QA ECSA Project

Kian Luke Sheppard

2022-10-07

Contents

List of Figures

1. Introduction

This report focuses on analyzing the data gained from a sales company. Having the correct business understanding to the problems given, analyzing/exploring the data by addressing data quality issues and using the correct data exploration techniques, and finally preparing this data in order to display the necessary features. Tasks such as the optimizing the delivery times, finding relationships through techniques such as Doe and Manova and the calculations of the reliability of the process could be achieved once all the data had been cleaned through data wrangling and statistical analysis.

2. Part 1: Data Wrangling

To avoid unnecessary future issues, data exploration is a crucial, initial step before diving into any type of data set. It is difficult to see at an initial glance that a data set has defects such as missing information, or incorrect data. The biggest mistake a company can make would be to make use of incorrect data - which in turn, returns incorrect predictions. Ultimately, the aim was to optimize the data quality issues like the missing and incorrect data to achieve the ideal final data transformation seen. We were given a data set with 18000 instances. After cleaning the data, it was clear that 22 instances had to be removed. Of these 22 instances, 17 were 'Not applicable' values and 5 were negative.

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

Figure 1: Summary of invalid data

After removing all the invalid data, 17978 instances remained, which was now known as our valid data. This data would then be used for further analysis of the project.

```
       Y                X                 ID               AGE              Class                Price                Year           Month             Day
 Min.   :     1   Min.   :     1   Min.   :11126   Min.   : 18.00   Length:179978      Min.   :    35.65   Min.   :2021   Min.   : 1.000   Min.   : 1.00
 1st Qu.: 44995   1st Qu.: 45004   1st Qu.:32700   1st Qu.: 38.00   Class :character   1st Qu.:   482.31   1st Qu.:2022   1st Qu.: 4.000   1st Qu.: 8.00
 Median : 89990   Median : 90005   Median :55081   Median : 53.00   Mode  :character   Median :  2259.63   Median :2025   Median : 7.000   Median :16.00
 Mean   : 89990   Mean   : 90003   Mean   :55235   Mean   : 54.57                      Mean   : 12294.10   Mean   :2025   Mean   : 6.521   Mean   :15.54
 3rd Qu.:134984   3rd Qu.:135000   3rd Qu.:77637   3rd Qu.: 70.00                      3rd Qu.: 15270.97   3rd Qu.:2027   3rd Qu.:10.000   3rd Qu.:23.00
 Max.   :179978   Max.   :180000   Max.   :99992   Max.   :108.00                      Max.   :116618.97   Max.   :2029   Max.   :12.000   Max.   :30.00
 Delivery.time    Why.Bought
 Min.   : 0.5   Length:179978
 1st Qu.: 3.0   Class :character
 Median :10.0   Mode  :character
 Mean   :14.5
 3rd Qu.:18.5
 Max.   :75.0
```

Figure 2: summary of the valid data

3. Part 2: Statistical Analysis

Process Capabilities:

Process Capability may be defined as the ability of a process to meet specifications. The Process Capability indices: Cp, Cpu, Cpl and Cpk for the process delivery times of technology class items were calculated. We assumed the USL = 24 hours and the LSL = 0. This LSL value is quite logical as the minimum days for delivery is 0.

Standard deviation = 3.501993

Mean = 20.01095

CP = (USL- LSL)/6σ = 1.142207

CPU = (USL - μ)/3σ = 0.3796933

CPL = (μ – LSL)/3σ = 1.90472

CPK = min(CPL, CPU) = 0.3796933

Price:

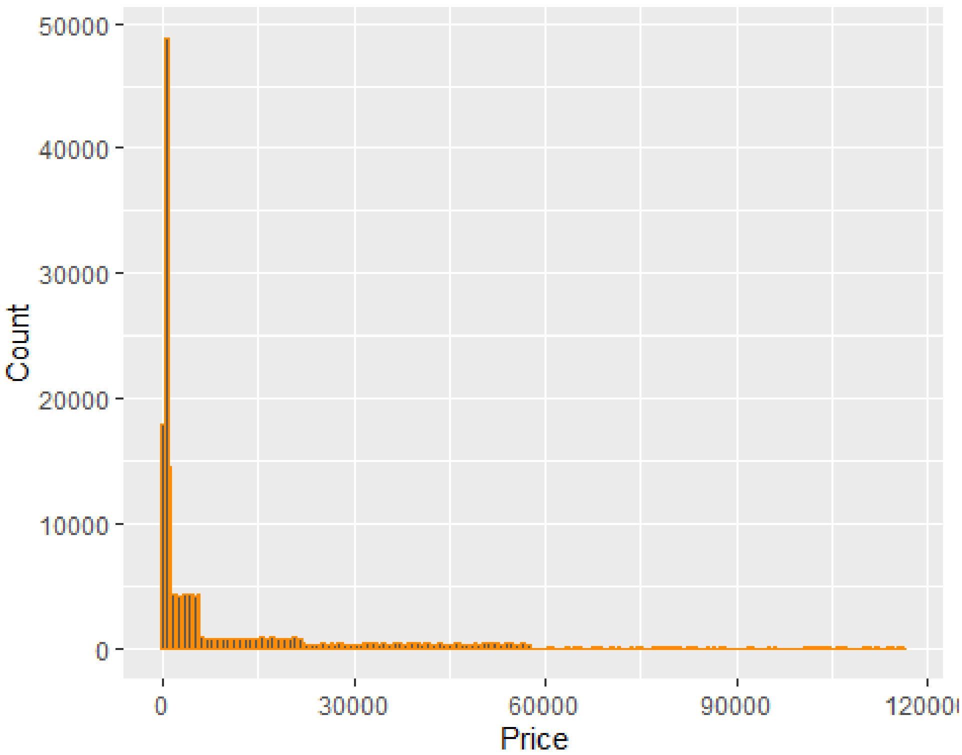

Figure 3: Price vs Count

By looking at the graph above it is clear to see that most of the class's pricing is relatively low. The data is skewed to the right with many outliers occurring past the price of 55000. The mean price is R12294,10.

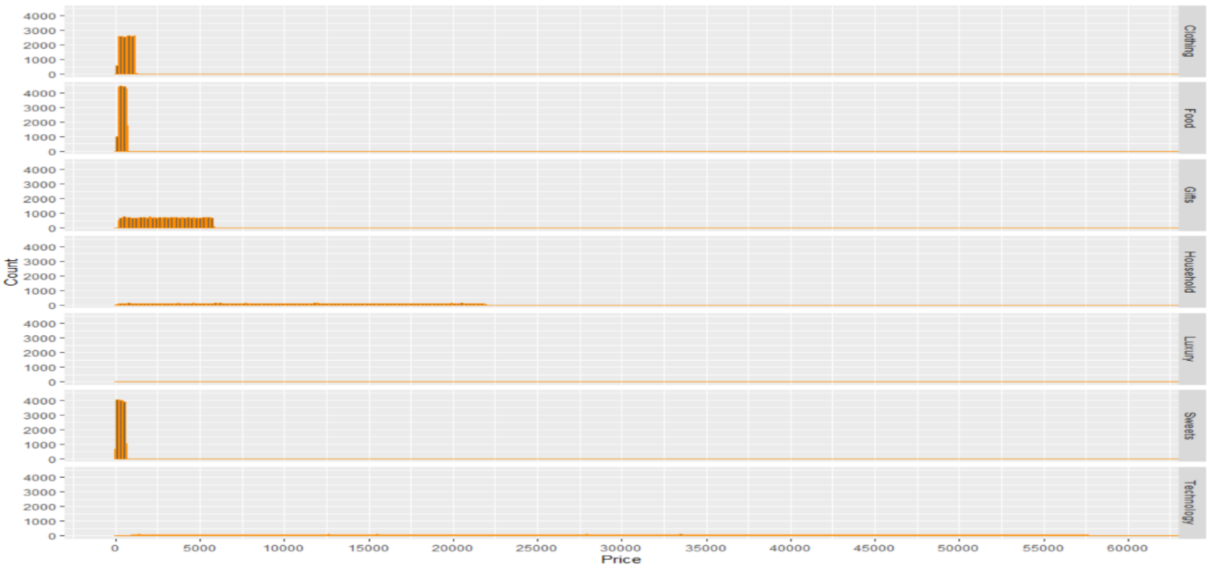

Figure 4: Individual classes Price vs Count

Technology and household are the two classes with the highest variety when pricing is considered. Their data is spread over a larger area than the other classes. The highest price is R116619, and it is under the Luxury class. Clothing, food and sweets tend to have relatively low prices. The lowest price is R35,65 and is found under the sweets class.

