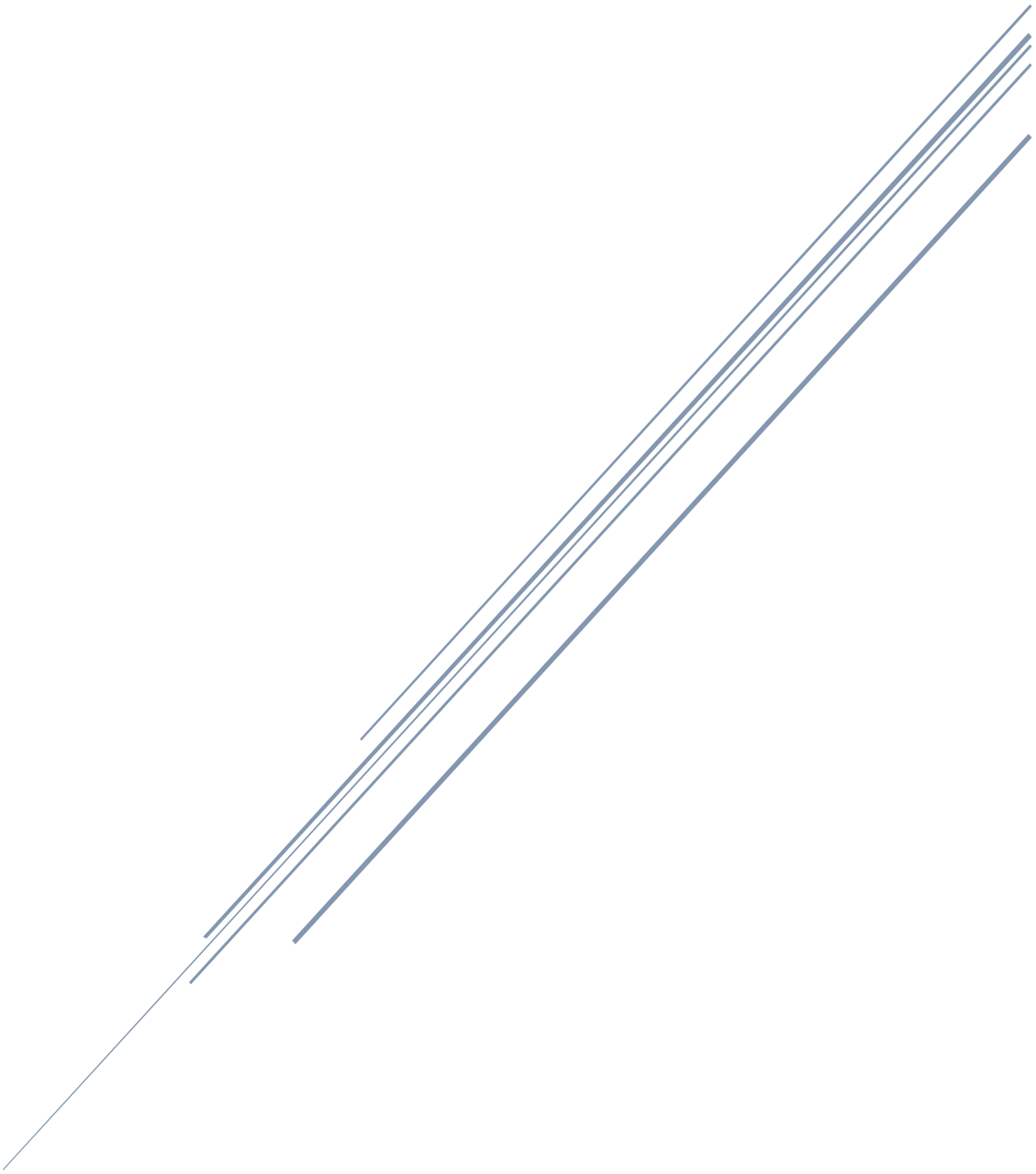
QUALITY ASSURANCE 344

ECSA GA4 REPORT



Ruan Rousseau
23690259@sun.ac.za

Table of Contents

List of figures

List of tables

Introduction

Client data for an online business has been given to be analysed, the data set contains 180 000 entries of different sales. The data will be used to determine trends and relationships between clients, product price, product classes, purchase dates and delivery times.

The data must be cleaned and manipulated before it can be used since there are invalid data in some cases. After cleaning the data descriptive statistics will be used to help us become more acquainted with the data set. After this the focus will move more intently to the delivery time and X&s – charts will be created for every class of sale. Finally, optimization will be done for the delivery process and the reliability of the service and products will be determined.

Part 1 – Data wrangling

The data set “salesTable2022” contained 180000 inputs of which various data had either missing values or negative prices which is considered as invalid data.

These data inputs were removed, and two new data sets were created. One which contained all valid data inputs, and another which contained all invalid data inputs.

The valid data set contained 179978 valid data inputs

The invalid data set contained 22 invalid data inputs

For the duration of the project, the valid data set will be used.

Part 2 – Descriptive statistics

In this part the cleaned data will be analyzed.

The cleaned data consists of 10 features.

1. X – Row number
2. ID – Identifies the Entry
3. Age – age of the client
4. Class – the category of the item purchased
5. Price – price paid for the item
6. Year – year that it was bought
7. Month - month that it was bought
8. Day – day that it was bought
9. Delivery Time - time it took for the item to be delivered.
10. Why Bought – Reason for the purchase

Firstly, an observation can be made about the primary type of customers.

