ECSA GA4 Report

HP KOTZé

24087351

Due date: 21 October 2022

Table of Contents

List of Tables

List of Figures

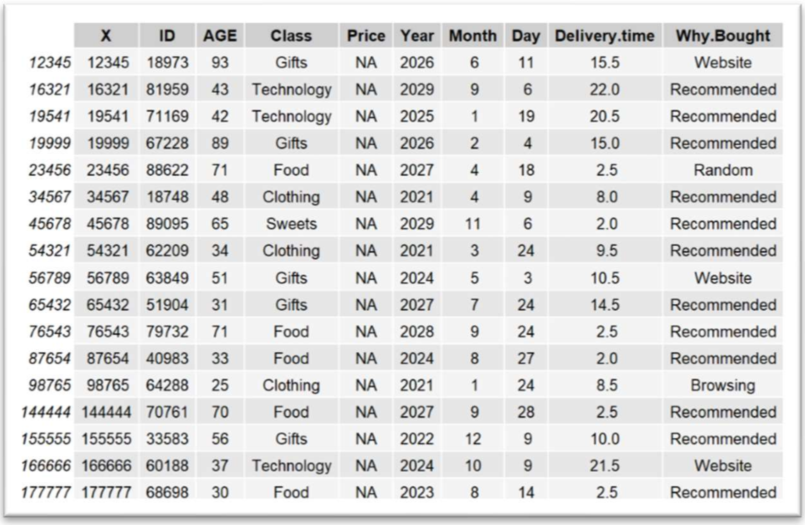
Introduction

An online business has compiled a dataset based on client data from 2021 to 2029. To gain a better understanding of the correlation between the different features, various plots and tables will be designed using R code. This will help the online business to plan for future sales based on the assumptions and tests done on the historical data. It will also improve the delivery time to the customer. Furthermore, descriptive statistics with the use of x-and s charts as well as DOE and MANOVA will be major components to help the business move forward with the decisions made from the results obtained. Finally, a brief use of Taguchi loss function will be demonstrated.

Part 1: Data Wrangling

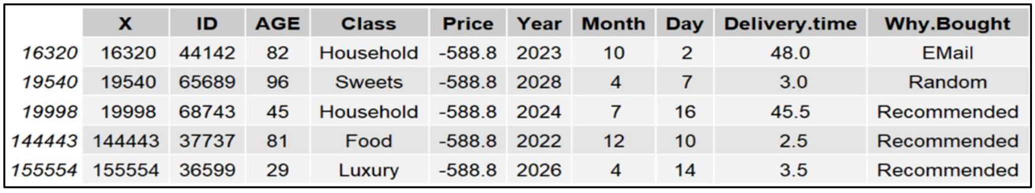
The original sales table had various instances that contained missing values or negative values. This is impossible since the minimum value for any of these features are 0 and cannot have a lower value. The data needs to be filtered and a new complete dataset should be created. In total there were 17 missing values. These instances will have a negative influence on the observations. The instances containing any missing values were removed since they do not contribute to the analysis of the online business. As mentioned above, the Price feature also had 5 negative values that needed to be removed. After filtering the sales table, 179978 instances were left for evaluation and the dataset was complete

All these instances that do not contribute to the data evaluation were removed with the use of the subset and complete cases functions. These functions ensured that data showing NA or negative values are removed in large datasets without having to do so manually for each instance. All the missing values are shown below:



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

Table 1:Missing Values



	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
16320	16320	44142	82	Household	-588.8	2023	10	2	48.0	EMail
19540	19540	65689	96	Sweets	-588.8	2028	4	7	3.0	Random
19998	19998	68743	45	Household	-588.8	2024	7	16	45.5	Recommended
144443	144443	37737	81	Food	-588.8	2022	12	10	2.5	Recommended
155554	155554	36599	29	Luxury	-588.8	2026	4	14	3.5	Recommended

Part 2: Descriptive statistics

Various plots were coded to give the needed insights to the relationships between the various features of the sales table.

2.1 Cardinality of dataset:

The cardinality is important for any dataset and especially those with many instances. Cardinality is known to be the number of distinct values in a column or for every descriptive feature when working with datasets such as the sales table (SolarWinds, 2020).

The cardinality for the different features whereas follow:

X: 179 978-this is irrelevant since the feature has no impact on the dataset and will not be used for any evaluation.

ID: 15000- This is the same as the X-feature and will have contain relevant information for the online business to use.

AGE: 91- This indicates that there are 91 different age groups that is used in the valid dataset. This can also indicate to the business that there is a wide variety of clients buying from the online business

Price: 78 832- A large cardinality for price is expected since there are several classes to buy from and a large volume of different products.

Year, Month and Day: 9, 12 and 30 respectively- The valid dataset works over a period of 9 years from 2021 to 2029. The months and days per month is known. The 12 months is per year and there are 30 days per month.

Delivery time: 148- There are numerous times when looking at delivery. This, again,

Class: 7- This shows there are several different items to choose from the online business which consists of luxury, food and household products to name a few.

Why. Bought:6- the number shows that there are unique ways to buy the items. This includes spam, website and browsing to name a few.

2.2 Price

Price variations count

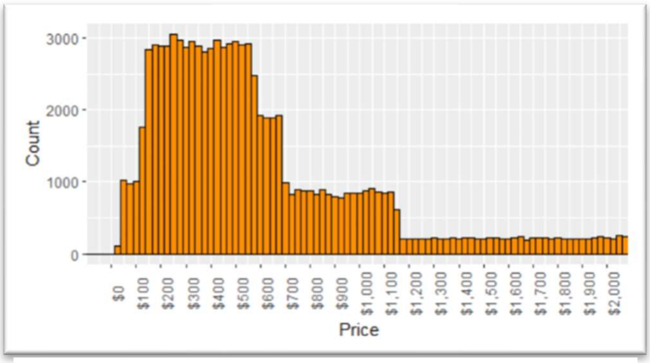


Figure 1:Price Vs Count

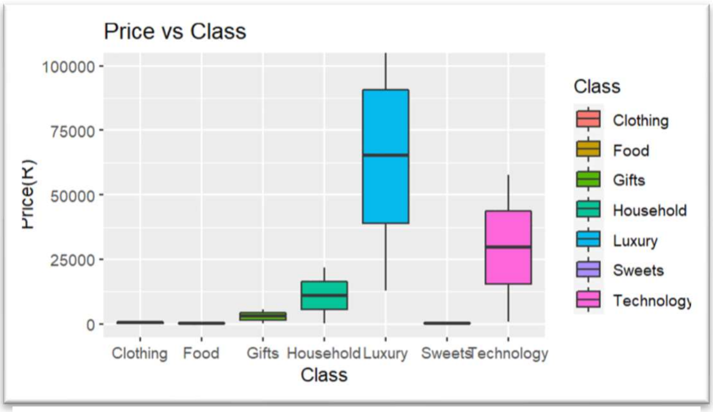
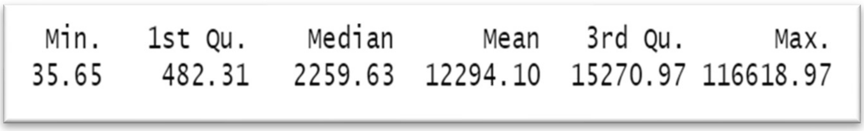


Figure 2:Price vs Class



```
  Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 35.65    482.31   2259.63  12294.10  15270.97 116618.97
```

The histogram clearly indicated that most of the products bought, range between $200 and $1100 with little instances more expensive. The summary and boxplot for the prices also shows the wide variety of products that are bought with most of the items costing roughly R2250 with a maximum way bigger at R116618.97. The prices for the luxury products range dramatically as shown on the boxplot and is much more expensive than the food, clothing and sweets as expected.

2.3 Age of customers

Age distribution of sales:

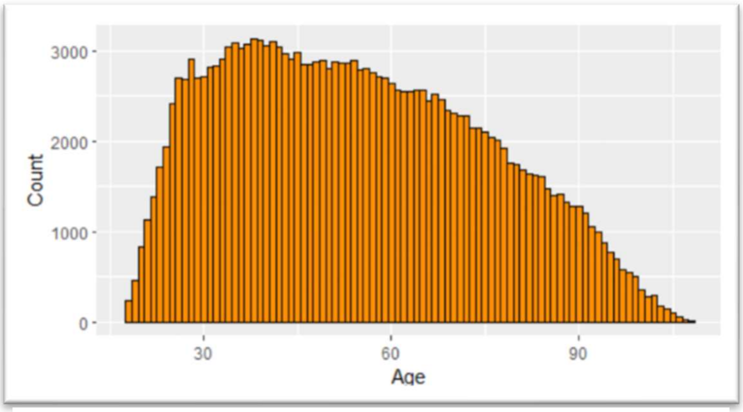
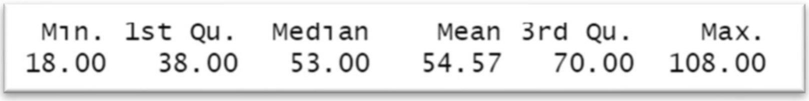


Figure 3: Age distribution of sales



```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 18.00   38.00   53.00   54.57   70.00  108.00
```

For any online business the target audience is an important factor to consider. The histogram is widely spread with most customers between 30 and 60 years old. The histogram is skewed to the right slightly. Overall, no dramatic changes in the target audience were found.