Age:

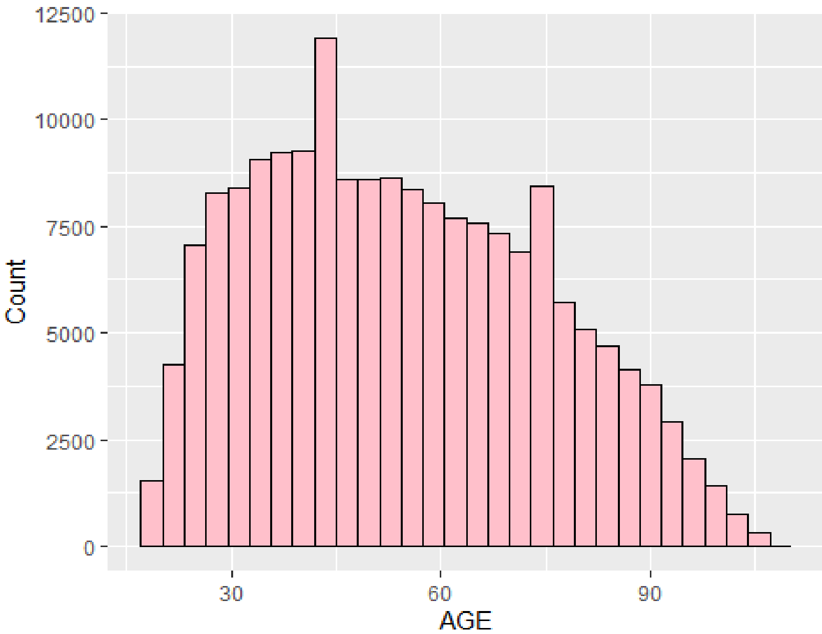

Figure 5: Age vs Count

The skewed to the shape of the graph indicates a middle-aged customer target market. Ages ranging from 30 to 60. The mean age is 54.56564 years with a max of 108 years and a minimum age of 18 years.

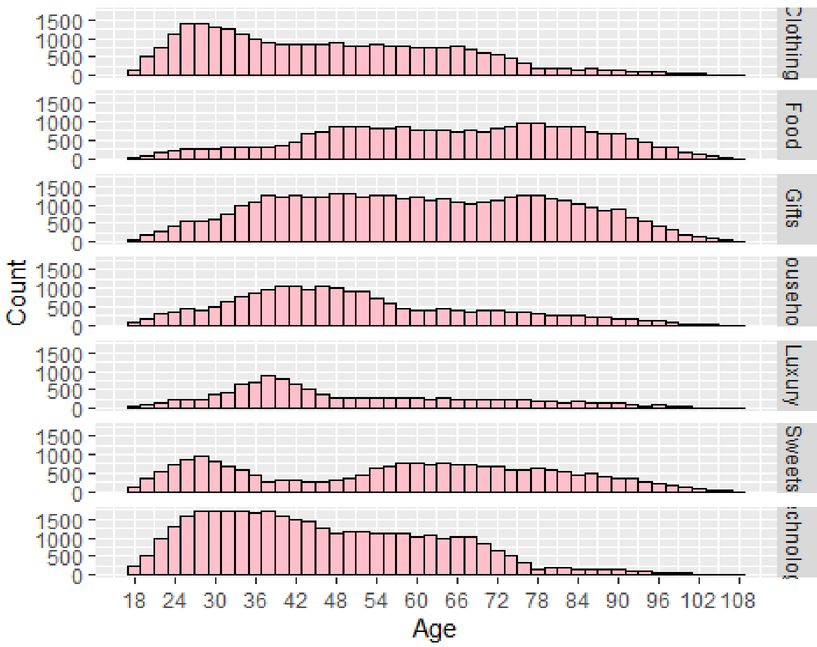

Figure 6:Individual age vs count

More clear indication of target markets for the different classes. With a younger age group purchasing Clothing, Household, Luxury and Technology. An older population purchasing Food and Gifts. Sweets are seen to be purchased by a younger and older population, with ages ranging from 36 to 54 not regularly purchasing sweets.

