ECSA Graduate Attributes Project

Name: Ruben Taylor

Student number: 23543744

Stellenbosch University

2022

Abstract

This report will analyse the sales of an online business to gain useful insight and information from the business. It will pre-process the data by removing instances that contains invalid data. These instances include missing values and negative values.

It will then use the historic data to identify trends and pattern as well as identifying where the business can make improvements. This information will be presented to the online business and recommendations will be made on decisions that should be made for the business to be successful and to adhere to customer requirements.

Table of Contents

Introduction

In this report the dataset of an online business is given and should be analysed. This will be done by pre-processing the data to make sure that all data within the dataset is useful and valid and that it contains no missing or negative values.

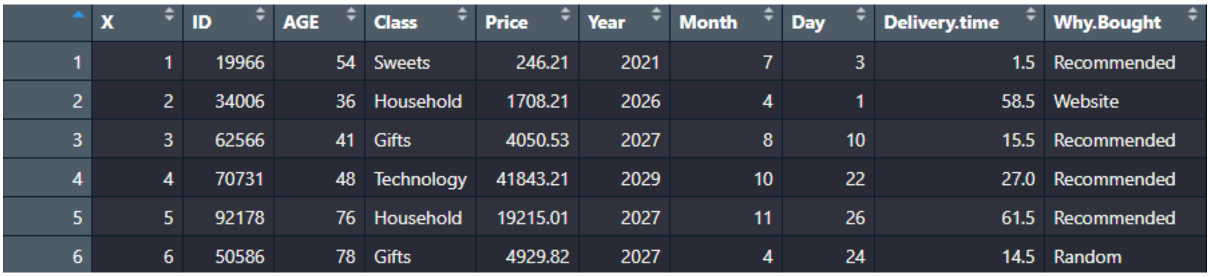
The following report will analyse the data by using historic data to identify useful information like patterns and trends that were previously unknown. It will also use the data to make recommendations and predictions on decisions that should be made to ensure that the business is successful.

Part 1: Data wrangling

This section aims to pre-process data in order to make it useful for analysis. The process will separate instances containing missing values and put them in a different dataset as these instances will lead to inaccurate analysis.

Original Data

The original dataset contains 180 000 instances with 10 different features. The dataset provides information on sales that was made during 2022.



	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

(Table 1: Original dataset with all instances)

Feature description:

X: Provides the instance within this dataset.

ID : Provides the identification number of customer.

Age: Provides the age of the customer.

Class: Provide the type of product that was purchased.

Price: Selling price of that specific product purchased.

Year: Year in which the product was purchased.

Month: Month in which the product was purchased.

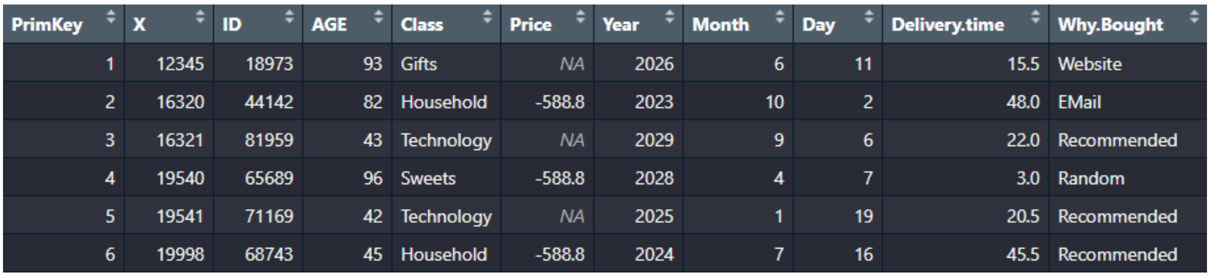
Day: Day on which the product was purchased.

Delivery time: Time from purchase to delivery.

Why Bought: Provides inside to why the product was bought and what advertisement worked.

Invalid Data

The invalid dataset contains 17 instances where the price of the item purchased was not included and contained NA in its cell. This means that there was no information or that information wasn’t entered into the system. It also contained 5 instances where the Price feature contained negative values. These instances should not be used as it can lead to inaccurate predictions. A primary key feature was included in the dataset to number the instances within the dataset. The final dataset contains 22 instances with 11 features.

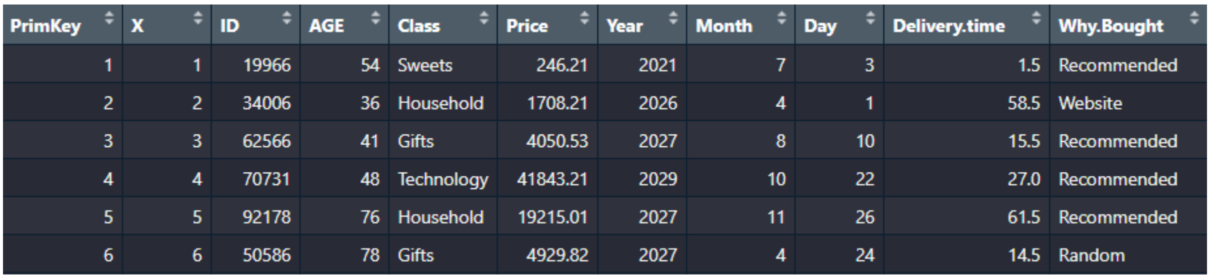


PrimKey	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	12345	18973	93	Gifts	NA	2026	6	11	15.5	Website
2	16320	44142	82	Household	-588.8	2023	10	2	48.0	EMail
3	16321	81959	43	Technology	NA	2029	9	6	22.0	Recommended
4	19540	65689	96	Sweets	-588.8	2028	4	7	3.0	Random
5	19541	71169	42	Technology	NA	2025	1	19	20.5	Recommended
6	19998	68743	45	Household	-588.8	2024	7	16	45.5	Recommended

(Table 2: Invalid data)

Valid Data

This dataset contains 179978 instances that will be used for analysis after invalid data was removed.



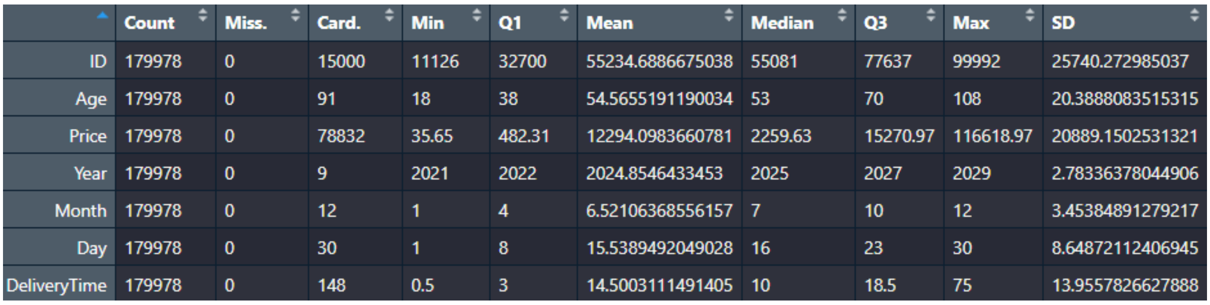
PrimKey	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	1	19966	54	Sweets	246.21	2021	7	3	1.5	Recommended
2	2	34006	36	Household	1708.21	2026	4	1	58.5	Website
3	3	62566	41	Gifts	4050.53	2027	8	10	15.5	Recommended
4	4	70731	48	Technology	41843.21	2029	10	22	27.0	Recommended
5	5	92178	76	Household	19215.01	2027	11	26	61.5	Recommended
6	6	50586	78	Gifts	4929.82	2027	4	24	14.5	Random

(Table 3: Valid data)

Part 2: Descriptive statistics

Descriptive statistics will make use of a data quality report as well as some other techniques to gain a better insight to the dataset and to extract useful information from it. Within this section only the valid dataset will be used as specified.