2.4 Delivery Times

Distribution of delivery times:

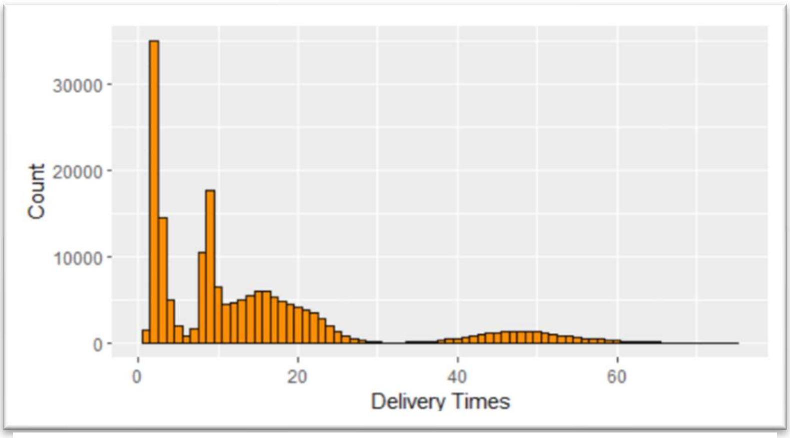


Figure 4: Delivery times count

The delivery time must be as short as possible to maintain a high level of customer service and loyalty. The histogram shows that most of delivery times (in hours) are less than 20 but then between 40 and 60 hours there is another small peak. The possible answer for the second peak is the order of the household items which has a longer delivery time.

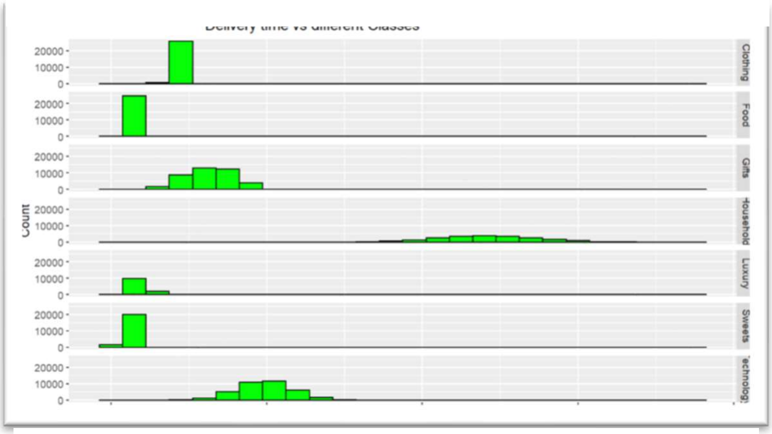


Figure 5: Delivery time vs Class

The possible answer for the second peak is the order of the household items as shown at the left. Everyday items such as sweets, clothing and luxury have shorter times to deliver.

2.5 Class

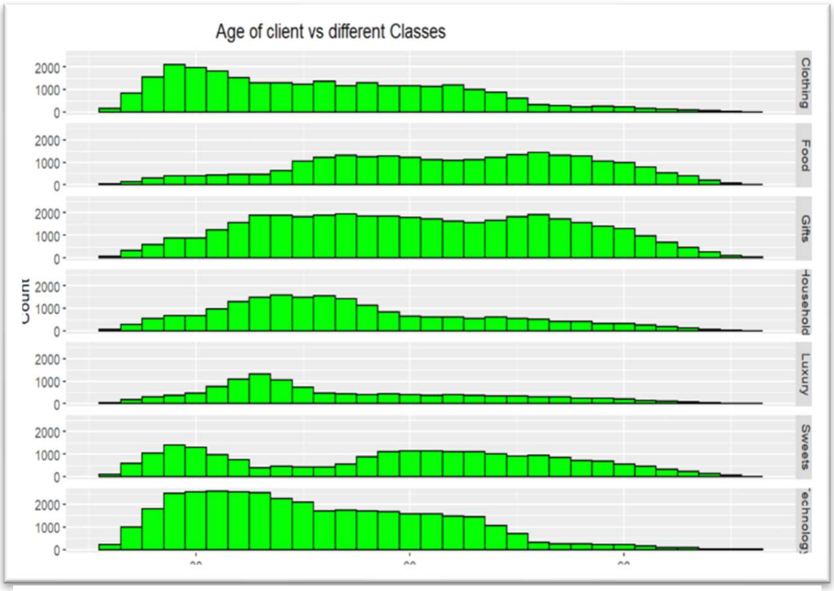


Figure 6: Age vs Class

The distribution for the different classes is accurate to what many would think before looking at the results visually. The distribution of the histograms ranges from 18 to 108 years old on the x-axis. Clothes for example become less relevant as you age and it shows with less older people buying clothes. Certain classes such as food and gifts remain consistent while technology as expected is more relevant to the younger generations with a peak early over the x-axis. The sweets items is distributed equally throughout the years.

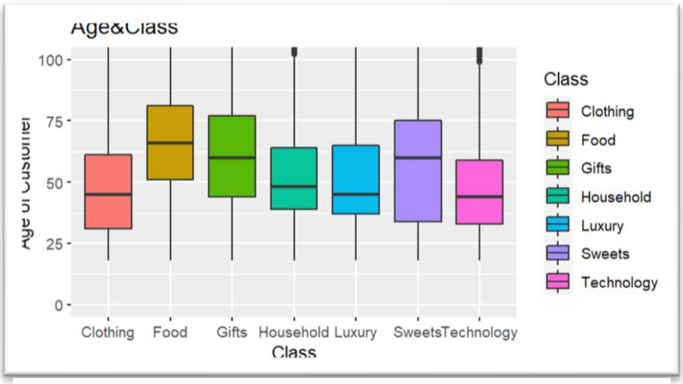


Figure 7: Age vs Class

The boxplot gives indication how certain classes are more relevant to a certain age than other classes might be.

Revenue generated by each class

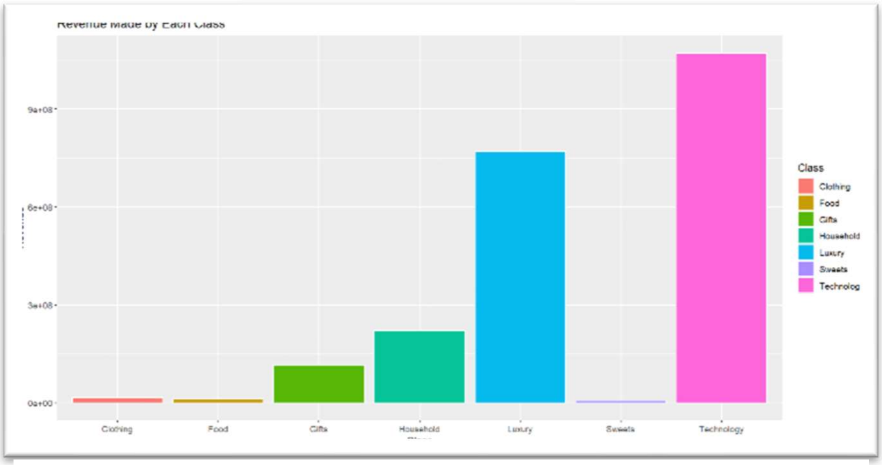


Figure 8: Revenue per Class

The result for each class shows what the best sellers are over the years in terms of revenue generated. Technology and luxury are the clear leaders in this category. It would be a smart move for the online business to invest in these 2 classes since they can help grow the financial needs of the business and make predictions easier although they are not necessarily sold the most.

2.6 Month of purchase

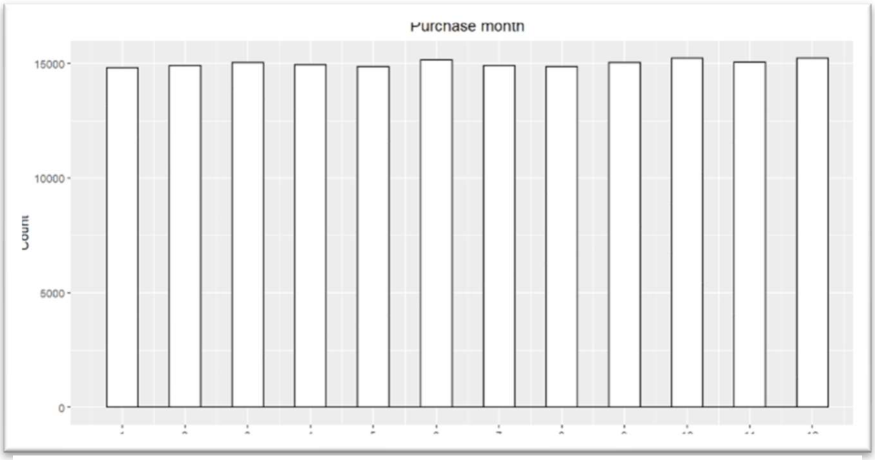


Figure 9:Month of Purchase

The month of purchase shows no clear outlier or difference in the quantity at which products are purchased. Month 1 to 12 are very close and after inspection month 6(June) and month 12 (December) have the highest values. This is usually in the holidays which will have more orders from customers.

2.7 Reason for purchase

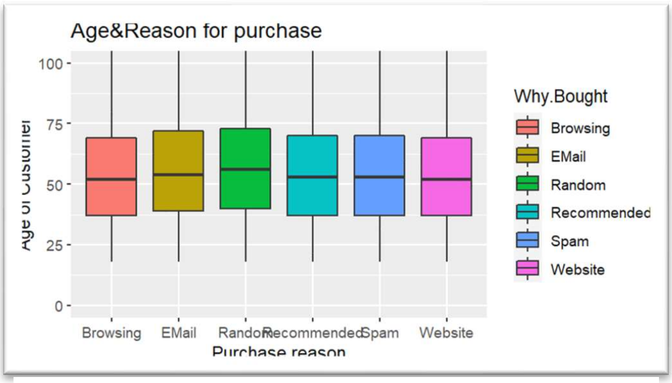


Figure 10: Age vs Why Bought

The boxplot shows that browsing is done by younger individuals while older people randomly found the online business.

