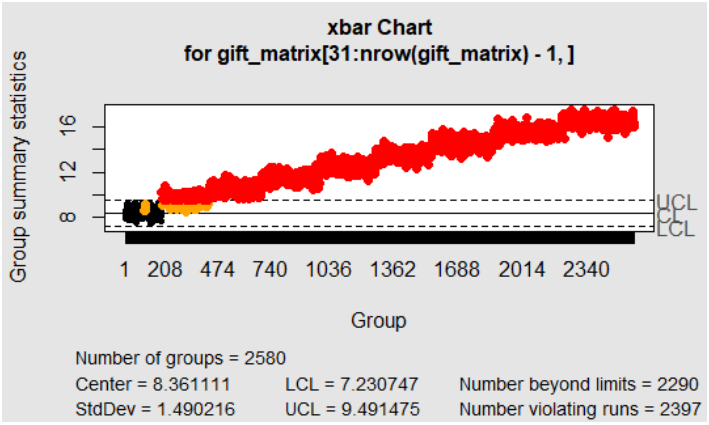


Figure 19

A lot of samples outside of control and an increasing trend of out-of-control samples. The process is outside of control.

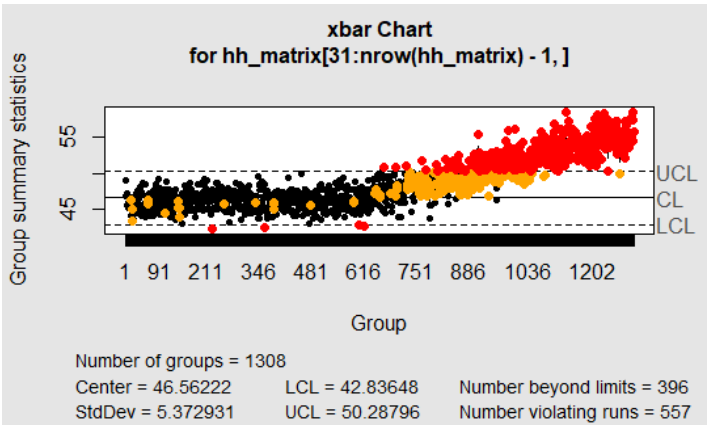


Figure 18

A lot of samples outside of control and an increasing trend of out-of-control samples. The process is outside of control.

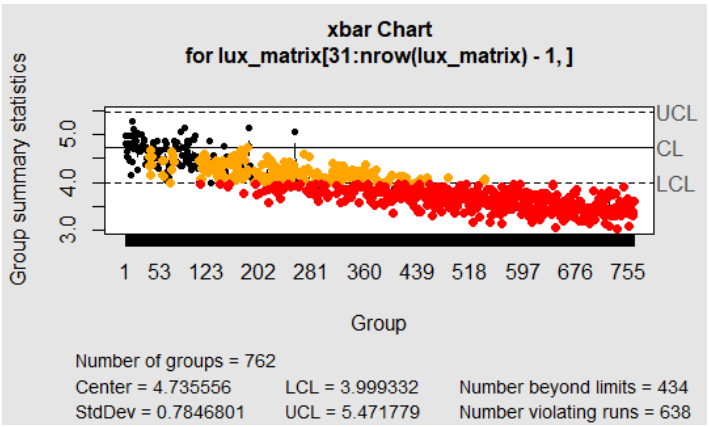


Figure 17

A lot of samples outside of control and an increasing trend of out-of-control samples. The process is outside of control.

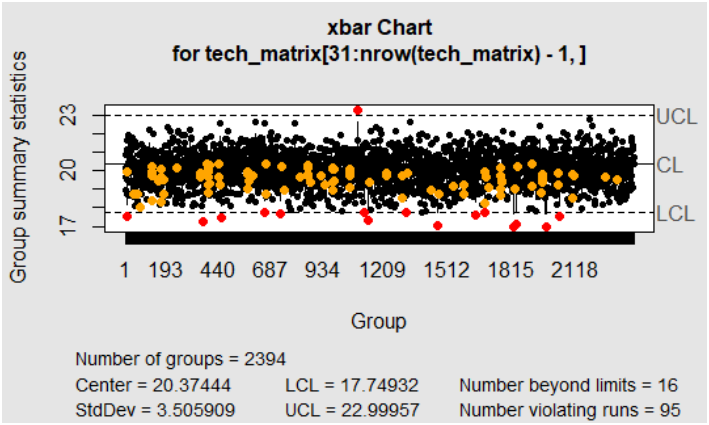


Figure 20

Not a lot of out-of-control samples, therefore the process is under control.