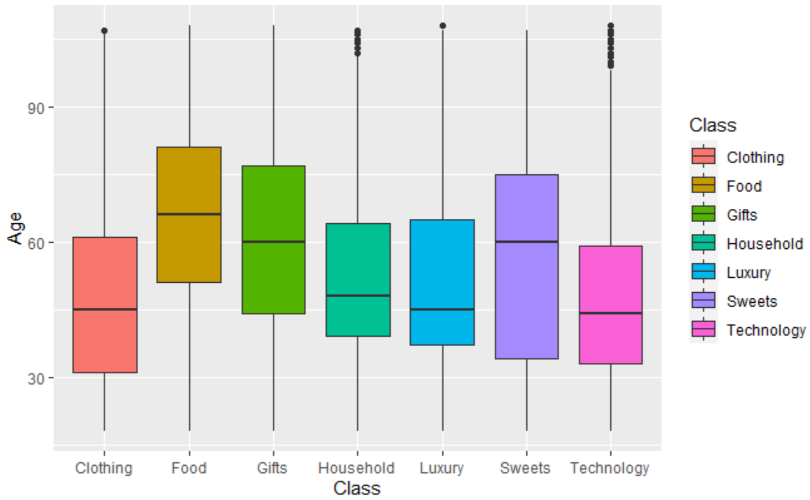

Figure 7: Boxplot Age vs Class

Boxplot of different ages purchasing respective classes. This is of course shown above in the form of a histogram; however, the boxplot gives a different angle to view the differentiation. Also indicates outliers represented by black dots in a few of the classes such as clothing, household, luxury and technology.

Class:

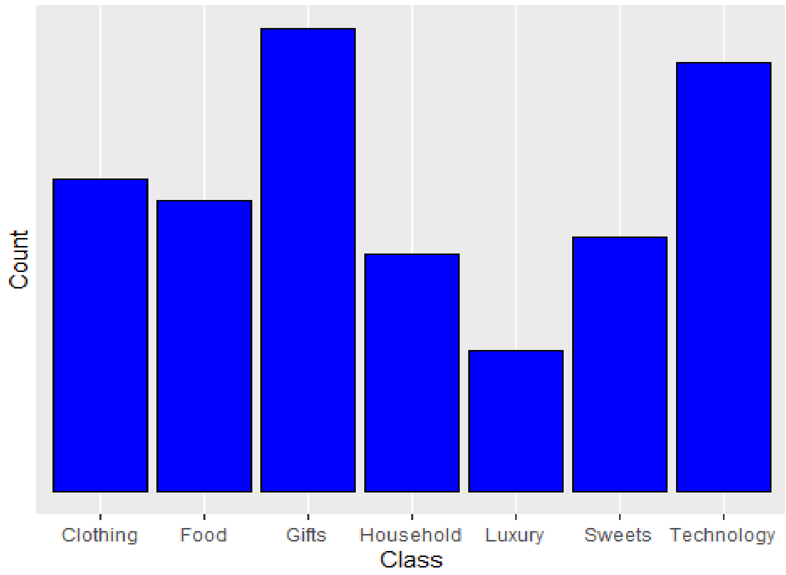

Figure 8:Class vs Count

Technology and gifts are the most popular class items purchased. The least amount of people are willing to invest in luxury items. This could suggest that only a certain number of people can afford them or because they are not seen as a necessity.

Why Bought:

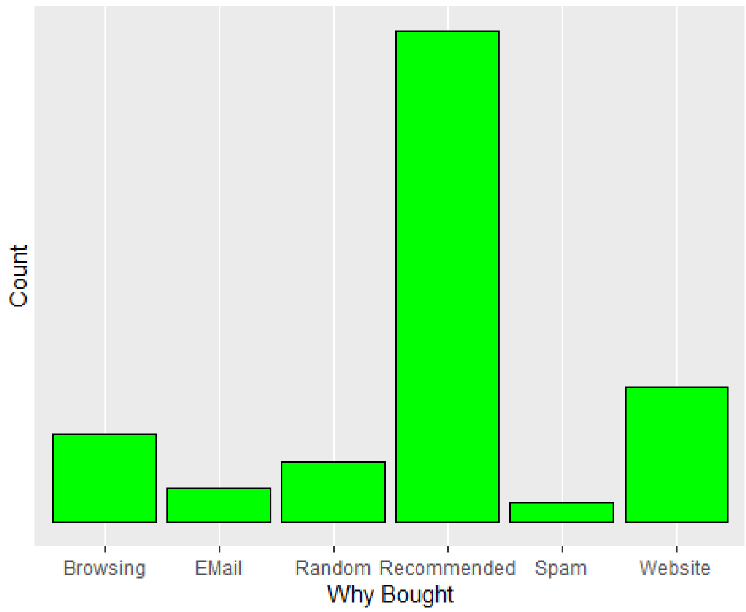

Figure 9:Why Bought vs Count

The ‘Why Bought’ feature represents valuable information as companies often want to know how their customers became aware of/ interested in their product. This aids them when discussing different marketing strategies. The graph above clearly indicates that most of the reasons as to why the different classes of products were bought was because it was recommended to the customers. This could be from indirect or direct relationships with the company.