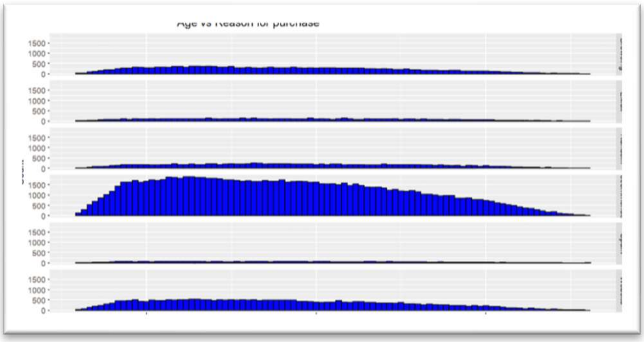


Figure 11: Age and Why bought boxplot

The different histograms for the reasons of purchased helped to support the boxplot above. The online business was recommended for most of the customers whereas spam was not the reason for purchase in many of the cases. All the graphs are skewed to the right

2.8 Process Capability calculations:

Technology:

USL= 24 hours

LSL = 0 hours

X = 19.76053 hours

σ = 3.50199 hours

Cp = (USL-LSL)/6*σ = 0.9439534

Cpu = (USL-X)/3*σ = 0.403529

Cpl =(X-LSL)/3*σ = 1.880884

Cpk =min (Cpu,Cpl) = 0.403529

The Cpk at the moment is too low. Higher values are always better and a value of 1 indicates marginally stable while 1.3 can be seen as stable. To improve the Cpk, we may either centre the process on the target with the adjustment of a dial or setting. Another option is to reduce the variation and spread of the process (Technology). This can be done by some fundamental redesign of the process.

Explain why an LSL of 0 is logical:

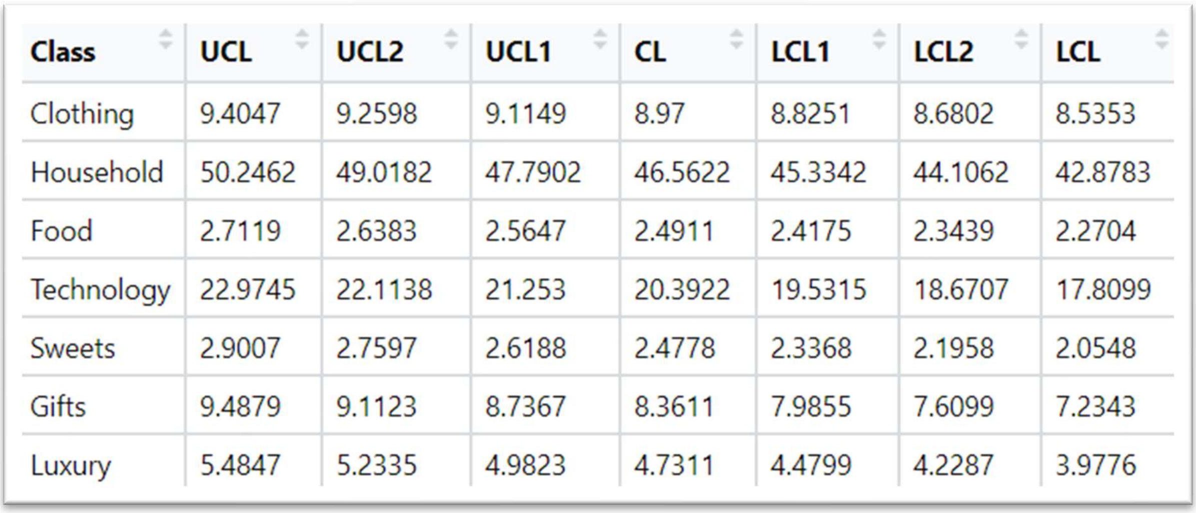
It is logical because the lower specification limit cannot be a negative number. The delivery time cannot be less than zero and that is why the lowest point is 0.

Part 3: Statistical Process Control (SPC)

3.1 First 30 Samples that will be encountered after ordering the data by year, month, day and X (row-index).

3.1.1 X-Chart limits:

The x chart limits are used to examine the process stability. The X chart indicates the changes of a process over the time in subgroup values.

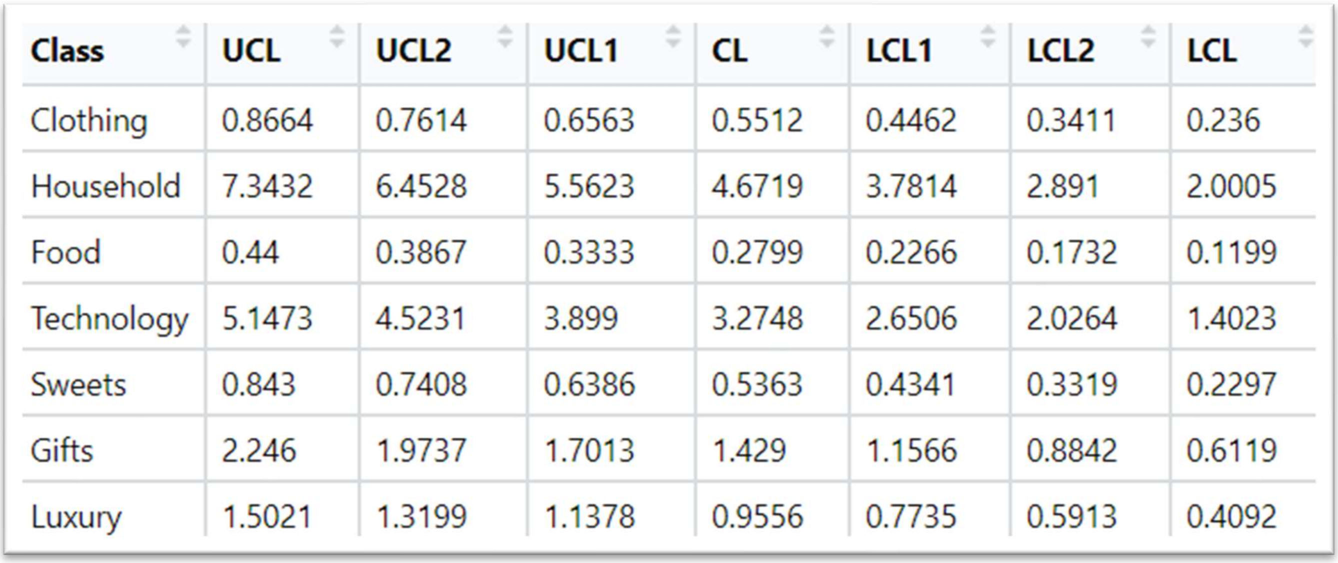


Class	UCL	UCL2	UCL1	CL	LCL1	LCL2	LCL
Clothing	9.4047	9.2598	9.1149	8.97	8.8251	8.6802	8.5353
Household	50.2462	49.0182	47.7902	46.5622	45.3342	44.1062	42.8783
Food	2.7119	2.6383	2.5647	2.4911	2.4175	2.3439	2.2704
Technology	22.9745	22.1138	21.253	20.3922	19.5315	18.6707	17.8099
Sweets	2.9007	2.7597	2.6188	2.4778	2.3368	2.1958	2.0548
Gifts	9.4879	9.1123	8.7367	8.3611	7.9855	7.6099	7.2343
Luxury	5.4847	5.2335	4.9823	4.7311	4.4799	4.2287	3.9776

Table 2: X-chart limits

3.1.2 S-Chart limits

S-charts are used to examine the standard deviation of the process stability. The standard deviation for each class will differ quite dramatically since they differ largely in delivery time.



Class	UCL	UCL2	UCL1	CL	LCL1	LCL2	LCL
Clothing	0.8664	0.7614	0.6563	0.5512	0.4462	0.3411	0.236
Household	7.3432	6.4528	5.5623	4.6719	3.7814	2.891	2.0005
Food	0.44	0.3867	0.3333	0.2799	0.2266	0.1732	0.1199
Technology	5.1473	4.5231	3.899	3.2748	2.6506	2.0264	1.4023
Sweets	0.843	0.7408	0.6386	0.5363	0.4341	0.3319	0.2297
Gifts	2.246	1.9737	1.7013	1.429	1.1566	0.8842	0.6119
Luxury	1.5021	1.3199	1.1378	0.9556	0.7735	0.5913	0.4092

Table 3: S-Chart limits

3.1.3 Control Charts

For the control charts the first thirty samples (subgroups) were used to estimate the mean as well as the standard deviation of the delivery process for every class. The mean and standard deviation is then used as control measures to calculate the limits for the various charts plotted underneath. The xBar SPC chart for most of the classes showed no real problems except technology which indicated two large deviations from the median. Clothing however remained consistently on the different classes are indicated as follow for the Xbar charts:

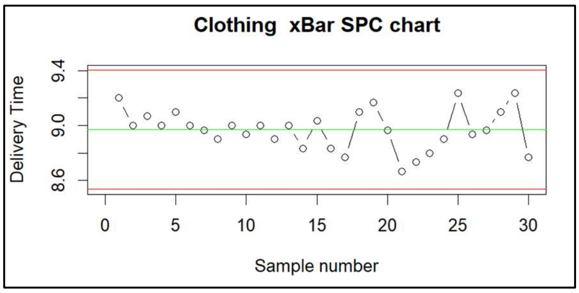


Figure 13: xBar chart- Clothing

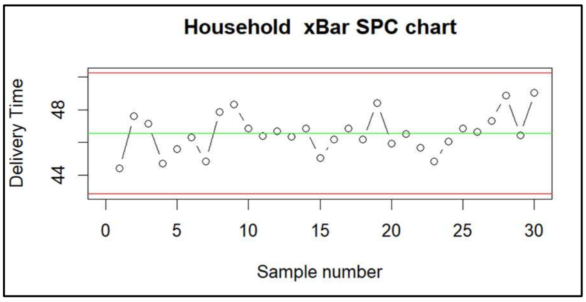


Figure 14:xBar chart-Household

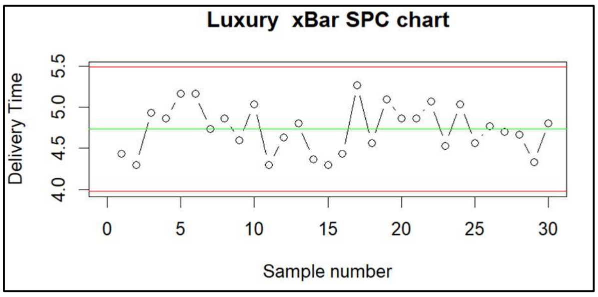


Figure 15: xBar chart-Luxury

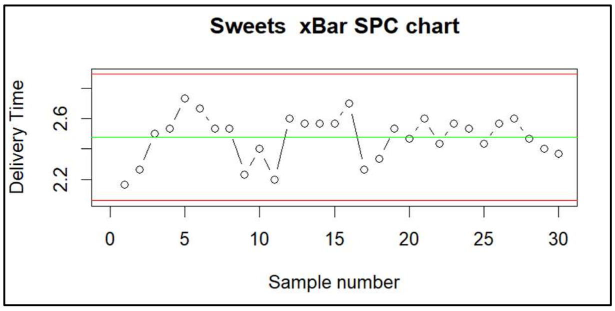


Figure 12: xBar Chart-Sweets

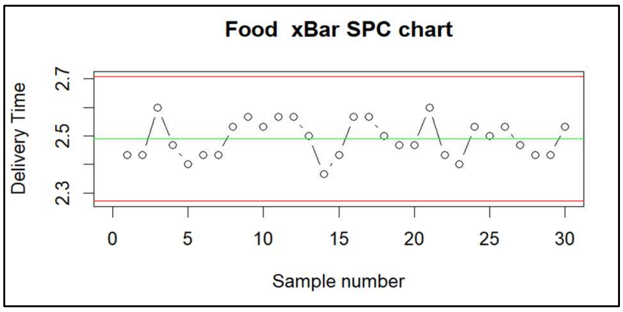


Figure 16: xbar Chart-Food

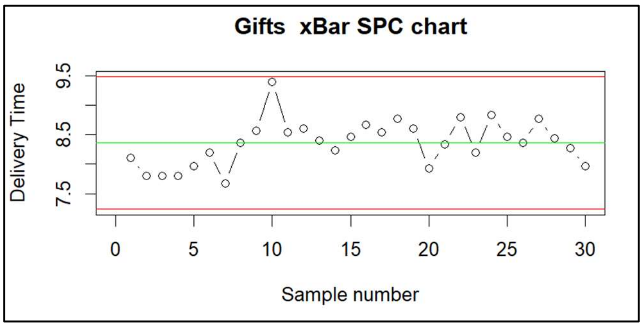


Figure 18: xBar chart-Gifts

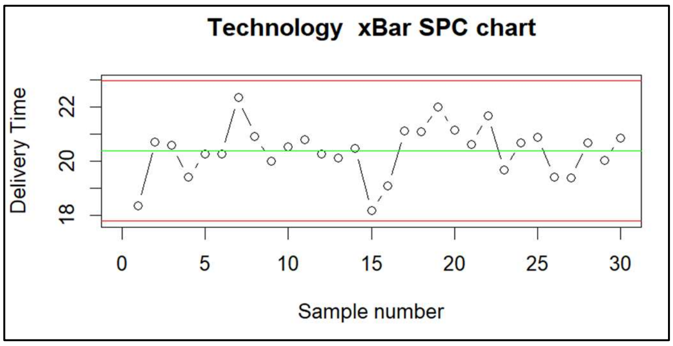


Figure 17:xBar Chart-Technology

3.2 Process Control

3.2.1 Out of control processes:

Looking at the gifts, household and luxury items there is a strong indication that the processes are out of control and if not adapted, will need to be removed. All three types show not control between the upper and lower bounds and deviates on a large scale from the mean. Both the X-chart and S-chart below shows the deviation from the mean for each type of item.

Gifts:

X chart:

S chart:

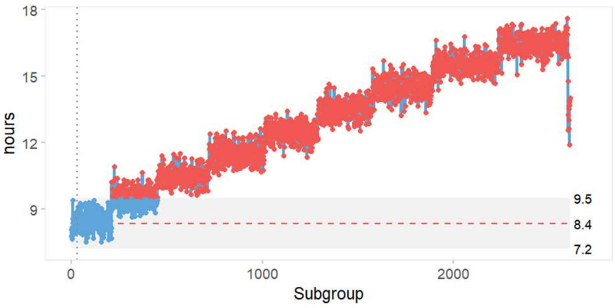


Figure 20: X Chart- Gifts

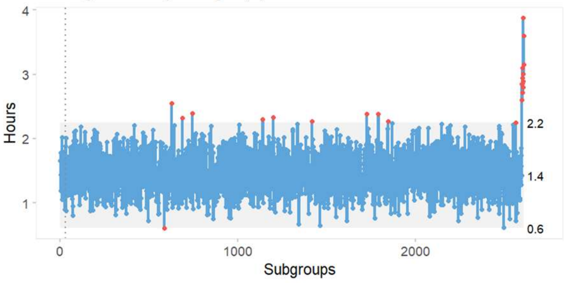


Figure 19: S Chart-Gifts

Gifts deviate far away from the median and there is no control in this process.

Household:

X Chart

S Chart

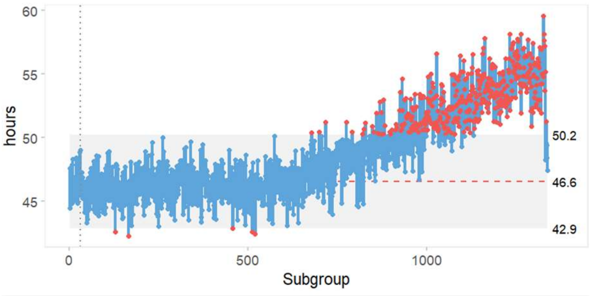


Figure 22: X Chart-Household

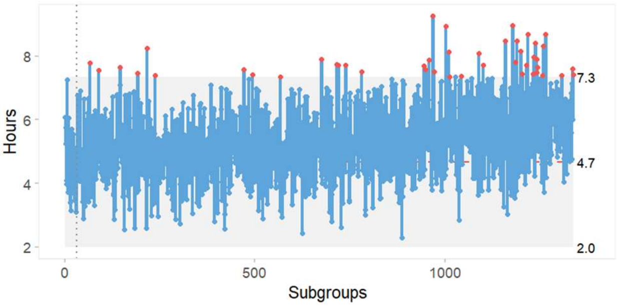


Figure 21: S Chart-Household

Household items started under control but after 750 subgroups (or samples) started to get out of control and move out of the upper limit of the controlled region.

Luxury:

X Chart

S Chart

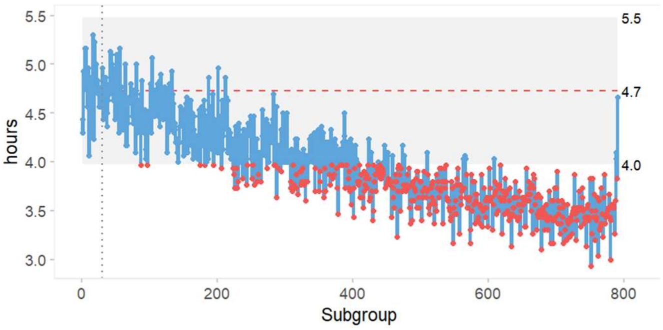


Figure 24: X Chart- Luxury

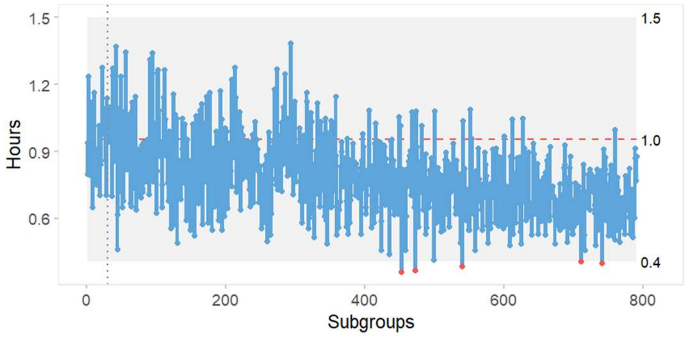


Figure 23: S Chart-Luxury

Luxury also shows no control and moves from the upper level at 4.7 hours to under the lower level of 4 hours. There is no control over the class.

3.2.2 In control processes:

The processes (Technology, Food, Sweets and clothing) are almost 100% stable with a few instances that are out of the indicated control limits. Nonetheless, the overall results have indicated that these 4 descriptive features are in control with very little outliers. The subgroups(samples) are on the x-axis and the delivery time in hours on the y-axis.

Technology:

X chart:

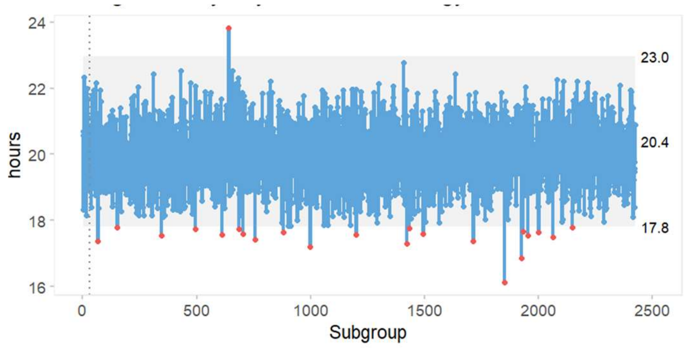


Figure 25: X Chart- Technology

S chart

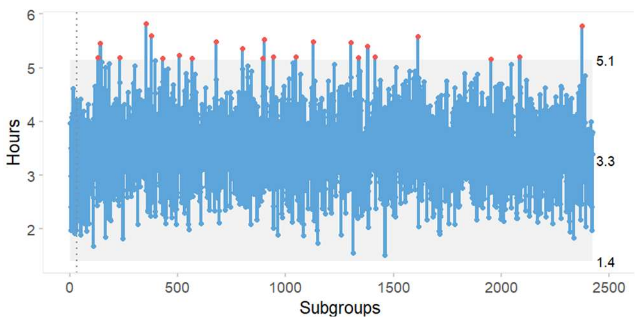


Figure 26: S Chart-Technology

From all the controlled process, technology has the most outliers (23) but remains in-between the upper limits and lower limits that is 17.8 and 23 hours.