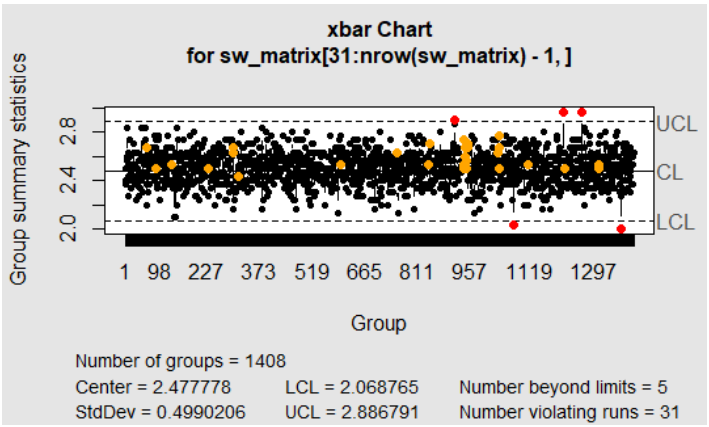


Figure 21

Not a lot of out-of-control samples, therefore the process is under control.

Part 4: Optimising the delivery processes

Sample means outside the outer control limits

The first 3, last 3 as well as the number of instances outside of the control limits of each class is summarised in the following table.

Class	1st	2nd	3rd	3rd last	2nd last	Last	Number
Clothing	119	188	426	1330	1528	1648	22
Food	64	309	403	1174	1438	1486	18
Gifts	184	187	189	2578	2579	2580	2290
Household	664	696	723	358	600	614	396
Luxury	113	142	155	760	761	762	434
Sweets	913	1214	1265	-	1075	1374	5
Technology	1093	88	369	1843	1980	2042	16

Table 5

Gifts and luxury have a lot of sample means outside of the lower and upper limits when looking at delivery times. These processes are therefore not under control and the causes should be investigated.

Most consecutive samples (-0.3 and 0.4 Sigma control limits)

Class	Most consecutive samples	Last sample
Clothing	24	1698
Food	16	1070
Gift	17	2495
Household	27	973
Luxury	86	746
Sweets	24	1355
Technology	16	336

Table 6

The class luxury has the most consecutive samples between 0.4 and -0.3 Sigma with the last sample of the sequence having a mean of 746.

Likelihood of making a type I error

A type I error is when the null hypothesis is falsely rejected. This is also called a false positive. In this case the null hypothesis is that is in control, therefor if a type I error is made it is calculated that the process is out of control, but it is in control. The likelihood of making a type I error is the statistical percentage of falsely rejecting a null hypothesis. (Banerjee, 2009)

For A the control limit is 3 Sigma. It is double sided therefor the pnorm function is used on 3 and then multiplied to account for both sides.

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

$$P(Type\ I\ error) = 0.002699796$$

$$P(Type\ I\ error) = 0.270\ \%$$

For B the control limits are 0.4 Sigma and -0.3 Sigma. The area of the pnorm graph between these two points are calculated as follow

$$P(Type\ I\ error) = pnorm(0.4) - pnorm(-0.3)$$

$$P(Type\ I\ error) = 27.33\%$$

Optimising the delivery process

When an order arrives late (more than 26 hours after the purchase is made) it will cost the company R329 per hour late. By reducing the average delivery time per item by an hour it will cost the company R2.5. The company currently loses R758 674 due to late deliveries in the Technology class.

A graph is constructed showing the effect of reducing the delivery time on the total cost. The global minimum of the graph is used as the optimal solution.

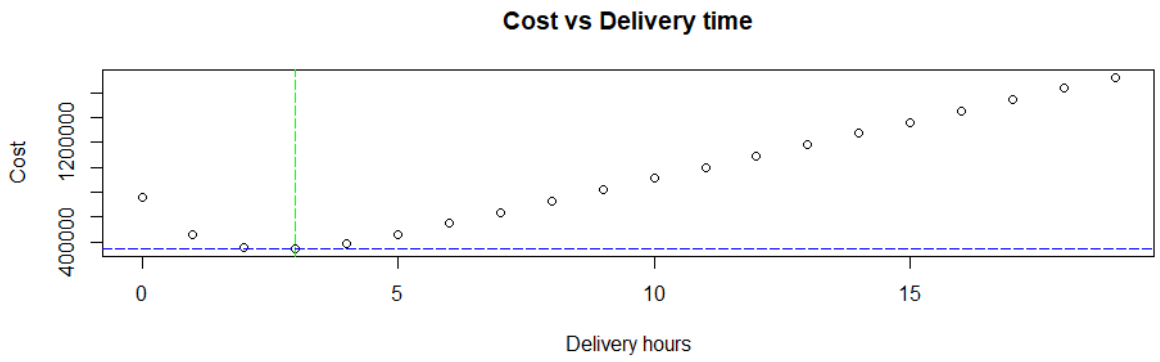


Figure 22

The optimal solution is to reduce all delivery times by 3 hours, this will result in a total cost of R340 870, therefor the company saves R417 804 by making this decision.