The following graph shows the number of customers in every age group between 18 and 108 years old. It's visible that the most customers are between 30 and 45 Years of age, after which interest slowly starts to decrease.

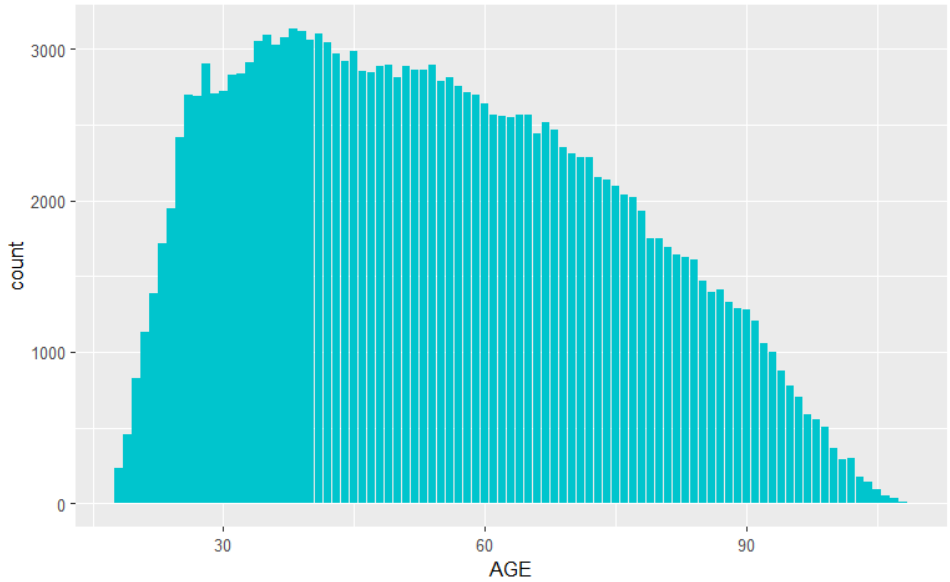


Figure 1: Customer age

Next, a closer look is taken at the products.

The following graph indicates the number of different delivery times

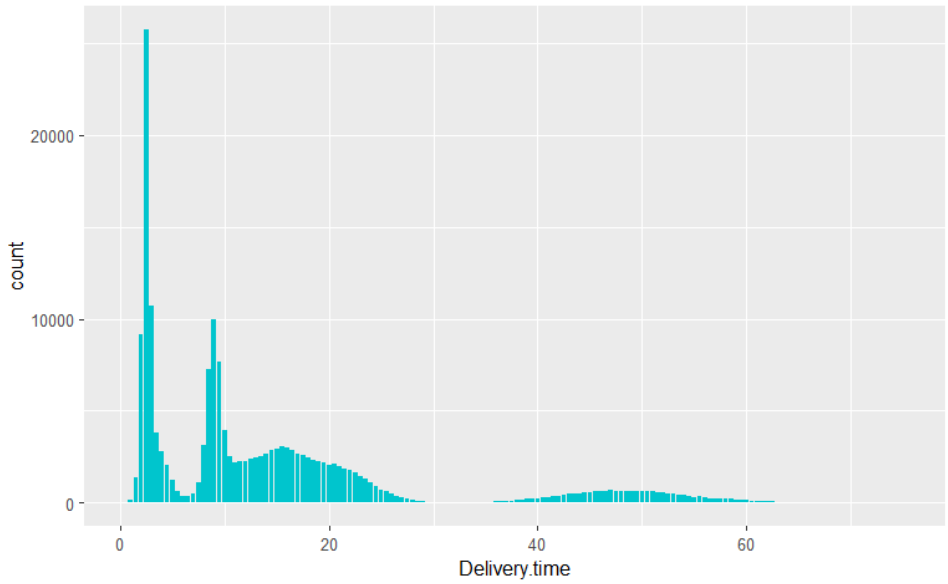


Figure 2: Delivery times

A connection is now made between the different classes of delivery times and other features such as the price of the products. The following graph compares the delivery time and the price of the product.

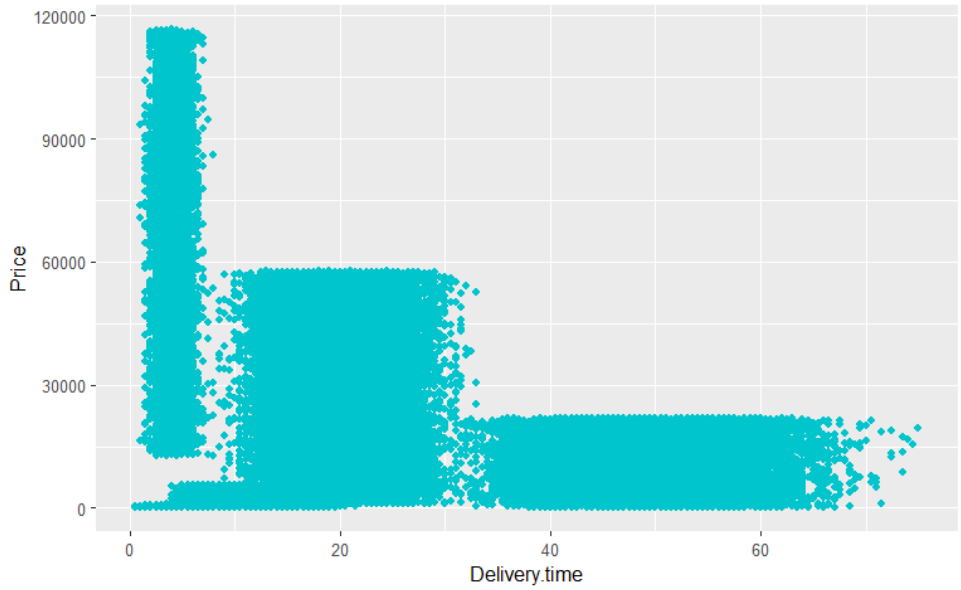


Figure 3: Price and Delivery Time

As observed, there is a clear correlation between an increase in delivery time and a decrease in price. **Products will fall into one of 3 classes**.