Food:

X chart:

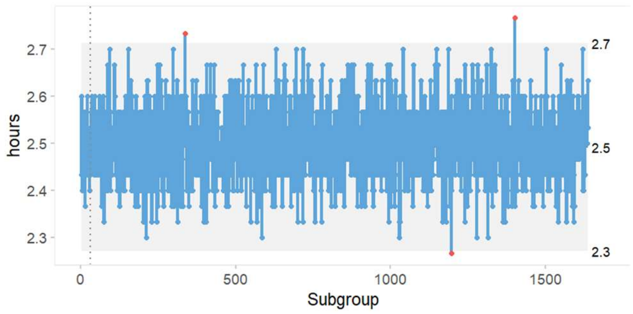


Figure 28: X Chart-Food

S chart:

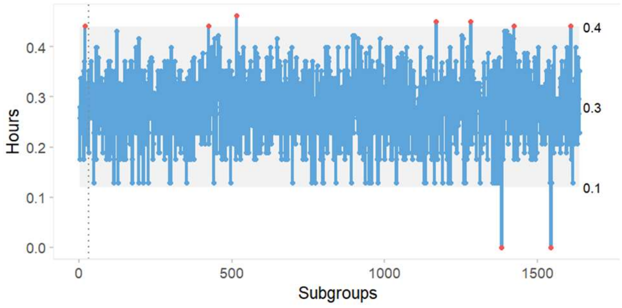


Figure 27: S Chart-Food

Food has minimum outliers for both charts. The X chart has 3 with the S chart at 9 outliers for all the subgroups (samples).

Sweets:

X chart:

S chart:

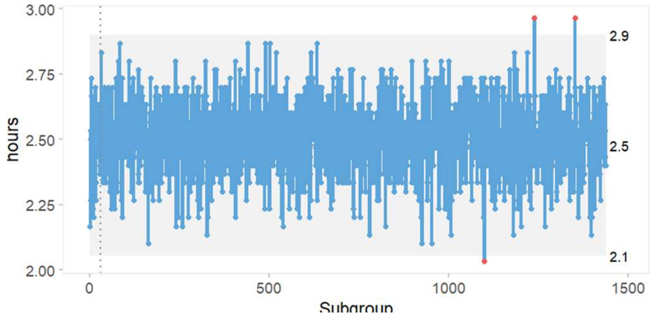


Figure 30: X Chart-Sweets

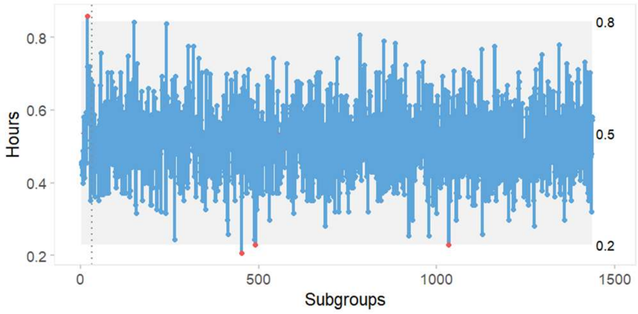


Figure 29: S Chart-Sweets

Clothing:

X chart

S chart

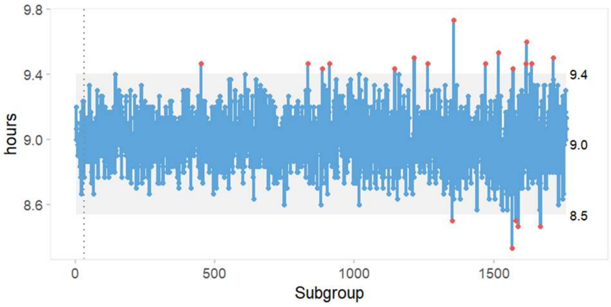


Figure 32: X Chart-Clothing

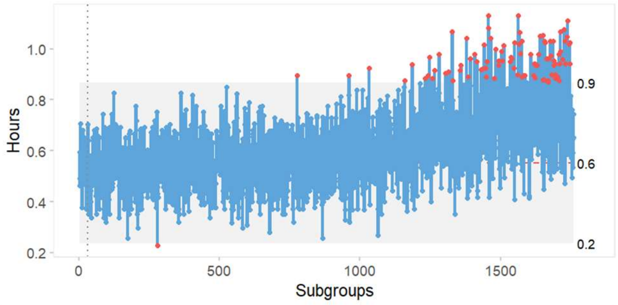
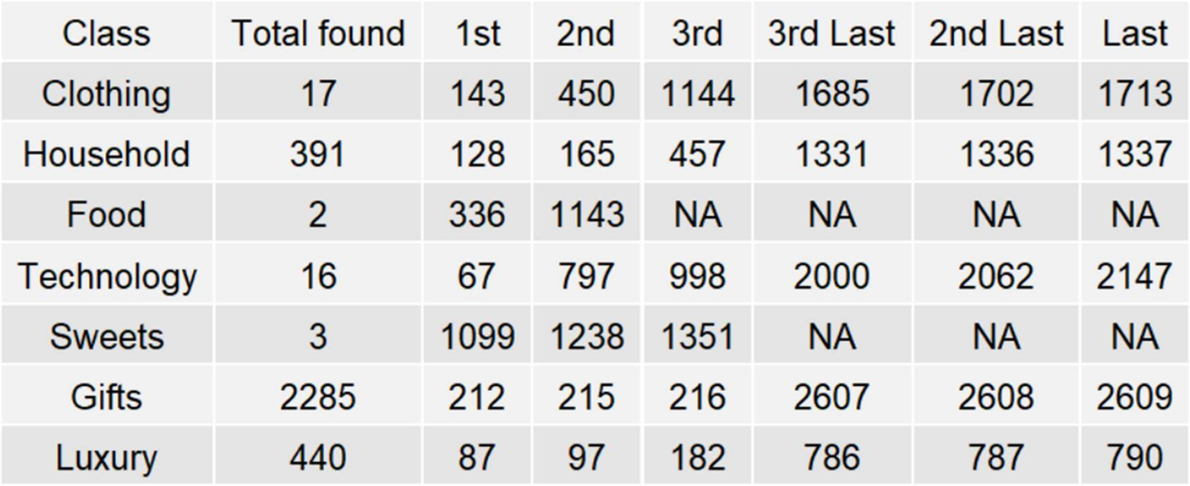


Figure 31: S Chart-Clothing

Part 4: Optimising the delivery processes

4.1 the Sample numbers that gave indications of out of control, for X-s charts using the rules:

4.1.A First and last three outliers



Class	Total found	1st	2nd	3rd	3rd Last	2nd Last	Last
Clothing	17	143	450	1144	1685	1702	1713
Household	391	128	165	457	1331	1336	1337
Food	2	336	1143	NA	NA	NA	NA
Technology	16	67	797	998	2000	2062	2147
Sweets	3	1099	1238	1351	NA	NA	NA
Gifts	2285	212	215	216	2607	2608	2609
Luxury	440	87	97	182	786	787	790

Table 4: X-bar outside outer control limits

As shown in the X and S charts the clear classes with a significant number of outliers were household, gifts and luxury products. The gifts class have a worrying number of outliers that need to be considered when reviewing the specific classes during a business meeting. Outliers show that the delivery times are not consistent for some of the classes and that there can be no specific conclusions made on the relationship between the classes and delivery times. The food and sweets have very little outliers which was indicated in part 3. They are very much in control process and fall in the UCL and LCL range.

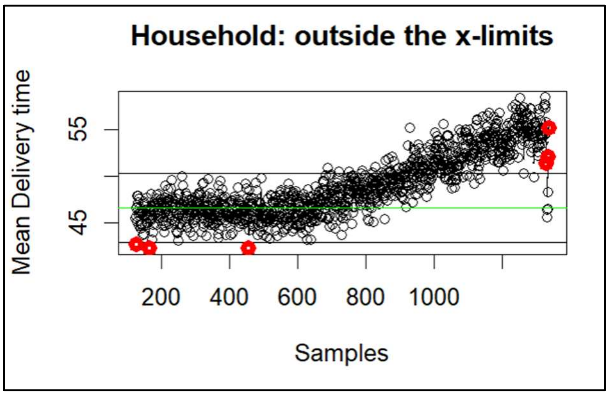


Figure 33: Outside x limits-Household

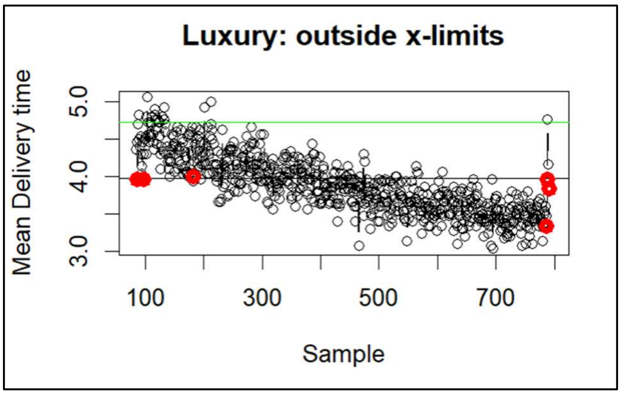
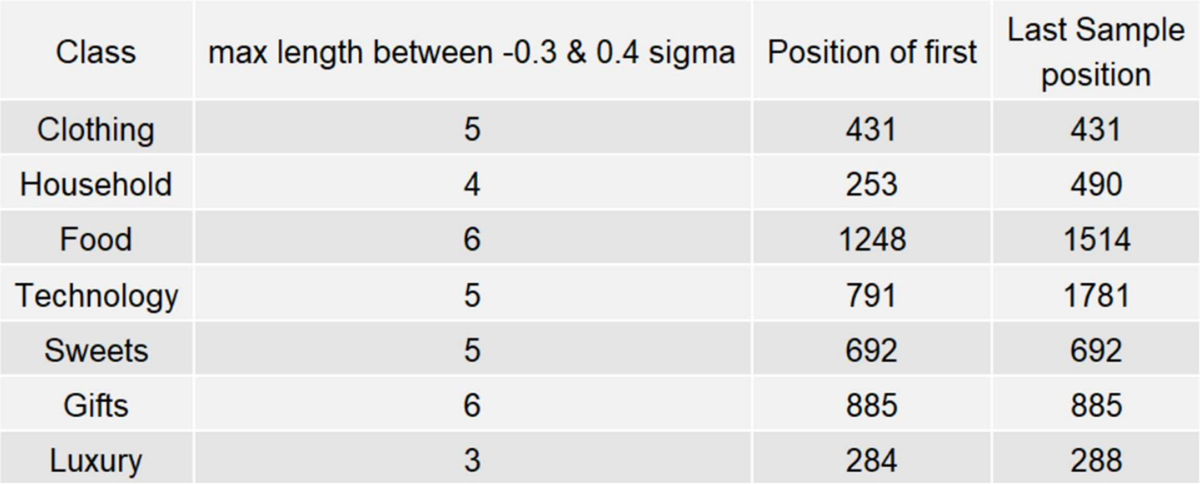


Figure 34: Outside x limits-Luxury

For Household the first 3 outliers are located before 500 samples with 2 of those samples early in the process and before 200 samples. For Luxury the first 3 outliers are located before 200 samples with 2 of those samples early in the process. The last 3 outliers are located close to sample 790.