Likelihood of making a type II error

By making a type II error you fail to reject the null hypothesis, but the null hypothesis had to be rejected. The probability of making a type II error can be calculated.

$$\mathrm{P}(\text{Type II error} = \mathrm{pnorm}(\mathrm{UCL}, \mathrm{mean}, \text{standard deviation}) - \mathrm{pnorm}(\mathrm{LCL}, \mathrm{mean}, \text{standard deviation})$$

$$\mathrm{P}(\text{Type II error}) = \mathrm{pnorm}(23, 23, 0.875) - \mathrm{pnorm}(17.75, 23, 0.875)$$

$$= 0.5$$

If the delivery time average moves to 23 hours, the probability for falsely not rejecting the null hypothesis for the technology class, is 50%.

Part 5: DOE and MANOVA

It was decided that the categorical features would be used in the MANOVA, “Class” and “why bought”. Price and delivery time were used in the analysis. A hypothesis test is done for both the categorical features using the MANOVA table and mean plots are used to support the result.

Class

Ho: The class of the item that is sold does not affect the price and delivery time.

H1: The class of the item that is sold does affect the price and delivery time.

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

 Response Price :
                Df     Sum Sq    Mean Sq F value    Pr(>F)
Class            6 5.7168e+13 9.5281e+12   80258 < 2.2e-16 ***
Residuals   179971 2.1366e+13 1.1872e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 23

The p value as seen in in the MANOVA table of class is very small (2.2e-16) which means Ho can be rejected. Therefor there is high probability that the class of the item that is sold will affect the outcome (price and delivery time).

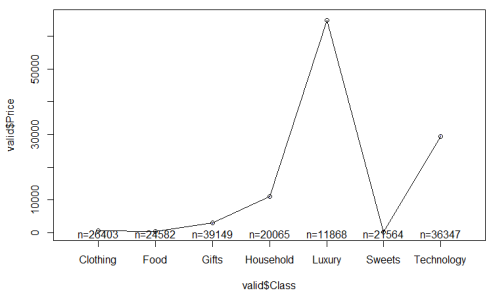


Figure 24

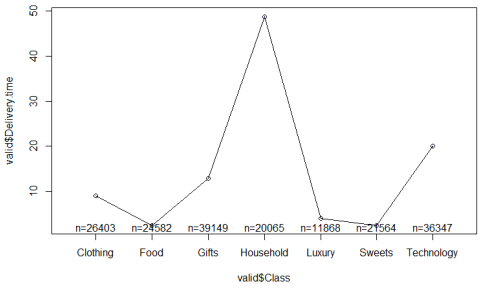


Figure 25

The mean plots shows that the average price and delivery time per class varies from class to class, this supports the conclusion that class influences price and delivery time.

Why Bought

Ho: The reason a customer buys an item will not affect the price and delivery time.