Continuous Features



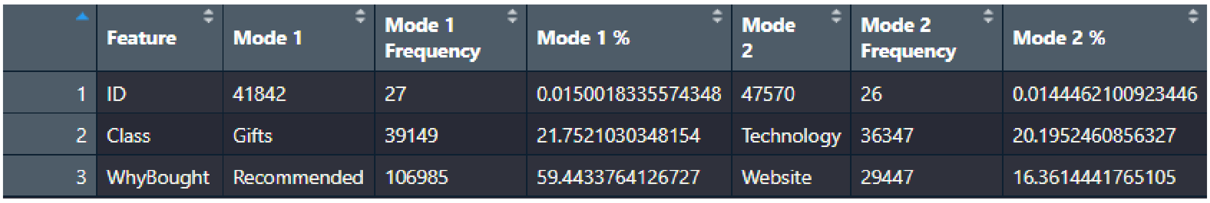
	Count	Miss.	Card.	Min	Q1	Mean	Median	Q3	Max	SD
ID	179978	0	15000	11126	32700	55234.6886675038	55081	77637	99992	25740.272985037
Age	179978	0	91	18	38	54.5655191190034	53	70	108	20.3888083515315
Price	179978	0	78832	35.65	482.31	12294.0983660781	2259.63	15270.97	116618.97	20889.1502531321
Year	179978	0	9	2021	2022	2024.8546433453	2025	2027	2029	2.78336378044906
Month	179978	0	12	1	4	6.52106368556157	7	10	12	3.45384891279217
Day	179978	0	30	1	8	15.5389492049028	16	23	30	8.64872112406945
DeliveryTime	179978	0	148	0.5	3	14.5003111491405	10	18.5	75	13.9557826627888

(Table 4: Continuous features)

From this table it can be seen that all missing- and negative values have been removed from the dataset. The cardinality of the ID feature shows that the business has 15000 customers that have bought products from them. The ages of these customers range from 18 to 108 with most customers being 53 years old. The price of items ranges from R35.65 to R116619 over 9 years.

There is no irregular cardinality for the features of month and day which indicates that daily sales occur. The delivery time for most products is within the specification limits of 0 and 24 but there are a few that is well above that needs to be addressed.

Categorical Features



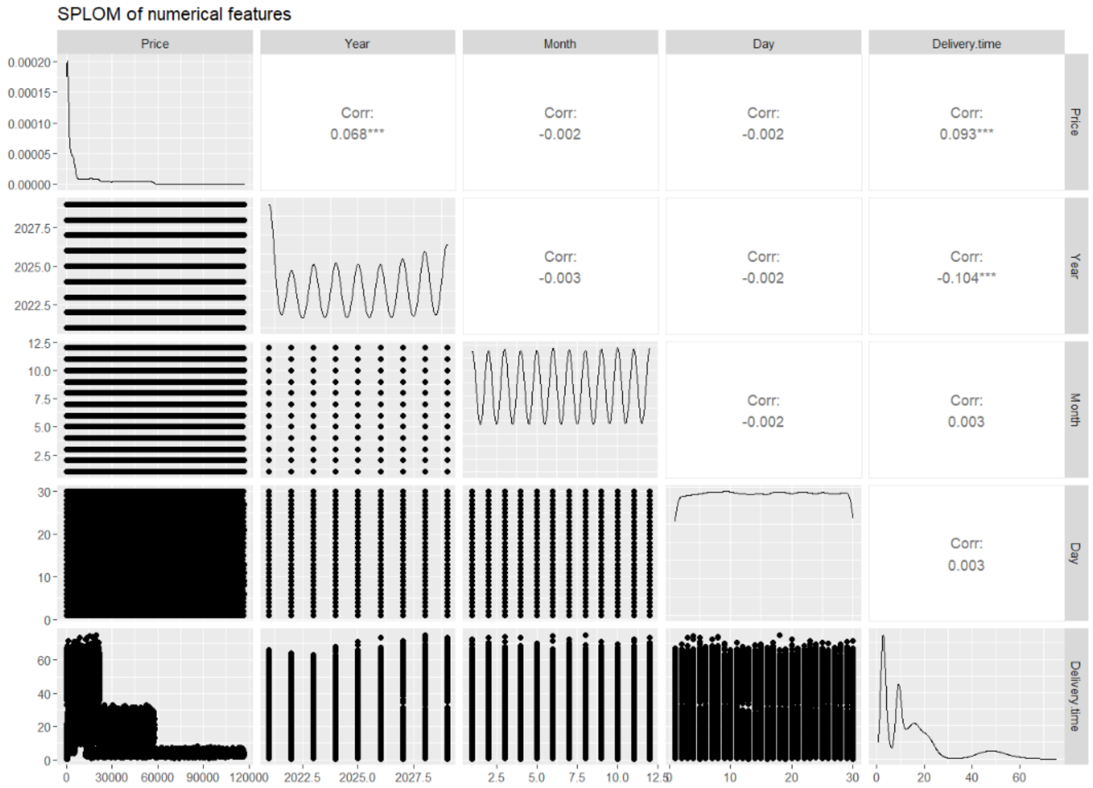
	Feature	Mode 1	Mode 1 Frequency	Mode 1 %	Mode 2	Mode 2 Frequency	Mode 2 %
1	ID	41842	27	0.0150018335574348	47570	26	0.0144462100923446
2	Class	Gifts	39149	21.7521030348154	Technology	36347	20.1952460856327
3	WhyBought	Recommended	106985	59.4433764126727	Website	29447	16.3614441765105

(Table 5: Categorical features)

This table shows the two IDs of the customers that has bought the most products from the online store. It shows that gifts are the class that is bought the most with 39149 instances. Gifts are 21.75% of all sales while Technology makes up 20.2% of sales with 36347 instances.