4.1.B Most consecutive samples



Class	max length between -0.3 & 0.4 sigma	Position of first	Last Sample position
Clothing	5	431	431
Household	4	253	490
Food	6	1248	1514
Technology	5	791	1781
Sweets	5	692	692
Gifts	6	885	885
Luxury	3	284	288

Table 5: S-bar between -0.3 and +0.4 sigma control limits

When searching for the most consecutive samples of the sample standard deviations between -0.3 and +0.4 the following results were obtained. Gifts and food have the most consecutive occurrences of samples between the two ranges. The Luxury have the smallest length which is 3 ranging from 284 to 288. Food is one of the classes with the largest consecutive instances inside the two sigma levels ranging from 1248 to 1514.

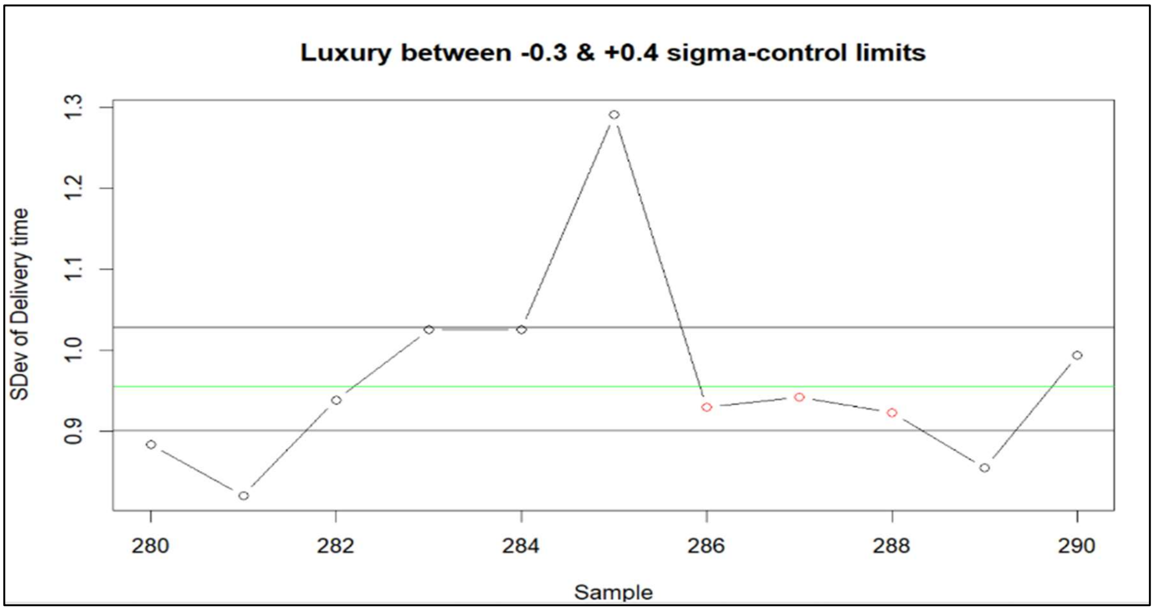
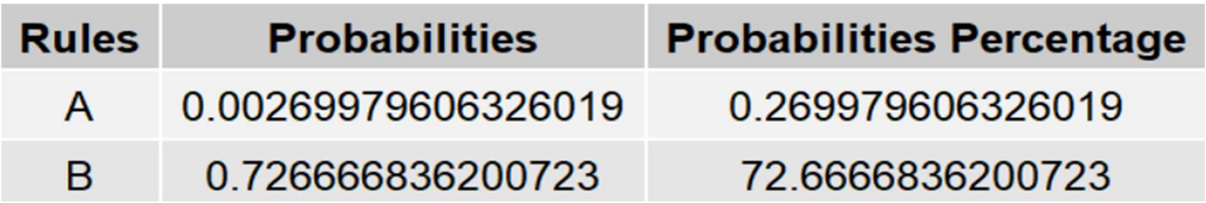


Figure 35: Luxury item max length

The 3 three instances for luxury moving between the UCL and LCL were found to be in the region between 280 and 290. This is the maximum length for the consecutive instances that fall between -0.3 and +0.4.

4.2 likelihood of making a Type I (Manufacturer's) Error for A and B



Rules	Probabilities	Probabilities Percentage
A	0.00269979606326019	0.269979606326019
B	0.726666836200723	72.6666836200723

Table 6: Rule A and B for Type 1 error

For rule A:

The rule for A remains consistent for any process and is not related to the dataset. The 0.2699% shows the chance of making the type 1 error is indeed very small. This is a good sign for any business because they think their process is not under control but in hindsight it is under control.

For rule B:

The values used is for the 2 sigma levels which is +0.4 and -0.3. The

4.3

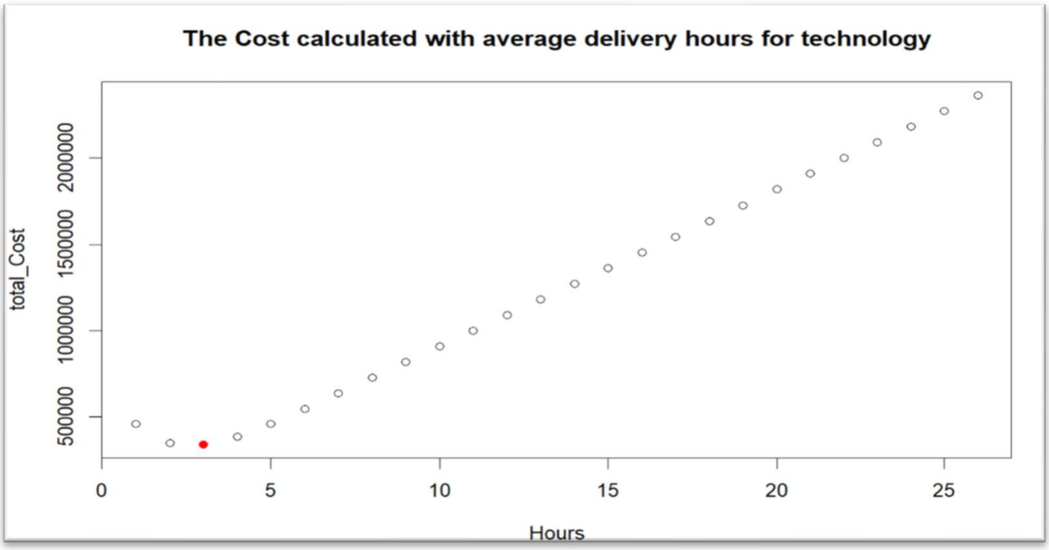


Figure 36: Cost vs Delivery time

As shown above the lowest cost for the delivery hours of the technology class was calculated. The hour on the x-axis indicates that the best cost was at the 3-hour mark. On the y-axis the cost is shown for the different hours from 0 to 25 hours. The goal is to have the least amount of hours of delivery time since it will reduce the cost to the online business and thus help the profitability. The best cost at the 3-hour mark was R 340877.5.

4.4 likelihood of making a type II (Consumer's) Error for A

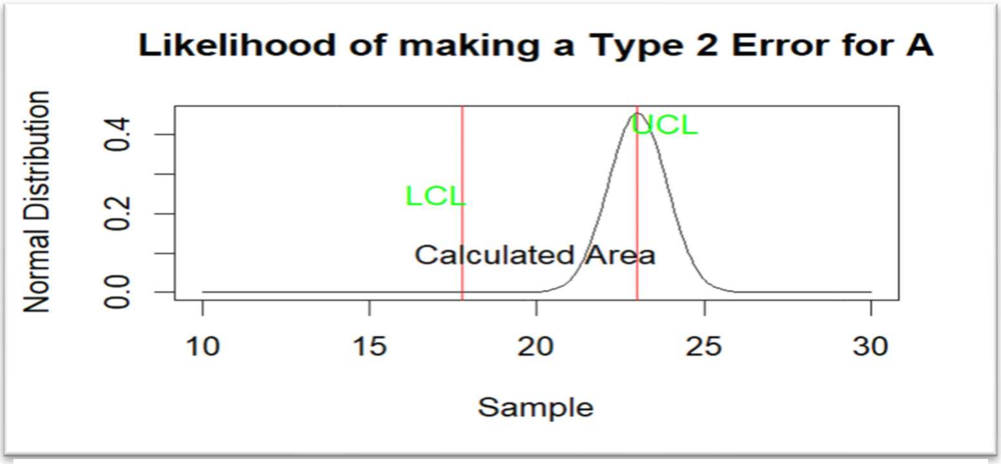


Figure 37: Likelihood of making type 2 error for A

The likelihood of making a type 2 error for technology is between the two levels indicated above. The normal distribution is plotted on the y-axis and the sample over the x-axis. The UCL has a value of 22.99129 and the LCL is 17.7468. The calculated area between the two levels will show the likelihood of making a type 2 error for A. The value of the calculated area is 49.6025% which is very high for the company under the circumstances. This is not a positive sign since this indicates the chance of thinking the process is under control when it is the complete opposite at almost 50% chance of error. If this mistake is made by the business, it can lead to additional costs.