Class 1 = delivery time of under 10, these products can have a maximum price of +- 11500

Class 2 = delivery time of roughly 10-30, these products can have a maximum price of +- 58000

Class 3 = delivery time of above 30, these products can have a maximum price of +- 25000

It is obvious that delivery time will play a big role on the price and other characteristics of the products. So, the impact of other features on delivery time are analyzed further. The following graph indicates the delivery time for different classes of products.

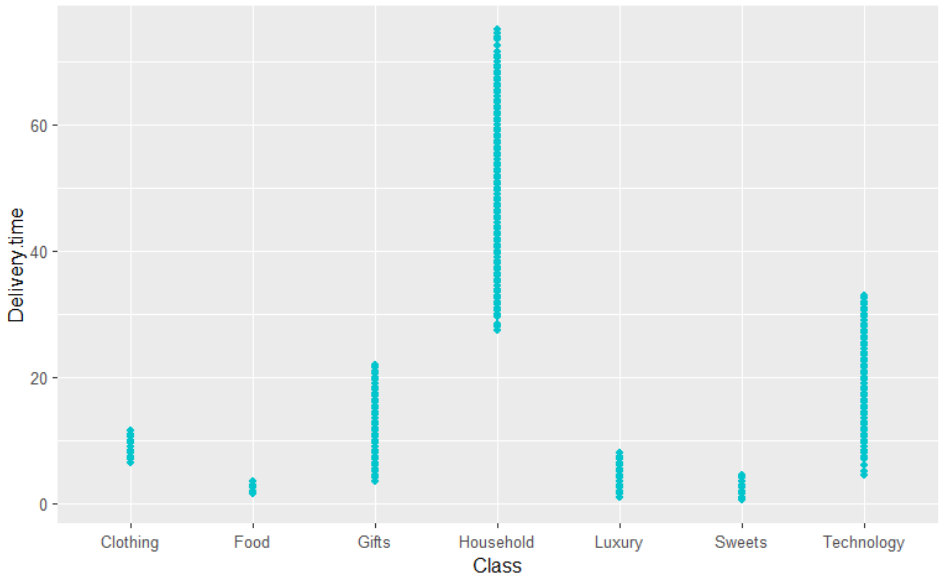


Figure 4: Delivery times of classes

The following graph looks at the prices of different classes.

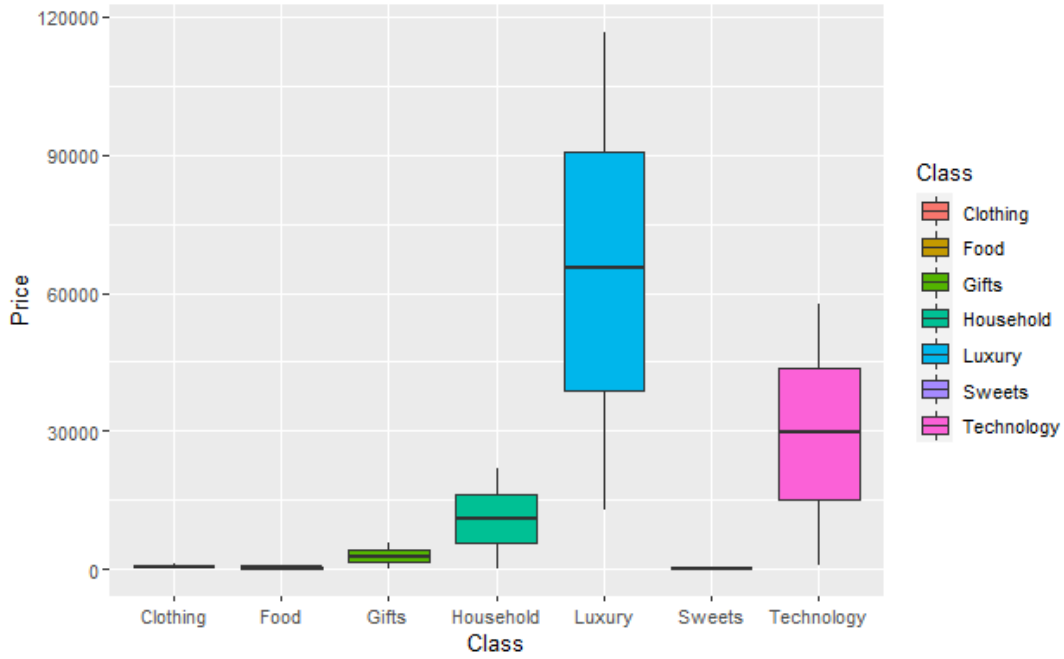


Figure 5: Prices of classes

An obvious observation can be made from this graph. Luxury and Technology sales will potentially be the way that the company can make the most revenue.

A follow up analyzation could be done to see the volume of sales for these products.