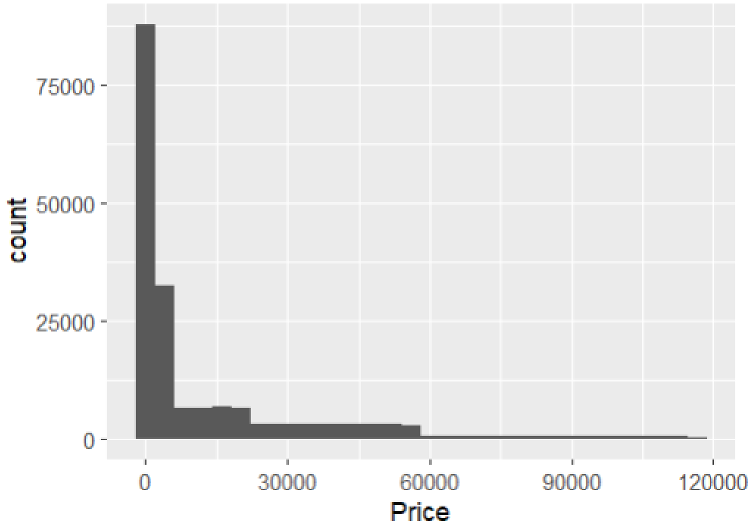
59.44% Of customers buy products from this online store because it was recommended while 16.36% buys products because of the comfort of the website. This indicates that advertisement is done by word of mouth as well as websites.

Descriptive graphs



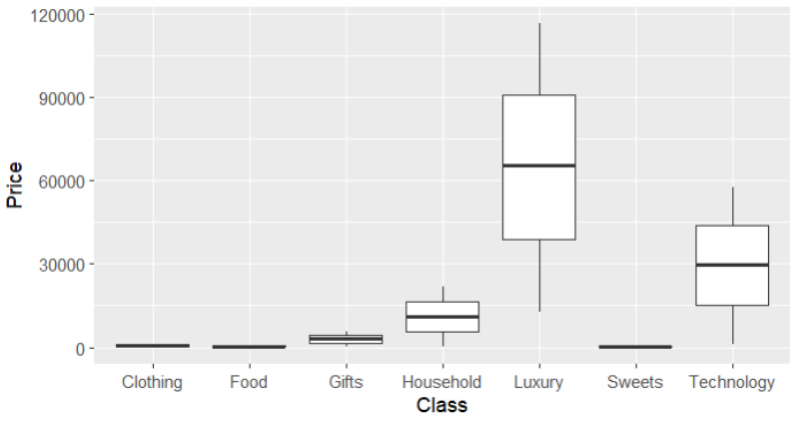
(Figure 1: SPLOM of numerical features)

The scatterplot matrix shows that features aren't strongly correlated but it is worth considering that instances within those features may be correlated to each other.



(Figure 2: Histogram that shows price distribution)

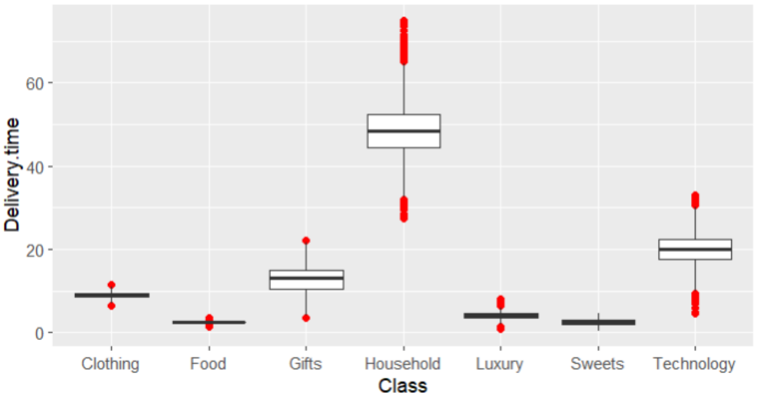
This histogram inspects the price of products that was sold at the online shop. The histogram is skewed to the right which indicates that more items were purchased in the lower price range.



(Figure 3: Boxplot showing Price vs Class)

(Hernandez, 2015)

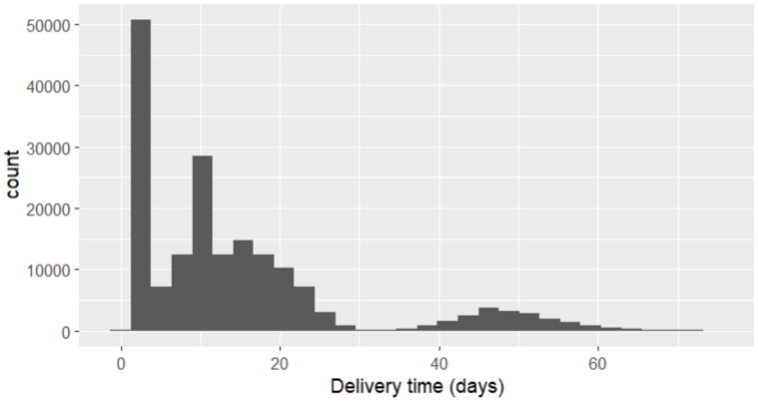
This boxplots within classes are symmetrically distributed with Luxury being the most. Technology is second most expensive and then Household. Clothing and food are the least expensive but is a necessity and will be purchased the most frequent. This explains the peak at the lower part of the histogram in figure 2.



(Figure 4: Boxplot showing Delivery time vs Class)

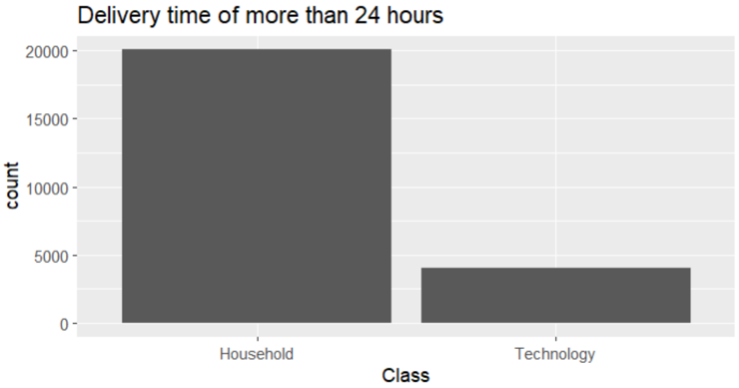
(Hernandez, 2015)

This boxplot shows the distribution of delivery times within the different classes. The distribution within classes is symmetrically distributed. The red dots indicate outliers if clamp transformation has been used. Household items has the longest delivery time while food as the shortest. This is a good sign as food can become expired and requires the shortest delivery time.



(Figure 5: Histogram showing frequency of delivery times)

This histogram is skewed to the right which indicates that most instances have a short delivery time. According to figure 4 food has a short delivery time and it has to be frequently purchased. This can cause the histogram to have higher peaks at lower delivery times.



(Figure 6: Bar plot showing frequency per class of instances where delivery is more than USL)

The USL is the upper specification limit that is based on customer requirements. The customer requires that the delivery time should be shorter than 24 hours. This bar plot shows the frequency that these requirements are not met within a class. Most instances within the technology class satisfies these requirements. Shortage of supply should be considered at technology as some products are hard to find. The high count of household items should be addressed.

Process capability

Process capability measures the business's ability to deliver parts within specific limits. These limits are specification limits and are based on the requirements of customers. The upper- and lower specification limits are given as 24 hours and 0 hours respectively. This means that the delivery time of products can't take longer than 24 hours and that it can't be delivered before it is paid for.

Delivery times for technology have a mean of 20.01 hours and a standard deviation of 3.5 hours.