Part 5: DOE and MANOVA

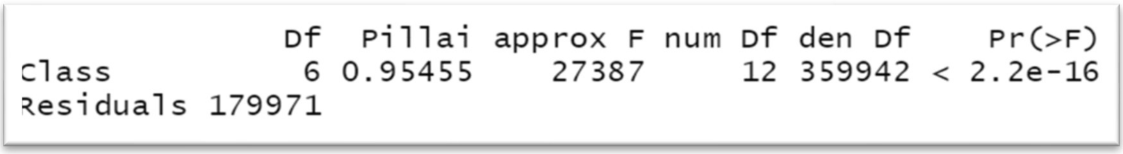
The MANOVA (Multivariate ANOVA/ Multivariate Analysis of Variance) is used to analyze the difference between more than one group of independent variables (Radečić, 2022) . For the online business the various hypotheses will be tested. For the two hypothesis tests the p-value of 0.05 was used. It is the most well-known and popular value it was decided that it can give a good indication if the test is valid or not.

MANOVA 1:

For the first hypothesis test:

H0: The reason for purchase and the type of class **depends** on the age of the client

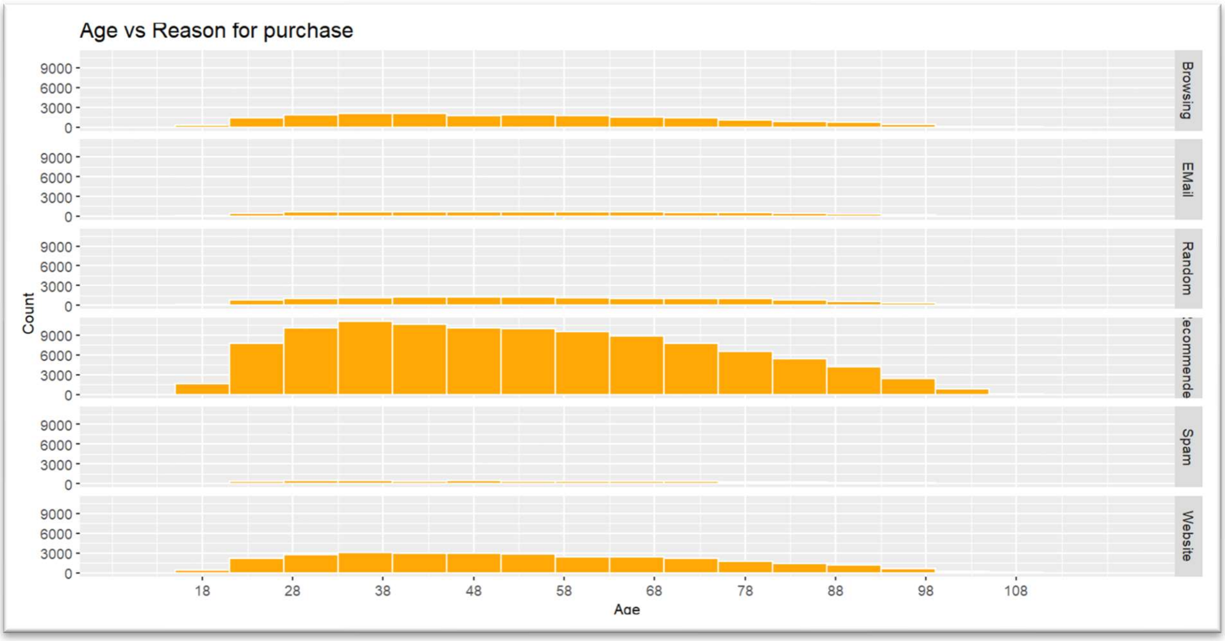
H1: The reason for purchase and the type of class **does not depend** on the age of the client



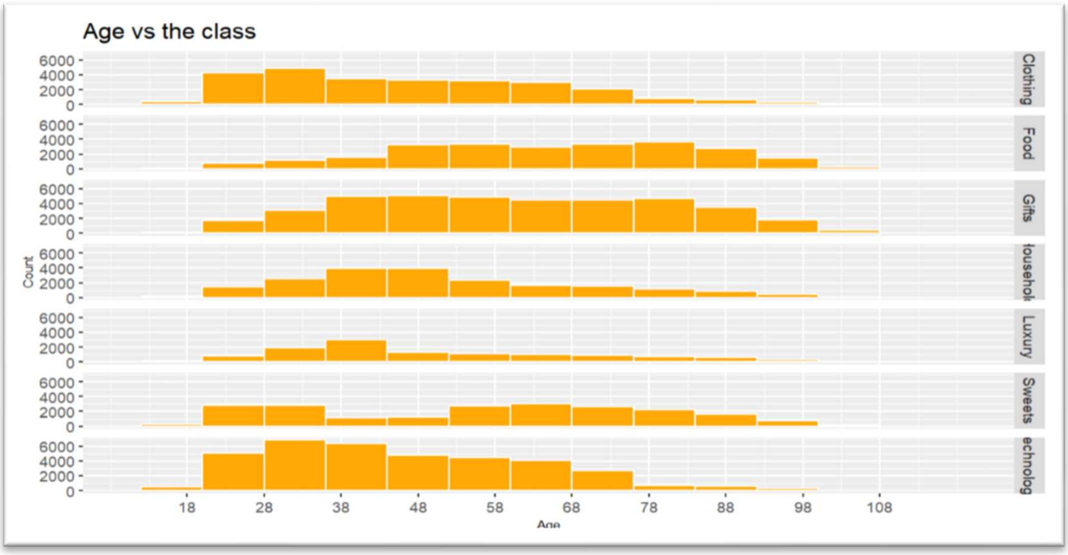
```
              Df  Pillai approx F num Df den Df    Pr(>F)
Class          6 0.95455    27387     12 359942 < 2.2e-16
Residuals 179971
```

Figure 38: MANOVA 1

The H0 is accepted since 0.05>2.2e-16. This shows that the type of class and the reason for purchase is dependent on the age of the client.



The age of the client and the reason for purchase has a clear relationship as shown in part 2 of the project. The older the client is the more inclined they are to react to recommendations and through browsing rather than by receiving email.



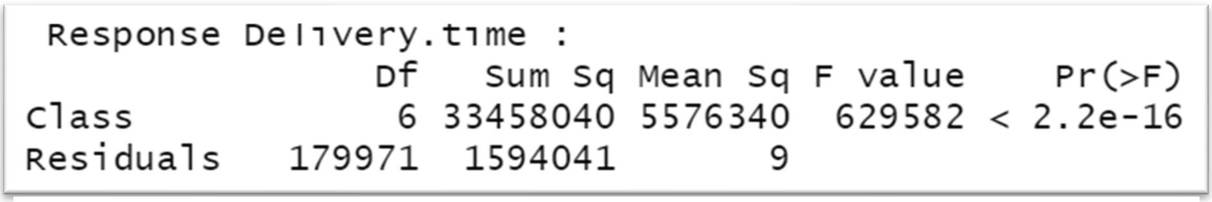
The age distribution shows also a clear relationship between the different classes and the age. Younger customers are more likely to purchase technology items while older ages use their money to buy basic needs such as food or gifts.

MANOVA 2:

For the second hypothesis test:

H0: The delivery time and the month of purchase **depends** on the different classes.

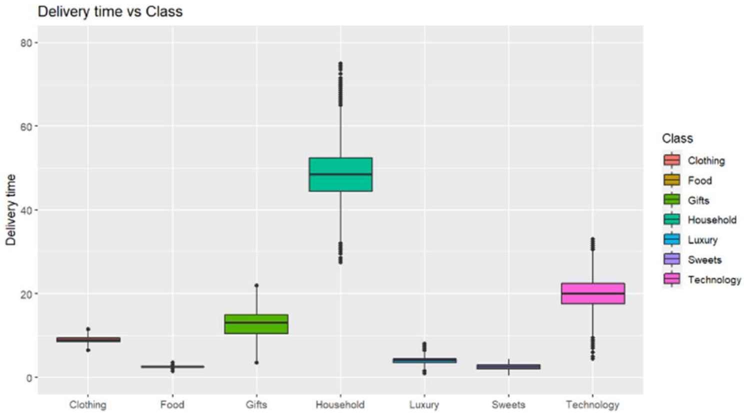
H1: The delivery time and the month of purchase **does not depend** on the different classes.



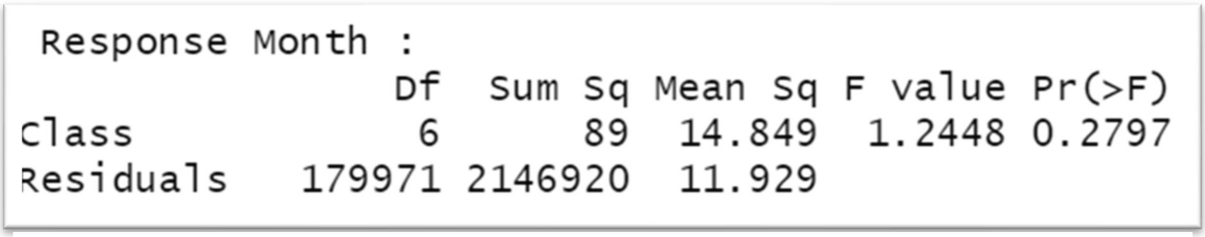
```
 Response Delivery.time :
                Df   Sum Sq Mean Sq F value    Pr(>F)
Class            6 33458040 5576340  629582 < 2.2e-16
Residuals   179971  1594041       9
```

Figure 39: MANOVA 2.1

The response for the delivery time shows that 0.05>2.2e-16 and that the relationship between the delivery time and the type of class is strong. The H0 is accepted and the delivery time is influenced by the type of class.