Delivery time:

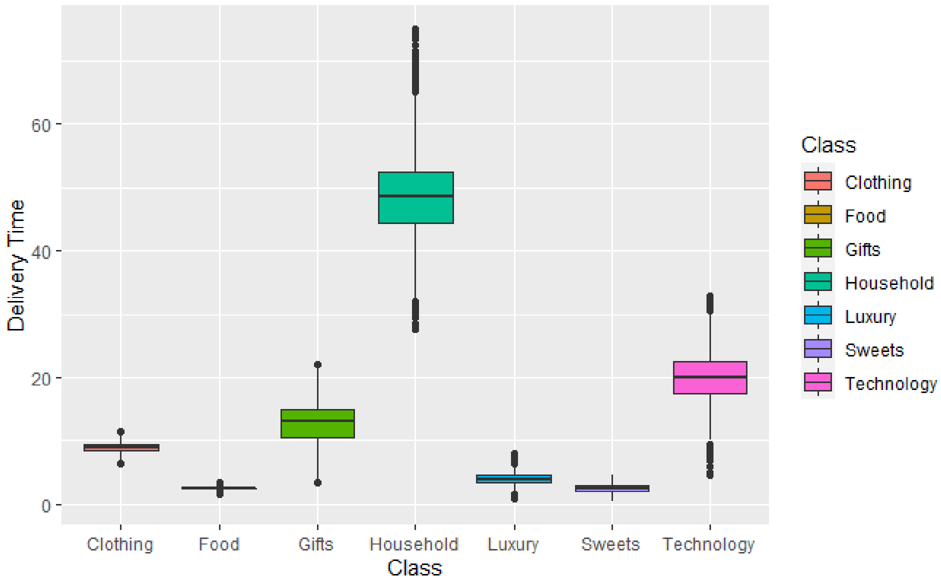

Figure 10:Boxplot of Delivery Time vs Class

	techdata	clothdata	housedata	luxdata	fooddata	giftsdata	sweetsdata
Min	4.50000	6.500000	27.50000	1.00000	1.500000	3.50000	0.500000
Max	33.00000	11.500000	75.00000	8.00000	3.500000	22.00000	4.500000
Mean	20.01095	8.999527	48.71956	3.97152	2.502014	12.89055	2.501206

Figure 11:Table summarizing Delivery time vs Class

By viewing the boxplot, we are able to see which of the different classes has the longest or shortest delivery times. It is informative as it displays the mean of the delivery time of each individual class as well as the various outliers represented by black dots. The total mean delivery time is 14.50031. this indicates that household goods have the longest delivery time. This could be due to the fact that they are not highly prioritized or because they are larger products and are thus more difficult to transport.