Cpu	**Cpl**	**Cpk**	**Cp**
0.3797	1.9048	0.3797	1.1422

(Table 6: Process capability indices)

Cp is the process potential that indicates the spread of variation while Cpl and Cpu shows one sided process potential. Cpk is the process capability index that summarizes how a process is running relative to its specification limits. This value is lower than one which indicates that the process is not capable of meeting the delivery time requirements of the customer.

(Tsui, 1999)

Part 3: Statistical process control (SPC) for the X&s-charts

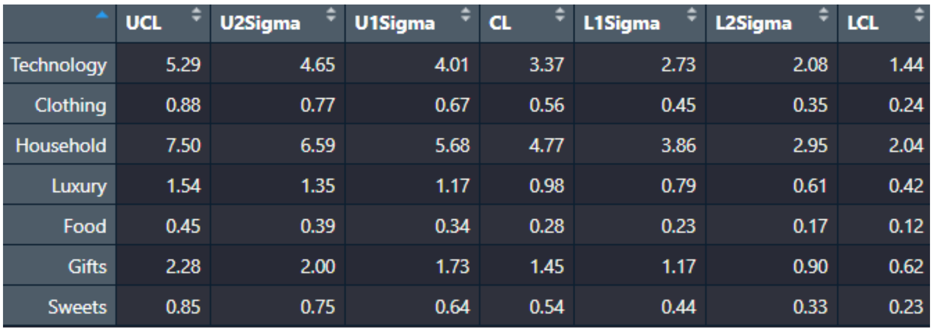
Statistical process control is a technique to analyse a process in order to monitor, control and improve that process. X-charts inspects averages of different delivery time samples while S-charts are used to evaluate standard deviation of delivery time within a process.

Chronological ordering of the dataset had to be done first. During initialization the oldest 450 instances were chosen. This resulted in 30 samples containing 15 instances each which were used to calculate the control limits.

(Anhoej, 2021)

S-Charts for different classes

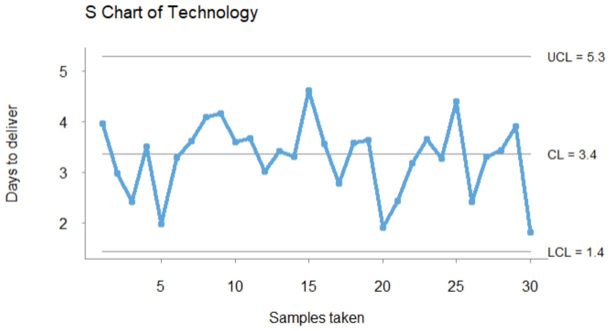
Control limits for S-chart:



	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.29	4.65	4.01	3.37	2.73	2.08	1.44
Clothing	0.88	0.77	0.67	0.56	0.45	0.35	0.24
Household	7.50	6.59	5.68	4.77	3.86	2.95	2.04
Luxury	1.54	1.35	1.17	0.98	0.79	0.61	0.42
Food	0.45	0.39	0.34	0.28	0.23	0.17	0.12
Gifts	2.28	2.00	1.73	1.45	1.17	0.90	0.62
Sweets	0.85	0.75	0.64	0.54	0.44	0.33	0.23

(Table 7: Control limits for the S-Chart)

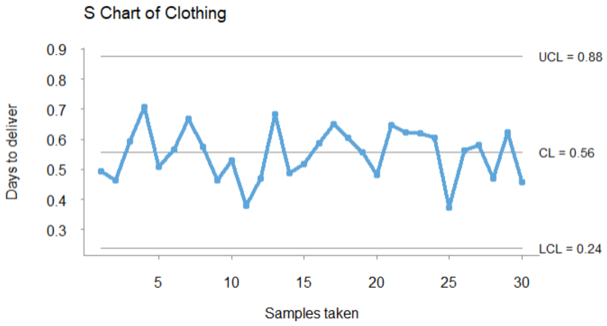
Technology



(Figure 7: S-Chart of Technology)

All samples are within limits which indicates that the process is stable and in control.

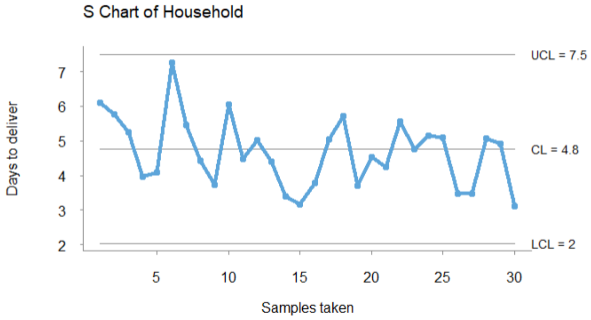
Clothing



(Figure 8: S-Chart of Clothing)

All samples are within limits which indicates that the process is stable and in control.

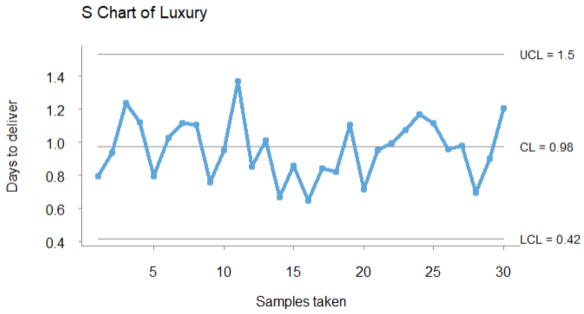
Household



(Figure 9: S-Chart of Household)

All samples are within limits which indicates that the process is stable and in control.

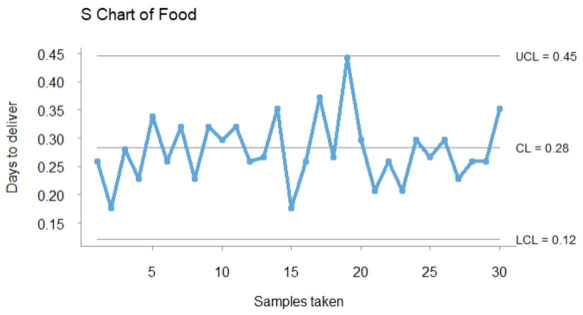
Luxury



(Figure 10: S-Chart of Luxury)

All samples are within limits which indicates that the process is stable and in control.

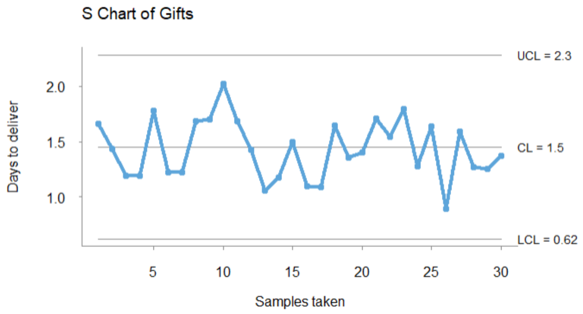
Food



(Figure 11: S-Chart of Food)

All samples are within limits which indicates that the process is stable and in control.