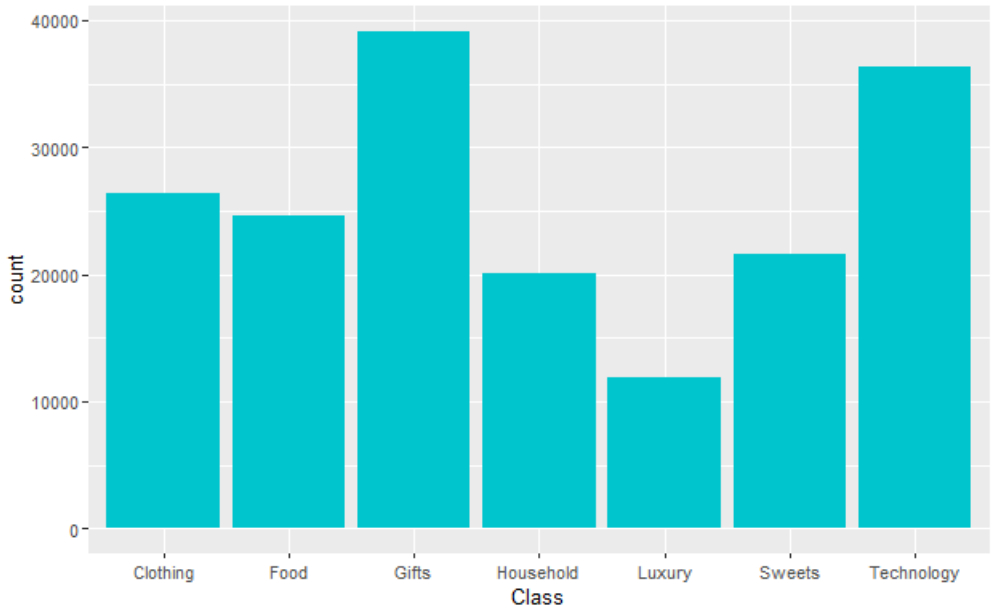


Figure 6: Class sales volume

This graph proves that luxury and technology will be the biggest revenue carriers, since these two classes of products are not a lot different in sales volume, but they are a lot more expensive than the other classes.

Process capabilities

Process Capability may be defined as the ability of a process to meet specifications.

The Upper Specification Limit (USL) has been given as 24 hours and the Lower Specification Limit (LSL) as 0. The LSL being 0 is logical because it is impossible to deliver a product in less than 0 hours.

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

$$CPU = (USL - \mu)/3\sigma$$

$$CPL = (\mu - LSL)/3\sigma$$

$$CPK = \min (CPL, CPU)$$

Process Capability indices

Cp	14.008
CPU	4.657
CPL	23.359
CPK	4.657

Table 1: Process Capabilities

Part 3 - Statistical process control

The charts were constructed with the first 30 samples of 15 sales each.

X – Table

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22.9746	22.1079	21.2412	20.3744	19.5077	18.6410	17.7743
Clothing	9.4049	9.2600	9.1150	8.9700	8.8250	8.6800	8.5351
Household	50.2483	49.0196	47.7909	46.5622	45.3335	44.1048	42.8761
Luxury	5.4940	5.2412	4.9884	4.7356	4.4828	4.2299	3.9771
Food	2.7095	2.6363	2.5632	2.4900	2.4168	2.3437	2.2705
Gifts	9.4886	9.1127	8.7369	8.3611	7.9853	7.6095	7.2337
Sweets	2.8970	2.7573	2.6175	2.4778	2.3380	2.1983	2.0585