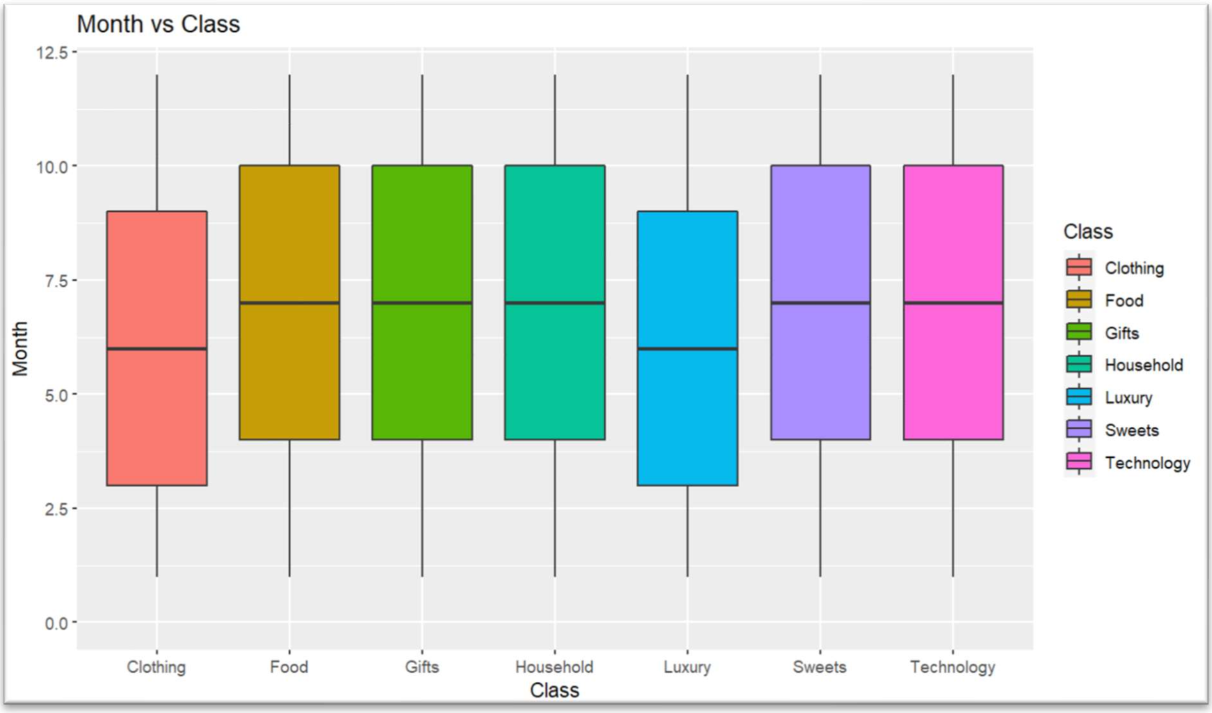
The different classes will have different delivery times as shown on the boxplot on the left. The delivery time is very specific for each class type. The conclusion is that H0 can be accepted.



```
 Response Month :
                Df  Sum Sq Mean Sq F value Pr(>F)
Class            6      89  14.849  1.2448 0.2797
Residuals   179971 2146920  11.929
```

Figure 40:MANOVA 2.2

For the month and the different classes there is not the same strong relationship that will help the online business to find a clear correlation. Although 0.05<0.2797, the difference is not significantly small enough to show that they are strongly linked to each other. The H0 is therefore not accepted since the null hypothesis is not true for this relationship.



The months from 1-12 are displaying almost identical distribution of the boxplots for every class. Only luxury and clothing have a lower median and first quartile.

Part 6: Reliability of services and products

6.1

For both the problems the Taguchi loss function was used. K is a constant, L is the loss, m is the theoretical target value and y is the actual size of the product.

Problem 6:

Information:

Specification for thickness: 0.06+-0.04

Costs $45 to scrap part

The function used: L(y) = k*(y-m) ^2

Where k = L(y)/ (y-m) ^2

Thus k = 45/ (0.04) ^2 = **28125**

Now Taguchi loss function: **L(y) = 28125*(y-m) ^2**

Problem 7:

Information:

$35 to scrap part

a) L(y) = k*(y-m) ^2

Where k = L(y)/ (y-m) ^2

Thus k = 35/ (0.04) ^2 = **21875**

Now Taguchi loss function: **L(y) = 21875*(y-m) ^2**

b)

Process deviation from the target is reduced to 0.027cm.

Thus, the Taguchi loss function:

21875*(0.027) ^2 = $15.94687

The cost is $15.947

6.2

Problem 27

a) **Reliability of machines:**

Series of machines: Ra*Rb*Rc

0.85*0.92*0.9 = **0.7038 / 70.38% reliability**

There is still a high level of unreliability when using only one machine at each stage.

b) Reliability of 2 machines:

Parallel: (1-(1-Ra) ^2) *(1-(1-Rb) ^2) *(1-(1-Rc) ^2

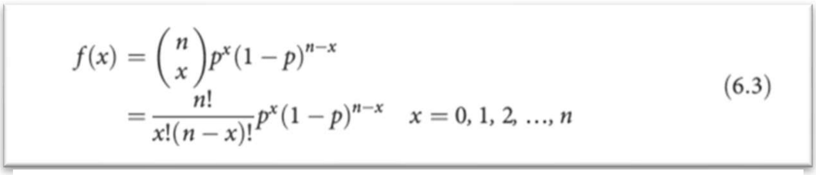
(1-(1-0.85) ^2) *(1-(1-0.92) ^2) *(1-(1-0.9) ^2) = **0.9615316 / 96.153% reliability**

There is a (96.15-70.38) 25.77% difference in the reliability when the machines are placed in parallel to each other rather than using one machine. The reliability has increased enough to ensure better certainty for the process when moving from machine A through to C. It is recommended to rather use these machines in parallel when making reliability the top priority.

6.3

Information:

The goal is to predict how many days of reliable delivery there will be in 1 year. The different probabilities for the vehicle and driver reliability will be calculated separately. Afterwards the two factors for reliability will be combined to find a clear prediction for the delivery process. The binomial distribution formula will be used and is indicated below.



$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad x = 0, 1, 2, \ldots, n \qquad (6.3)$$

Figure 41:Binomial distribution formula

Vehicle:

For the delivery process:

- ✓ Always 20 vehicles available for delivery.
- ✓ 19 required to be operating at any time.
- ✓ In the past 1560 days:
    - ➢ 20 vehicles available for 190 days
    - ➢ 19 vehicles available for 22 days
    - ➢ 18 vehicles available for 3 days
    - ➢ 17 vehicles available for 1 day.

The number of days that vehicles were available were calculated using the above-mentioned binomial distribution.

The overall number of days out of the 1560 that 21 vehicles are available is 1344.311 or roughly 1345 days. The number of days only 20 vehicles were available is 200.7493 or rounded up to 201 days. The vehicle availability during the last 1560 days were then used to calculate how many days in a single year we should expect reliable delivery times.

The number of days 20 and 21 vehicles were available was used and divided by the number of days which is 1560. Then the reliability which is 99.0423% is multiplied with 365 days (one year) and the answer is 361.5 or roughly 362 days of reliable delivery times. For 21 vehicles, the availability in one year is 314.54 or 315 half completed days. For 20 vehicles, the available days in one year is 46.97 or 47 days. This shows that moving from 20 to 21 vehicles is much more reliable.

Driver:

The same procedure for the drivers were followed.

For the delivery process:

- ✓ In the past 1560 days:
  - ➢ 20 drivers available for 95 days
  - ➢ 19 drivers available for 6 days
  - ➢ 18 drivers available for 1 day

The number of days that drivers were available are calculated using the above-mentioned binomial distribution.

The overall number of days out of the 1560 that 21 drivers are available is 1457.72 or roughly 1458 days. The number of days only 20 drivers were available is 99.01 or rounded down to roughly 99 days. The driver availability during the last 1560 days was then used to calculate how many days in a single year we should expect reliable delivery times.

The number of days 20 and 21 drivers were available, was used and divided by the number of days which is 1560 to get a probability(P-value). Then the P-value (reliability which was calculated to be 99.79%) is multiplied with 365 days (one year). The final answer for the drivers is 364.2352 or roughly 365 days of reliable delivery times. Thus, for only 21 drivers, the available days is 341.0693 or rounded down 341 complete days while the availability for 20 drivers is only 23.16582 or 23 complete days. There is definitely an increase in the availability of drivers when 21 is used instead of the 20 previous drivers.

Final reliability:

The overall reliability of the delivery times will depend on both the driver and the vehicles availability. Thus, the reliability of drivers available is higher but the overall number of days depend on the last number both are available and that is the two variables divided by 365^2. The number of days is thus 360.747 or roughly 361 days of reliable delivery time for both the vehicle and driver.

Conclusion

After filtering the client data for the online business and removing the irrelevant instances, a thorough review was conducted. This review looked at the relationship between the different descriptive features and what clear points stood out that will help to better predict the delivery times. For part 3, the SPC and X and s-charts were used to gain a better understanding of the different classes. This was indicated by plotting the different charts. Part 4 was used to optimize the delivery process. Part 5 made use of the MANOVA to do the 2 appropriate hypothesis tests. Both tests were accepted and discussed after the results were found. In part 6 the various word problems were solved, and a clear solution was found for problem 6,7 and 28.

References

Hernandez, F., 2015. *Data Analysis with R - Exercises.* [Online]
Available at: http://fch808.github.io/Data-Analysis-with-R-Exercises.html
[Accessed 10 October 2022].

Radečić, D., 2022. *MANOVA in R – How To Implement and Interpret One-Way MANOVA.* [Online]
Available at: https://www.r-bloggers.com/2022/01/manova-in-r-how-to-implement-and-interpret-one-way-manova/#google_vignette
[Accessed 17 October 2022].

SolarWinds, 2020. *What Is Cardinality in a Database?.* [Online]
Available at: https://orangematter.solarwinds.com/2020/01/05/what-is-cardinality-in-a-database/
[Accessed 14 October 2022].

STHDA, n.d. *MANOVA Test in R: Multivariate Analysis of Variance.* [Online]
Available at: http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance
[Accessed 9 October 2022].