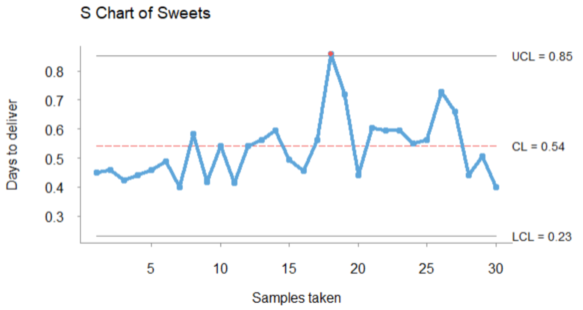
Gifts



(Figure 12: S-Chart of Gifts)

All samples are within limits which indicates that the process is stable and in control.

Sweets

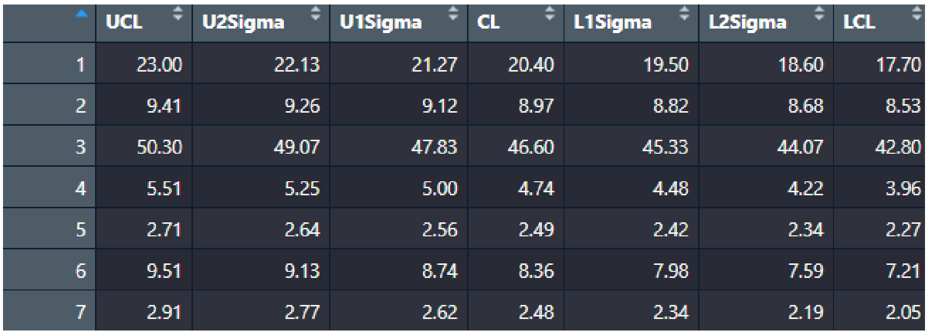


(Figure 13: S-Chart of Sweets)

One sample is above the upper control limit which indicates that a system is unstable. This can be fixed by minimising variability within this class. This can be done by having more stock or making use of more reliable suppliers.

Xbar-Charts for different classes

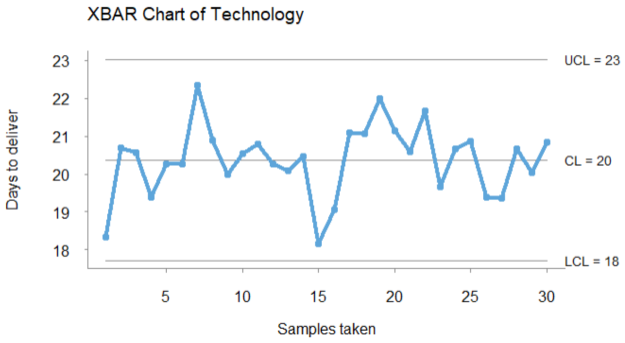
Control limits for Xbar-chart



	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
1	23.00	22.13	21.27	20.40	19.50	18.60	17.70
2	9.41	9.26	9.12	8.97	8.82	8.68	8.53
3	50.30	49.07	47.83	46.60	45.33	44.07	42.80
4	5.51	5.25	5.00	4.74	4.48	4.22	3.96
5	2.71	2.64	2.56	2.49	2.42	2.34	2.27
6	9.51	9.13	8.74	8.36	7.98	7.59	7.21
7	2.91	2.77	2.62	2.48	2.34	2.19	2.05

(Table 8: Control limits for the Xbar-chart)

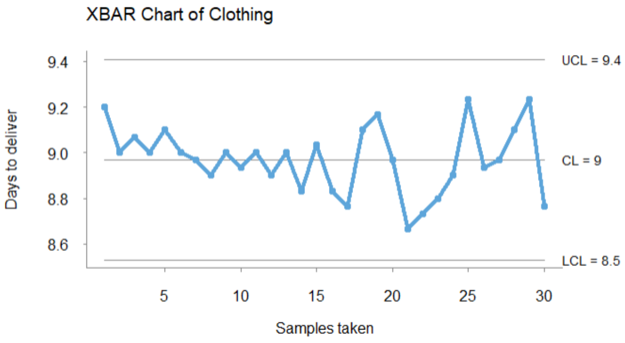
Technology



(Figure 14: Xbar-Chart of Technology)

All samples are within limits which indicates that the process is stable and in control.

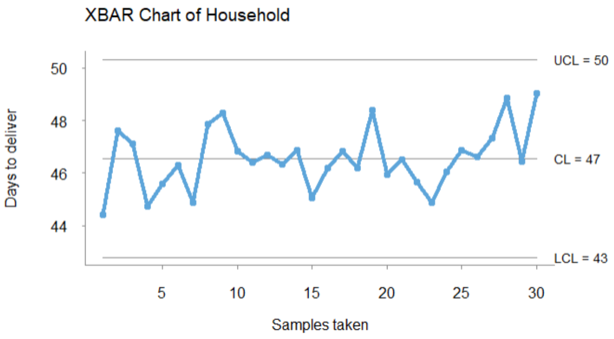
Clothing



(Figure 15: Xbar-Chart of Clothing)

All samples are within limits which indicates that the process is stable and in control.

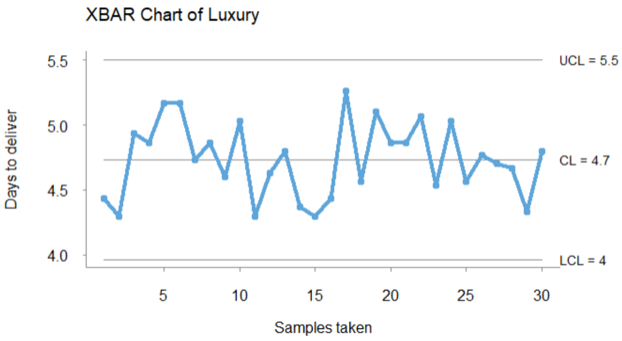
Household



(Figure 16: Xbar-Chart of Household)

All samples are within limits which indicates that the process is stable and in control.

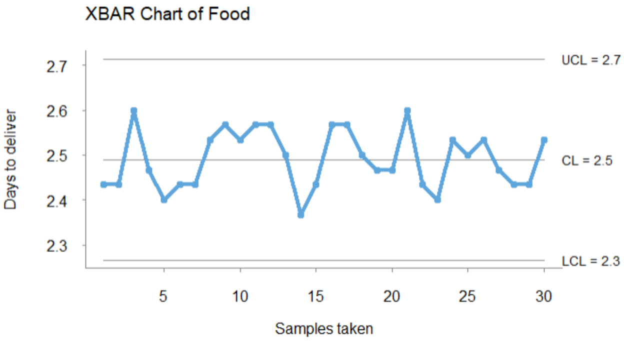
Luxury



(Figure 17: Xbar-Chart of Luxury)

All samples are within limits which indicates that the process is stable and in control.

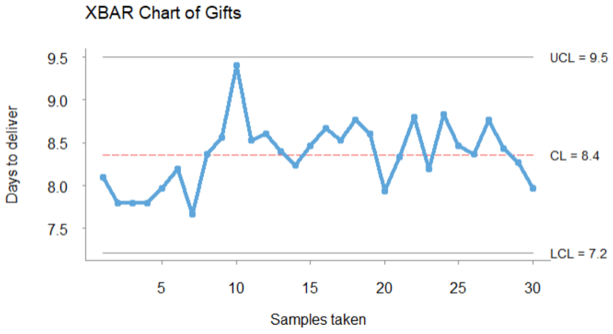
Food



(Figure 18: Xbar-Chart of Food)

All samples are within limits which indicates that the process is stable and in control.