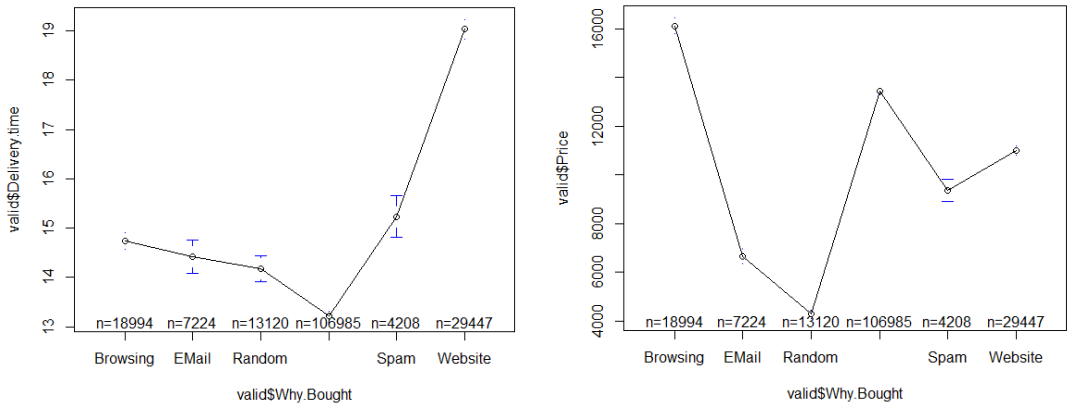
H1: The reason a customer buys an item will affect the price and delivery time.

```
 Response Delivery.time :
                     Df   Sum Sq Mean Sq F value    Pr(>F)
valid$Why.Bought      5   783320  156664  822.74 < 2.2e-16 ***
Residuals        179972 34269697     190
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Price :
                     Df     Sum Sq    Mean Sq F value    Pr(>F)
valid$Why.Bought      5 1.5742e+12 3.1484e+11  736.26 < 2.2e-16 ***
Residuals        179972 7.6960e+13 4.2762e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 26

The p value as seen in in the MANOVA table of why bought is very small (2.2e-16) which means Ho can be rejected. Therefor there is high probability that the reason an item is bought will affect the outcome (price and delivery time).



The mean plots shows that the average price and delivery time per why bought varies, this supports the conclusion that the reason an item is bought influences price and delivery time.

Part 6: Reliability of the service and products

Problem 6

6. 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 the specifications. Determine the Taguchi loss function for this situation.

Thickness = 0.06 +/- 0.04

Cost = \$45 to scrap a part outside of the given specifications

Taguchi loss function basic form: $\boldsymbol{L(x) = k\,(x - T)}$ 2

L(x) = 45

T = 0.06

K is determined by substituting the given values into the equation

45 = k (0.06) 2

K = 28 125

The final equation is therefor:

$$\boldsymbol{L(x)} = 28\ 125(\boldsymbol{x} - 0.06)\ 2$$

A graph representing the Taguchi's loss function can be constructed comparing the loss and the part size.

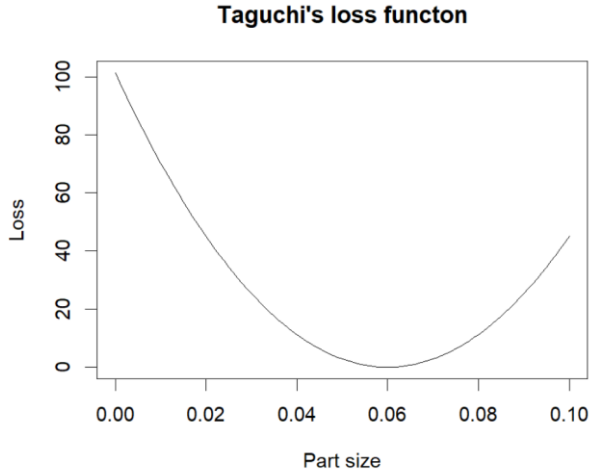


Figure 27

Problem 7

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 cost to $35 per part.
   a. Determine the Taguchi loss function for this situation.
   b. If the process deviation from target can be reduced to 0.027 cm, what is the Taguchi loss?

A) Thickness = 0.06 +/- 0.04

Cost = $35 to scrap a part outside of the given specifications

Taguchi loss function basic form: $\boldsymbol{L(x) = k\,(x - T)\,2}$

L(x) = 35

T = 0.06

K is determined by substituting the given values into the equation

35 = k (0.04) 2

K =21 875

The final equation is therefor:

$$\boldsymbol{L(x)} = 21\ 875(\boldsymbol{x} - 0.04)\ 2$$

A graph representing the Taguchi's loss function can be constructed comparing the loss and the part size.

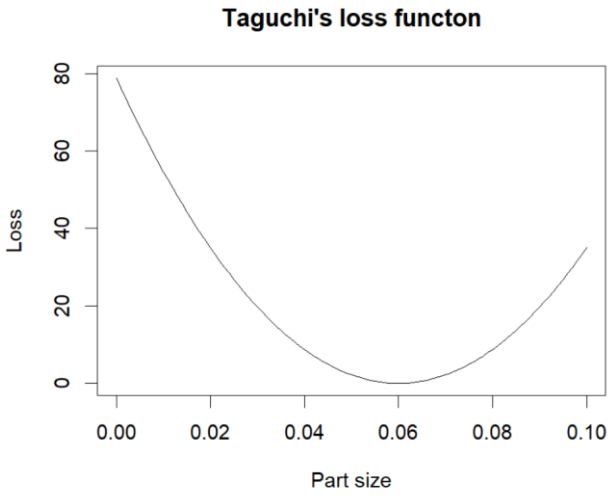


Figure 28

B)

By using the equation in question 7.B and constructing a vertical line it can be determined that the Taguchi's loss will be $15.947 per part

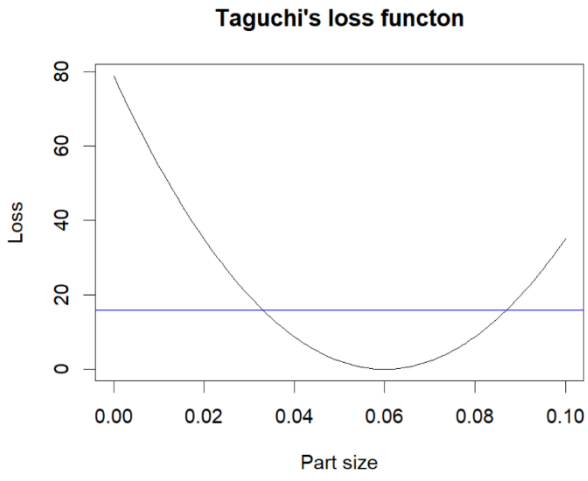


Figure 29

Problem 27

A)

Reliability = Ra x Rb x Rc

=0.85 x 0.92 x 0.90

= 70.38%

B)

$Reliability$ = [1 − (1 − RA) × (1 − RA)] × [1 − (1 − RB) × (1 − RB)] × [1 − (1 − RC) × (1 − RC)]

$Reliability$ = [1 − (1 − 0.85) × (1 − 0.85)] × [1 − (1 − 0.92) × (1 − 0.92)] × [1 − (1 − 0.90) × (1 − 0.90)]

= 96.15%

Using two machines in parallel at each station will improve the overall reliability of the process by 25.77%. This is because when one machine breaks down the process can still go on. The probability that both machines at a station breaks down is very low, therefor the reliability is much higher.