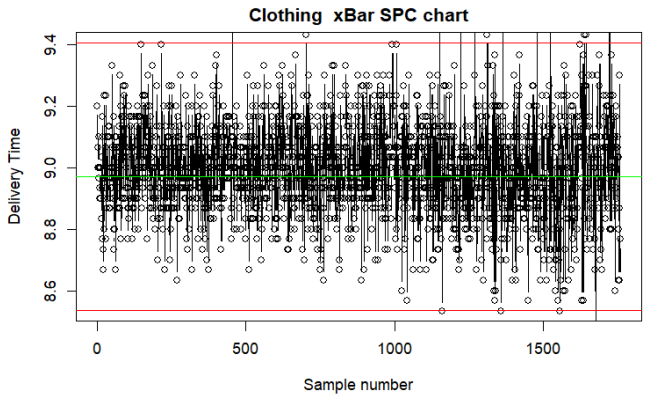
Table 2: X-Table

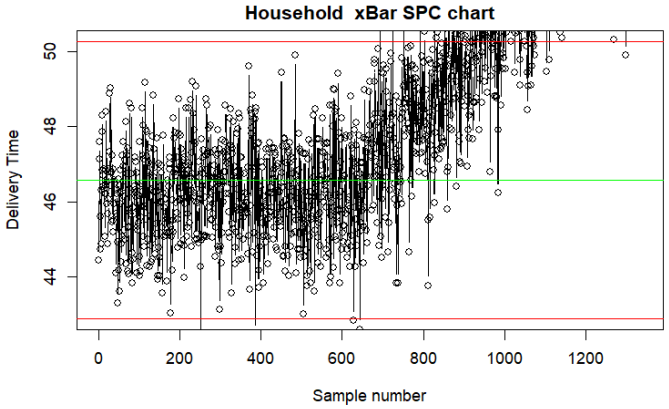
S – Table

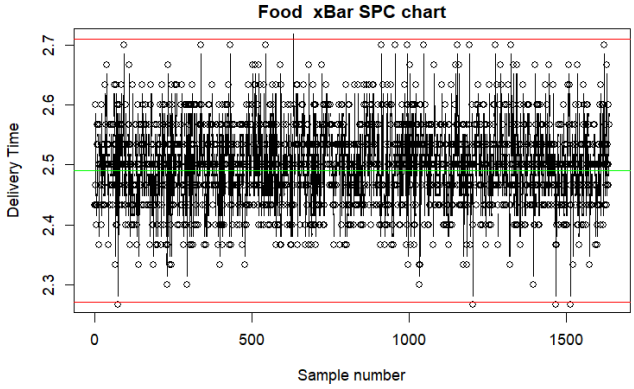
Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.1806	4.5522	3.9239	3.2955	2.6672	2.0388	1.4105
Clothing	0.8666	0.7615	0.6564	0.5512	0.4461	0.3410	0.2359
Household	7.3442	6.4534	5.5626	4.6719	3.7811	2.8903	1.9996
Luxury	1.5111	1.3278	1.1445	0.9612	0.7780	0.5947	0.4114
Food	0.4372	0.3842	0.3312	0.2781	0.2251	0.1721	0.1190
Gifts	2.2463	1.9739	1.7014	1.4290	1.1565	0.8841	0.6116
Sweets	0.8353	0.7340	0.6327	0.5314	0.4301	0.3288	0.2274

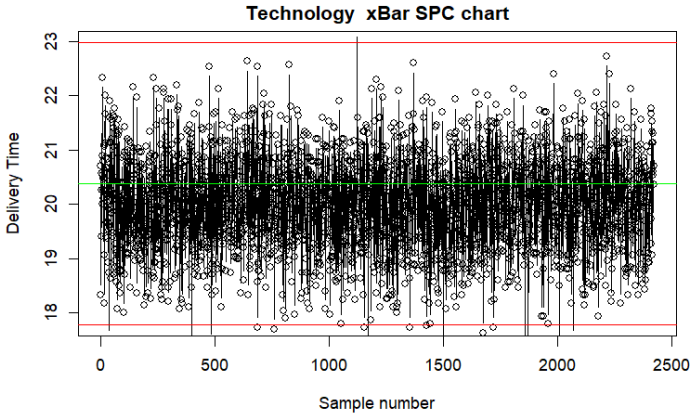
Table 3: S-Table

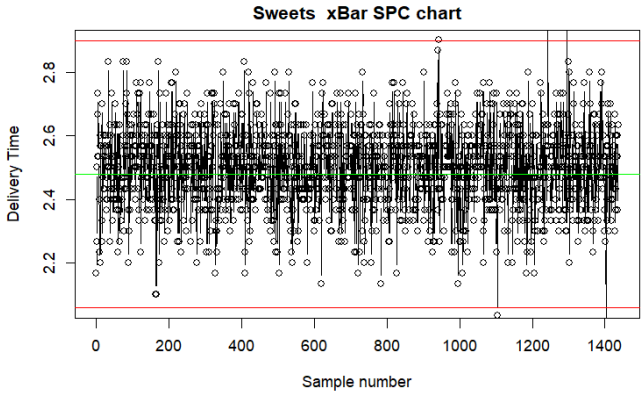
Part 3.2 Drawing samples to “control” the delivery process times

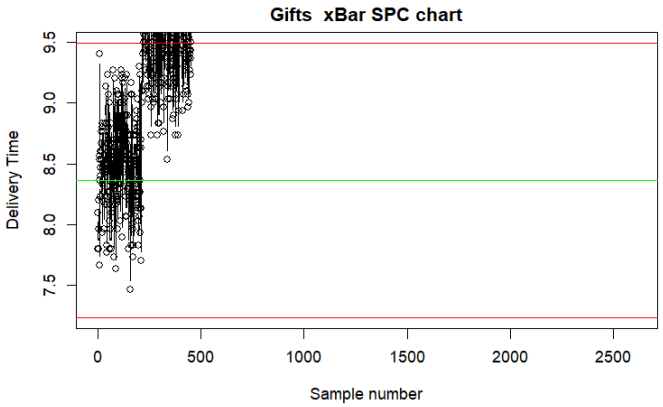












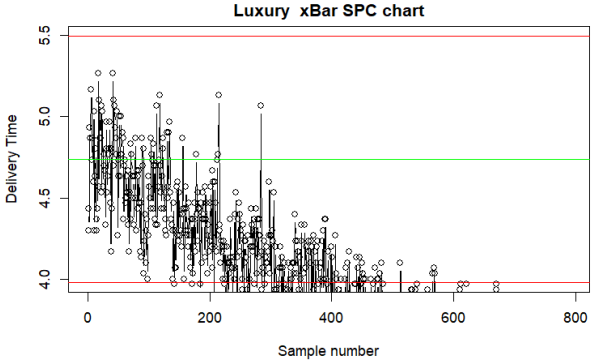


Figure 7: All samples X-Charts

The graphs that have been displayed portrays the delivery times of all samples in the data set. The green line represents the mean of the first 30 sample's delivery times.

3 of the classes show significant change in delivery times as the sample number increases. Because it is known that the data is ordered from oldest to newest, it indicates that these three classes have a change of delivery time as time went on.

The three classes that show these changes are household, gifts, and luxury.

In part 4 a closer in dept look will be taken at these three classes.