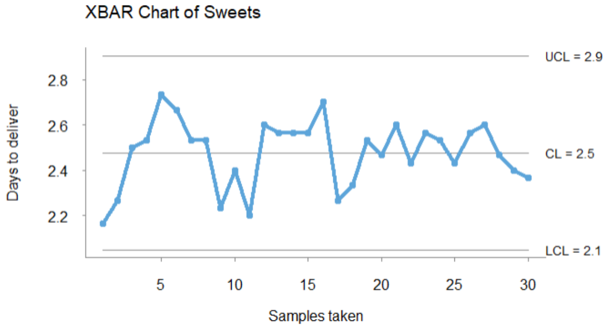
Gifts



(Figure 19: Xbar-Chart of Gifts)

All samples are within limits which indicates that the process is stable and in control.

Sweets



(Figure 20: Xbar-Chart of Sweets)

All samples are within limits which indicates that the process is stable and in control.

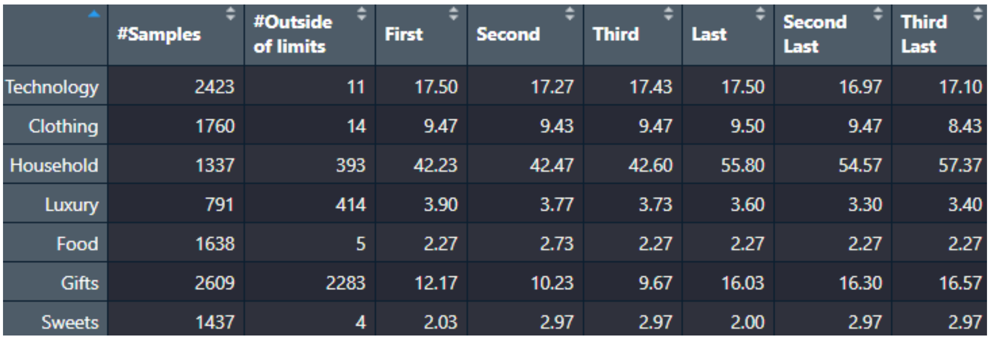
Part 4: Optimising the delivery processes

All delivery time samples will be used in this section.

4.1

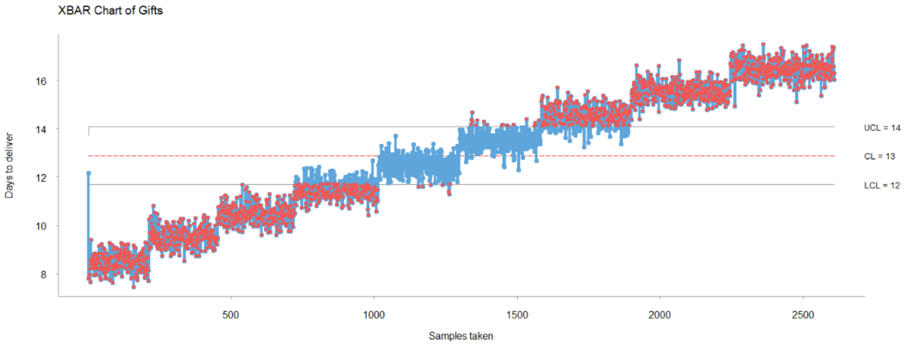
Sample means outside of the outer limits

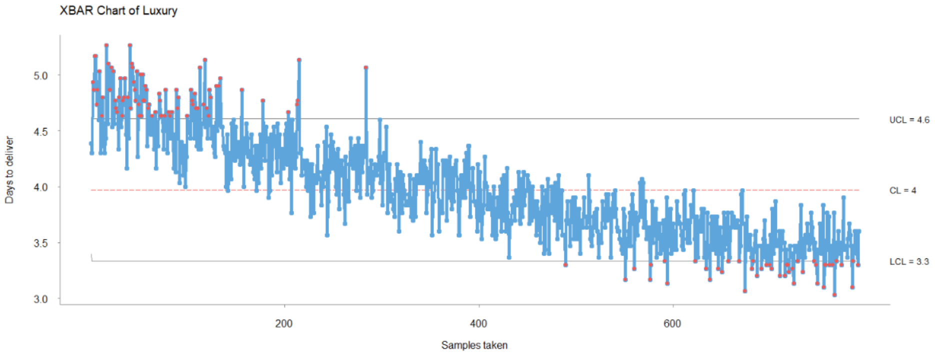
The number of samples used is different for every Class to ensure that the maximum number of samples is used and control limits of part 3 was used.

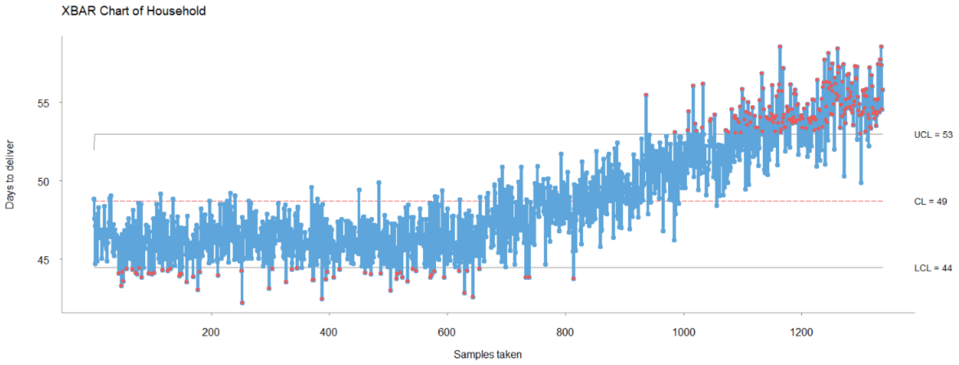


	#Samples	#Outside of limits	First	Second	Third	Last	Second Last	Third Last
Technology	2423	11	17.50	17.27	17.43	17.50	16.97	17.10
Clothing	1760	14	9.47	9.43	9.47	9.50	9.47	8.43
Household	1337	393	42.23	42.47	42.60	55.80	54.57	57.37
Luxury	791	414	3.90	3.77	3.73	3.60	3.30	3.40
Food	1638	5	2.27	2.73	2.27	2.27	2.27	2.27
Gifts	2609	2283	12.17	10.23	9.67	16.03	16.30	16.57
Sweets	1437	4	2.03	2.97	2.97	2.00	2.97	2.97

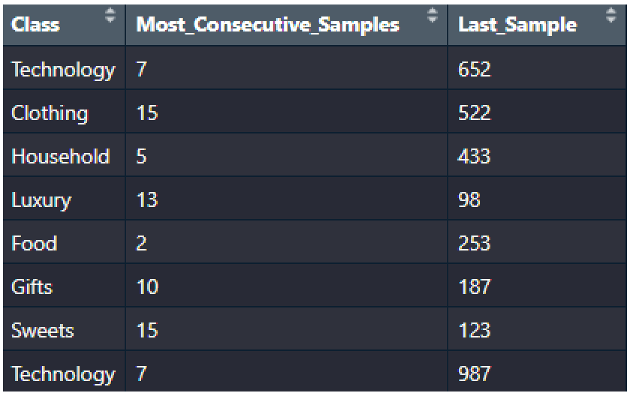
From the table it can be seen that the most samples out of limits are Gifts, Luxury and Household. This indicates that delivery time for these classes are out of control and unstable.