Binomial probability

Probability of x vehicles available based on the availability in the last 1560 days

$$P(x = 20) = \frac{190}{1560}$$

$$P(x = 19) = \frac{22}{1560}$$

$$P(x = 18) = \frac{3}{1560}$$

$$P(x = 17) = \frac{1}{1560}$$

$$P(x = 21) = 1 - \frac{190 + 22 + 3 + 1}{1560}$$

Probability of y drivers available based on the availability in the last 1560 days

$$P(y = 20) = \frac{95}{1560}$$

$$P(y = 19) = \frac{6}{1560}$$

$$P(y = 18) = \frac{1}{1560}$$

$$P(y = 21) = 1 - \frac{95 + 6 + 1}{1560}$$

P (≥ 20 $vehicles\ and\ drivers\ available$) = P (x = 20, y = 20) + P (x = 20, y = 21) + P (x = 21, y = 20) + P (x = 21, y = 21)

= 0.9789209402

The number of reliable days in the year can be calculated by multiplying this probability by the number of days in a year (365).

$$\mathrm{E}(\text{reliability}) = \mathrm{P}(\geq 20 \text{ vehicles and drivers}) * 365 = 357.30$$

Therefor 8 days of the year the delivery times won't be reliable, and 357 days of the year the delivery times would be reliable.

By adding another vehicle, there will be no change to the reliability of the drivers but only the vehicles. Therefor the unreliable probability can be calculated as:

$$p(unreliable) = \frac{3+1}{1560} = 0{,}00192308$$

$$p(reliable) = 1 - 0.00192308 = 0.99807692$$

Therefor the total days that the company will have a reliable delivery process in the year by adding another driver would be:

$$days = \ 0.99807692 * 365$$

= 364 Days

Conclusion

The report started by splitting the data into two separate datasets, Valid data and invalid data. The valid dataset was used for the rest of the report. A better understanding of the features was gained by using descriptive statistics. Correlations, distributions and densities was observed to find trends in the dataset. It was concluded that gifts were the highest selling products and that customers buys the most products through recommendations. The process capability indices of the class technology (referring to delivery times) were then calculated, and it showed that the process is not very capable.

A Statistical process control was done usi8ng S-Charts and X-Charts, and it was concluded that the system is fairly under control. The delivery process of the class technology was later improved by reducing the average delivery time by 3 days to prevent late deliveries at additional costs.

Probabilities of the companies, such as machine reliability, delivery reliability were calculated to prove that the decisions made were the best decisions for the company.

References

1factory. (n.d.). Retrieved from 1factory: https://www.1factory.com/quality-academy/guide-to-process-capability-analysis-cp-cpk-pp-ppk.html

Banerjee, A. (2009). Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal*, 127-131. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996198/#:~:text=A%20type%20I%20error%20(false,actually%20false%20in%20the%20population.

PQsystems. (n.d.). Retrieved from https://www.pqsystems.com/qualityadvisor/DataAnalysisTools/capability_4.6.1.php