Part 4.1 A - X-bar outside of the outer control limits

Class	Total found	1st	2nd	3rd	3rd last	2nd last	Last
Clothing	17	455	702	1152	1677	1723	1724
Household	400	252	387	629	1335	1336	1337
Food	5	75	633	1203	1467	1515	NA
Technology	17	37	398	483	1872	2009	2071
Sweets	5	942	1104	1243	1294	1403	NA
Gifts	2290	213	216	218	2607	2608	2609
Luxury	434	142	171	184	789	790	791

Table 4: Outlier table

The number of outliers is extremely high for household, gifts, and luxury. Just as predicted in Part 3. The following graphs shows the outliers for the luxury and household classes.

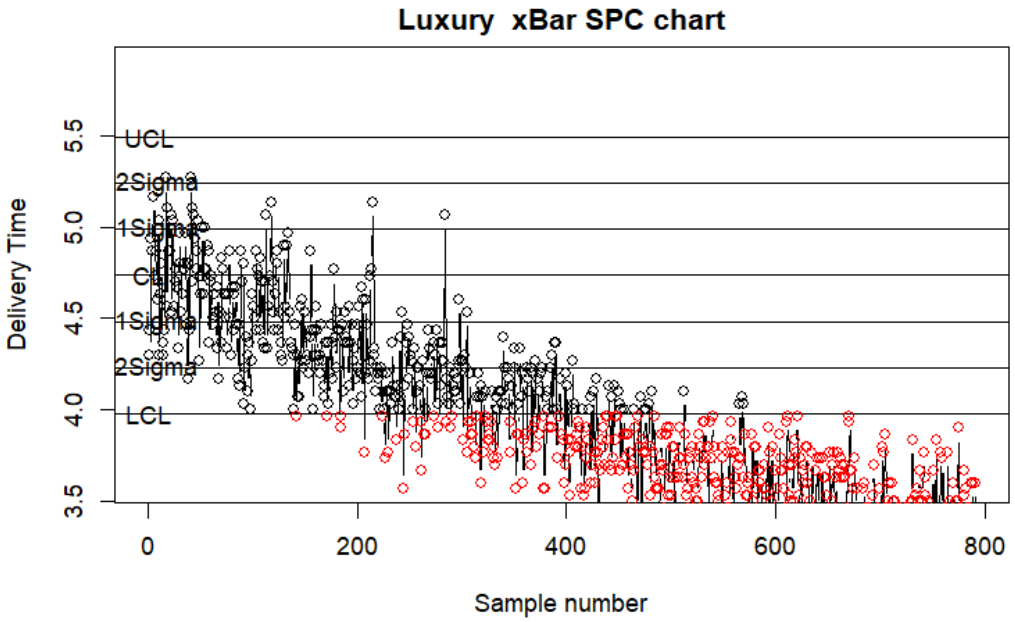


Figure 8: Outliers for Luxury

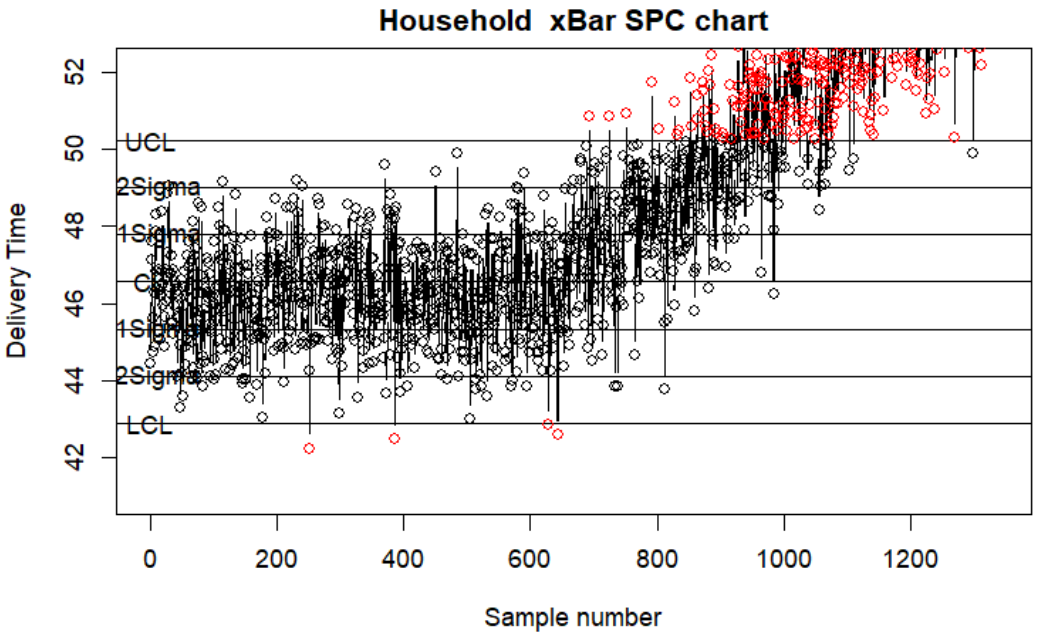


Figure 9: Outliers for household

Part 4.1 B

Find the most consecutive samples of "s-bar or sample standard deviations" between -0.3 and +0.4 sigma-control limits and the ending sample number (last sample to be in the given range)

The following table shows the total amount of errors in sequence and the ending sample number.

Class	Ending sample number	Total length
Clothing	223	4
Household	45	3
Food	43	4
Technology	63	6
Sweets	55	5
Gifts	77	7
Luxury	63	6

Table 5: Errors in sequence

Part 4.2 - Estimate the likelihood of making a Type 1 Error for A and B

The null hypothesis assumption is used to calculate a type 1 error.