Find the most consecutive samples of standard deviations between -0.3 and +0.4 sigma-control limits and the ending sample number



Class	Most_Consecutive_Samples	Last_Sample
Technology	7	652
Clothing	15	522
Household	5	433
Luxury	13	98
Food	2	253
Gifts	10	187
Sweets	15	123
Technology	7	987

4.2 Estimation of likelihood to make a Type I error

A manager’s error occurs when the system is in control and the samples are within the specified control limits. The managers however thinks that something is wrong and stops the process to inspect the system. This error is also known as a false positive as the null hypothesis will be rejected incorrectly. Type I errors occur when control limits are set incorrectly.

P(Type I error for A) = pnorm(-3) * 2 = 0.002699796 = 0.27%

P(Type I error for B) = pnorm(0.4) - pnorm(-0.3) = 0.2733332 = 27.33%

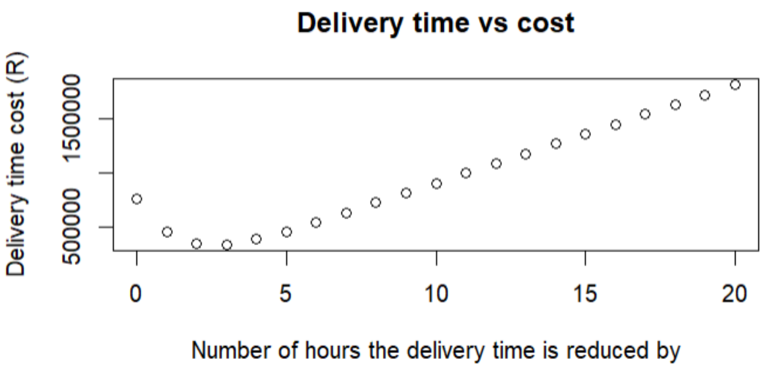
4.3 Delivery time optimization

In this section all data from the Technology class will be used which includes 36347 instances. R329 will be paid per item for every hour more than 26 hours. The cost to reduce the average time by one hour will be R2.50 per item for every hour that it is reduced by.

A brute-force method was used to determine by how much the average delivery time had to be reduced. This was done by calculating the cost at different delivery times to see where it was the lowest.

The current average of delivery time = 20 hours

The current cost of delivery time = R758 674



(Figure 21: Cost as delivery time is reduced with iterations)

From the graph it can be seen that the delivery time should be reduced by 3 hours to result in a minimum cost of R340 870.

4.4 Estimation of likelihood to make a Type II error

A Customer’s error occurs when the system is out of control but management fails to stop the system and the processes are not inspected. This error is known as a false negative as the null hypothesis is accepted when it should be rejected. Type II errors occur when control limits are too forgiving or if outliers are left out of calculations.

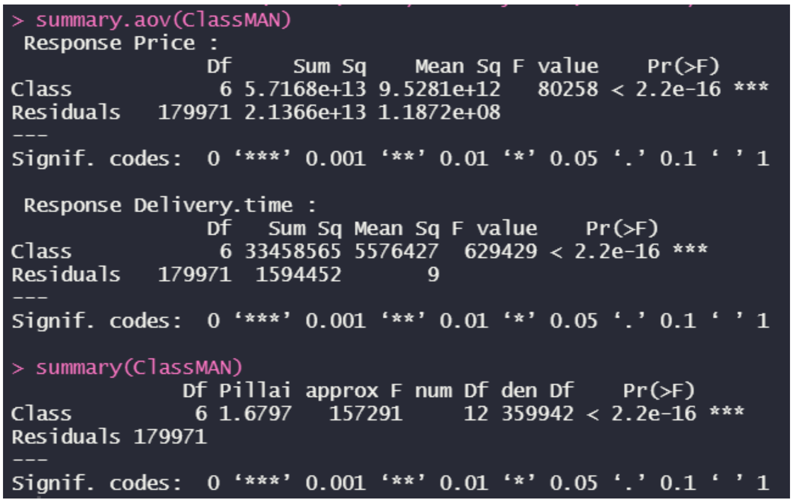
P(Type II error) = pnorm(UCL, mean, standard deviation) – pnorm(LCL, mean, standard deviation)

= 0.4999

= 49.99%

Part 5: DOE and MANOVA

MANOVA is multivariate analysis of variance that is used it situations where multiple response variables are tested simultaneously. MANOVA assumes multivariate normality and for this reason Delivery.Time and Price will be compared with Class as they are normally distributed within classes according to figure 3 and 4.



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

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

> summary(ClassMAN)
              Df Pillai approx F num Df den Df    Pr(>F)
Class          6 1.6797   157291     12 359942 < 2.2e-16 ***
Residuals 179971
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

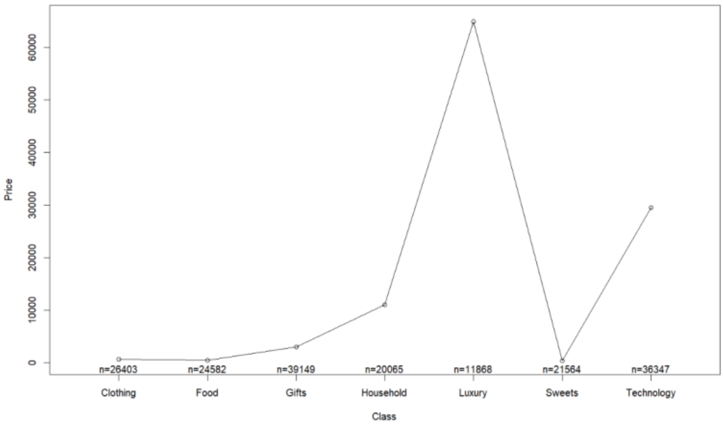
(Table 11: MANOVA comparing Price and Delivery time to Class)

H0: The null hypothesis states that price and delivery time is not affected by the class that it is in and that the class of the product does not matter.

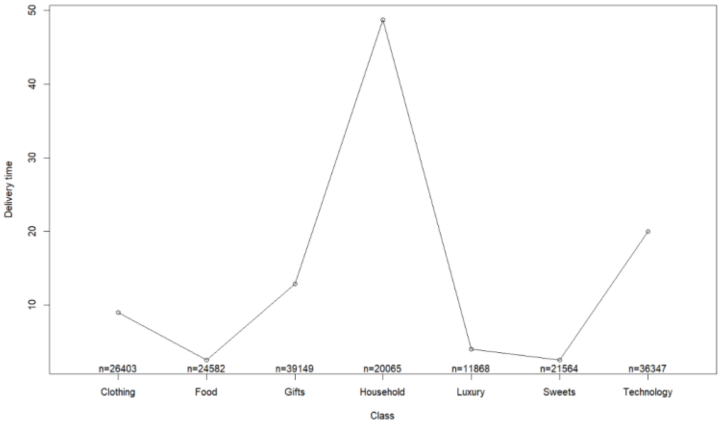
H1: The alternative hypothesis is that price and delivery time are dependent on the class that they are in and that these values will change according to their respective classes.

The small p-value of 2.2 x 10^{-16} indicates that the null hypothesis will be rejected while the alternative hypothesis will be accepted. This means that the price and delivery time will change as classes change.

(Friedrich, 2019)



(Figure 22: Plot of Price vs Class)



(Figure 23: Plot of Delivery time vs Class)

It can be seen from figure 22 and 23 that price and delivery time changes significantly within different classes. This confirms that the alternative hypothesis is indeed accepted and that these features are dependent on the class that they are in.

Part 6: Reliability of the service and products

Reliability of supply is an important aspect to a business. Unreliable supply can cause uncertainty and will make decision making difficult. Transportation plays a big role in the reliability of suppliers and is an aspect that needs to be acknowledged.

6.1

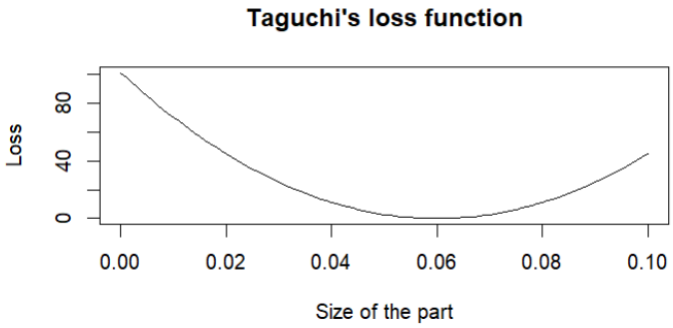
Problem 6:

L(x) = k(x-t)^2

45 = k(0.04)^2

Loss coefficient (k) = 28125

Taguchi loss function L(x) = 28125(x – 0.06)^2



(Figure 24 : Taguchi’s loss function)

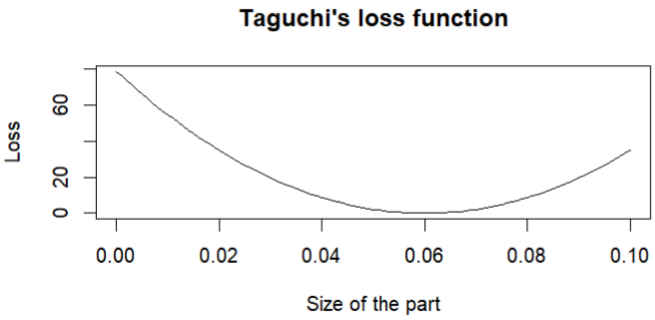
Problem 7:

a) L(x) = k(x-t)^2

35 = k(0.04)^2

Loss coefficient (k) = 21875

Taguchi loss function L(x) = 21875(x – 0.06)^2



(Figure 25 : Taguchi’s loss function)

The loss increases as the observed value is moving away from the target value. By decreasing the scrap cost, it could help to reduce the loss function.

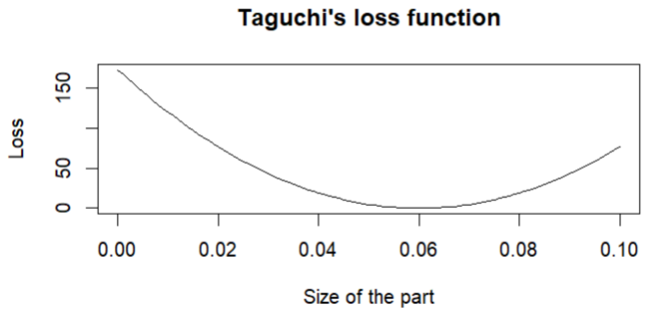
b) $L(x) = k(x\text{-}t)^2$

$35 = k(0.027)^2$

Loss coefficient (k) = 48010.97

Taguchi loss function $L(x) = 48010.97(x – 0.06)^2$

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



(Figure 26: Taguchi’s loss function)

If the process deviation is reduced to 0.027 cm the loss coefficient will be higher which will result in an increase of the loss function.

(Chin-Nung Liao, 2010)

6.2

Problem 27

The of machines working in parallel are independent of each other while machines in series are dependent on previous machines.

a) Reliability = Ra * Rb * Rc

= 0.85 * 0.92 * 0.9 = 0.7038 = 70.38%

b) Reliability = [1- (1-R(A)) * (1-R(A))] * [1- (1-R(B)) * (1-R(B))] * [1- (1-R(C)) * (1-R(C))]

= [1- (1-0.85 * (1-0.85)] * [1- (1-0.92) * (1-0.92)] * [1- (1-0.90) * (1-0.90)]

= 0.96153156 = 96.15%

The reliability improves by 25.77% when having two machines at each stage.

6.3

For the process there are 21 vehicles available while 20 vehicles are required to give a reliable service. There are also 21 drivers available while 20 drivers are required to give a reliable service.

P(unreliable service of the vehicle) = (22+3+1)/1560 = 0.01667

P(reliable service of the vehicle) = 0.9833

P(unreliable service of the driver) = (6+1)/1560 = 0.00449

P(reliable service of the driver) = 0.99551

Reliability = 0.9833 * 0.99551 = 0.9789

Expected number of reliable days = 0.9789 * 365 = 357 days

If the number of vehicles increase to 22 the reliability of the drivers would stay the same but the reliability of the vehicles will increase.

P(unreliable service of the vehicle) = (3+1)/1560 = 0.002564

P(reliable service of the vehicle) = 0.9974

Reliability = 0.99551 * 0.9974 = 0.99296

Expected number of reliable days = 0.99296 * 365 = 362 days

Conclusion

During pre-processing 17 instances that contained missing values was removed while 5 instances that contained negative values was removed. The remaining instances were used for further analysis. Descriptive statistics was then used to gain information about the business. This helped to identify aspects like product distribution between different classes as well as cost distribution between different products. The process capability indices indicates that the process is not capable of meeting the delivery requirements of the customer.

Statistical process control showed that the system is mostly stable and in control. From the MANOVA it was clear that the delivery time and price of a product is dependent on the class that it is in. Lastly the Taguchi loss function showed that scrap cost and variation within a process can significantly influence the number of defects within a system.

References

- Anhoej, J., 2021. *Control Charts with qicharts for R.* [Online] Available at: https://cran.r-project.org/web/packages/qicharts/vignettes/controlcharts.html

- Chin-Nung Liao, H.-P. K., 2010. Supplier selection model using Taguchi loss function, analytical hierarchy process and multi-choice goal programming. *Elsevier*, 4 May, pp. 571-577.

- Friedrich, K. a. P., 2019. *Introduction to MANOVA.RM,* s.l.: s.n.

- Hernandez, F., 2015. *Data Analysis with R - Exercises.* [Online] Available at: http://fch808.github.io/Data-Analysis-with-R-Exercises.html

- Tsui, K. P. &. K.-L., 1999. A review and interpretations of process capability indices. *Annals of Operations Research*, 1 April, pp. 31-47.