H0 = The process is in control and centred on the centreline calculated using the first 30 samples

Ha = the process in not in control and has moved from the centreline

For A

The Probability of A is equal to 0.002699796. The probability of performing a type 1 error is equal to 0.27%.

For B

The Probability of B is equal to 0.2733332. The probability of performing a type 1 error is equal to 27.33%

Making a Type I error means that you make an assumption that the process is fine, until you get an indication that something may be wrong.

Therefor A and B has probabilities 0.27% and 27.33% that the process will be in error while you assume it is not in error.

Part 4.3 – Optimize delivery time

If you lose R329/item-late-hour in lost sales if you deliver technology items slower than 26 hours, and it costs you R2.5/item/hour to reduce the average time by one hour, on how many hours should you centre the delivery process for best profit? Assume the process output distribution keeps the same shape when you move the centre, and it costs you less (-R2.5/item/hour) if you increase the delivery time.

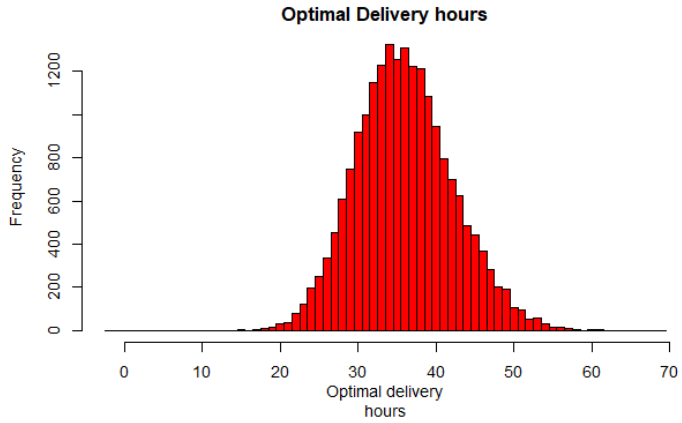


Figure 10: Optimal delivery hours

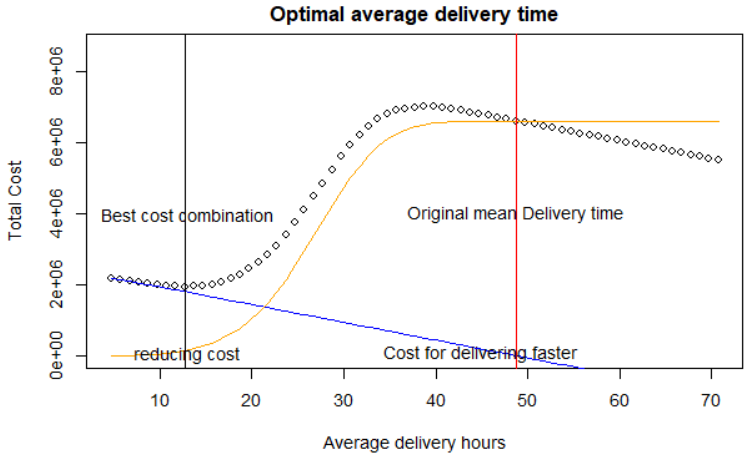


Figure 11: Optimal delivery time

In the graph different delivery times are visualized by making use of straight lines. The original average delivery time for technology was 48.718 hours and is represented in red. A blue line represents the gradient and relationship between the total cost for faster deliveries and delivery time. The difference between the average delivery time and the optimized delivery time is determined to be 36 hours. This means that the optimal delivery time is 36 hours faster than the average of 48.718 hours.

Which means that the optimal delivery time is equal to 12.718 hours.

Part 4.4 – Likelihood of making type II error for A in Class=Technology

A Type II error occurs when the company thinks that the delivery will be on time, but in reality, the product is being delivered late.

The red lines represent the outer control limits. It has been calculated that the likelihood of making a type II error is equal to 0.4883. This also represents the area of the graph that is between the outer limits.

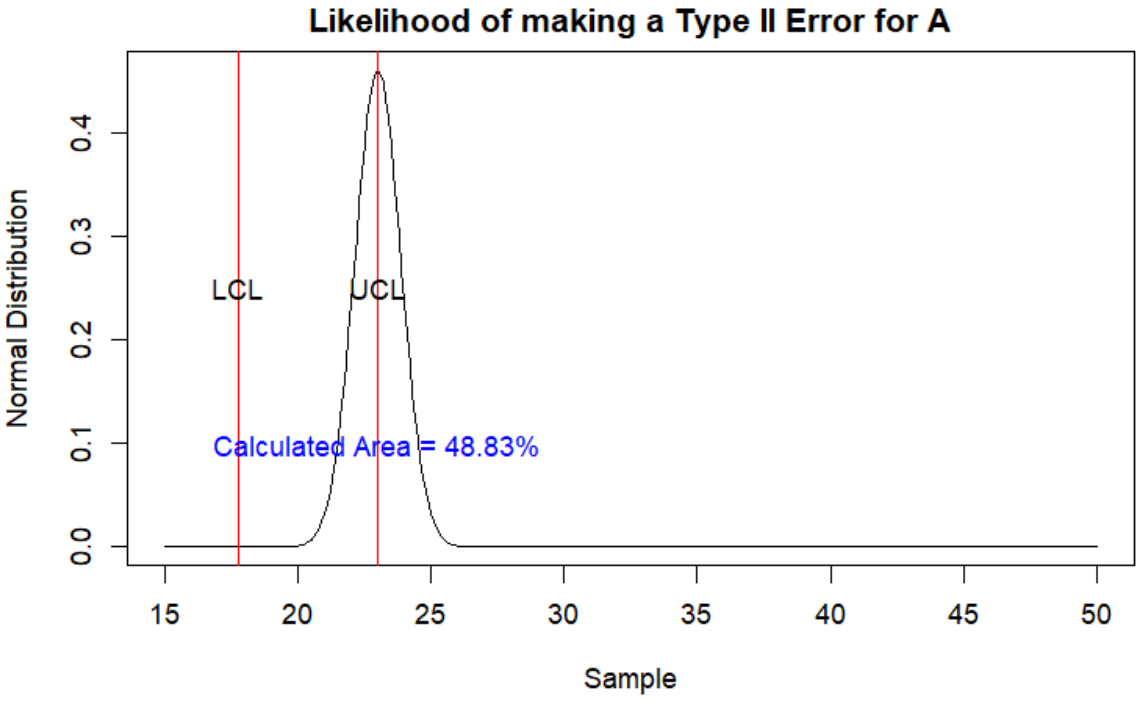


Figure 12: Type II error

Part 5: DOE and MANOVA

Dependant variables: Delivery time

Independent variables: Class, Price

The Null hypothesis (H0) is that the dependant variable of an item is not influenced by the independent variables of that specific item.

The Alternative hypothesis (H1) is that the dependant variable of an item is influenced by at least one of the independent variables.

A p value of 0.05 will be used to determine whether the dependant variable will be influenced by the independent variable.

If the p value found through MANOVA is more than 0.05, it means that the dependant variable is not influenced by the independent variable.

```
 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 13: MANOVA summary

Results

the p value for Class and delivery time is 2.2e-16

the p value for Price and delivery time is 2.2e-16

This means that the null hypothesis will be rejected, and the alternative hypothesis will be accepted, since both independent variables influence the dependant variable.

Part 6: Reliability of the service and products

Part 6.1

Problem 6

A blueprint specification for the thickness of a refrigerator part is 0.060+-0.04cm.

It costs $45 to scrap a part that is outside the specifications. Determine the Taguchi loss function.

The formula for the Taguchi loss function is $L(y) = k(y–m)^2$

The values L(y) = 45, (y-m) = 0.04 have been given, with which k can be determined

$$k = 45/ (0.04^2) = 28125$$

It is also known that m = 0.06 which means that the values can now be subbed in.

The Taguchi loss function

$$L(y) = 28125(y – 0.06)^2$$

Problem 7 (A)

For A the exact same procedure is followed as in problem 6 but with a scrap value of 35 instead of 45.

The values L(y) = 35, (y-m) = 0.04 have been given , with which k can be determined

$$k = 35/ (0.04^2) = 21875$$

The Taguchi loss function is therefor:

$$L(y) = 21875(y – 0.06)^2$$

(B)

The process deviation is now 0.027 which means that (y-m) = 0.027, therefor the loss can be worked out to be:

$$L= 21875*0.027^2 = 15.95$$

Part 6.2

Problem 27

(A) If only one machine worked at each stage

Reliability = Reliability (Machine A) x Reliability (Machine B) x Reliability (Machine C)
Reliability = 0.85 x 0.92 x 0.9
Reliability = **0.7038**

(B) How much better is two machines at each stage?

Reliability = Reliability (Set A) x Reliability (Set B) x Reliability (Set C)
Reliability = $(1 – (1- 0.85)^2)$ x $(1 – (1- 0.92)^2)$ x $(1 – (1- 0.90)^2)$
Reliability = **0.9615**

Which means that the reliability will be 25.77% better when both machines are running

The conclusion to management would be that running both machines makes the company significantly more reliable and reduces the chance of breakdowns. Meaning that it is not simply a waste to run both machines but crucial for good performance.

Part 6.3

19 Vehicles are required for reliable delivery times, which means a minimum of 19 drivers are required for reliable delivery service.

The probability that 2 vehicles break is 3/1560.

The probability of 2 vehicles breaking will be used since the total amount of vehicles will be 21 and only 19 Is required for reliable delivery time.

The probability that only 19 drivers were available was 6/1560.

Therefor the following calculation can be done to determine how many days will be expected to have reliable delivery times in the next year.

Number of reliable days per year = (1 – 3/1560 – 6/1560) * 365

= 362.89 days

Conclusion

After analysing the data some observations have been made. Technology and revenue are the biggest revenue creators for the business. This is currently good news for the business because the analysis of the data indicates that technology has a very stable delivery time and the luxury items delivery time is decreasing, meaning that there could potentially be more revenue available.

Two concerns is that the delivery time for products from classes household and gifts have increased drastically. This will lead to less income from these classes, because we know that price of products are dependant on the delivery time of products. This determined in the MANOVA test done in part 5.

Apart from these two concerns, the company is stable and growing steadily. If they are dealt with , the company will most likely continue to grow.

Bibliography

D.R. Kiran, in Total Quality Management, 2017.Process Capability. [Online] Available at: https://www.sciencedirect.com/topics/engineering/process-capability-index [Accessed 20 October 2022].

Fernando Hernandez, January 10, 2015. Data analysis with R-exercises. [Online] Available at: http://fch808.github.io/Data-Analysis-with-R-Exercises.html [Accessed 20 October 2022].

Taguchi Loss Function. [Online] Available at: https://www.whatissixsigma.net/taguchi-loss-function/ [Accessed 20 October 2022].

QA344 statistics, 2022. *Course document on SunLearn*, Stellenbosch.