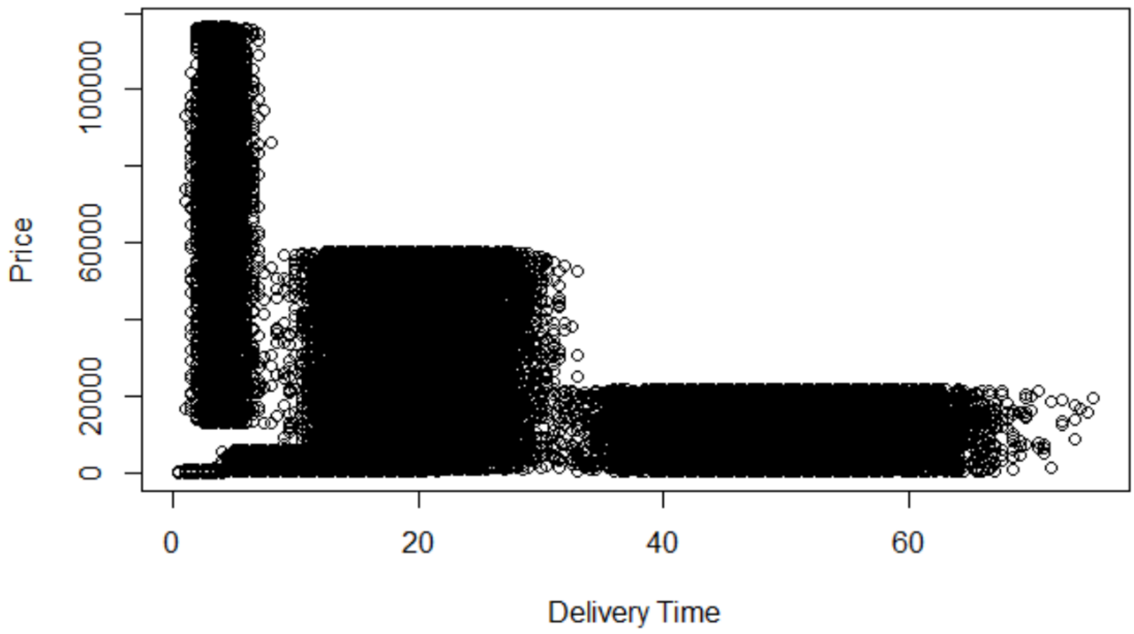

Figure 12:Delivery time vs Price

This delivery time vs Price graph indicates how the priority of the delivery goes up when the item is more expensive. There is a clear indication that more expensive class items have the quickest delivery time.

4. Part 3: Statistical Process Control

For the SPC we constructed control charts for the delivery process times. We initialized x and s charts for each individual class of sale. If the control limits of the s charts are within the acceptable range of +- 3sigma, we can accurately plot the x graphs within the correct limits. Centre lines, outer control limits, 2 sigma control limits and the 1 sigma control limits for both charts of the seven processes were determined using the first 30 samples of 15 sales each.

Initialization charts for Luxury Class:

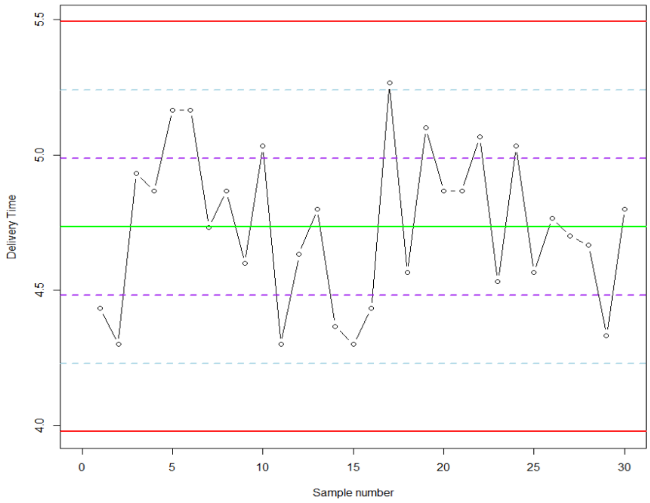

Figure 14:Luxury xBar graph

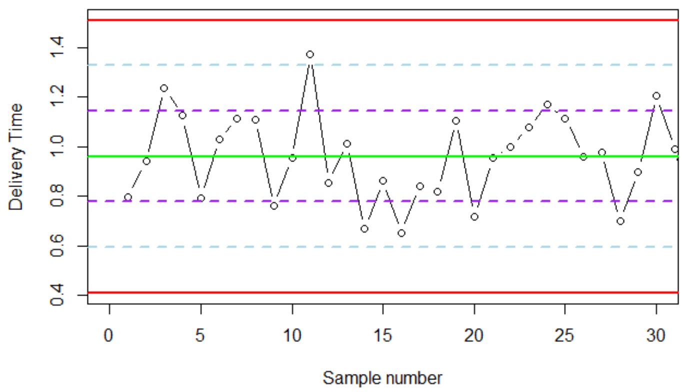

Figure 13:Luxury sBar graph

Six-sigma values for x-chart

	UCL	U2	U1	CL	L1	L2	LCL
Clothing	9.40	9.26	9.11	8.97	8.83	8.68	8.54
Household	50.25	49.02	47.79	46.56	45.33	44.10	42.87
Food	2.71	2.64	2.56	2.49	2.42	2.34	2.27
Technology	22.97	22.10	21.24	20.37	19.50	18.64	17.77
Sweets	2.90	2.76	2.62	2.48	2.34	2.20	2.06
Gifts	9.49	9.11	8.74	8.36	7.98	7.61	7.23
Luxury	5.50	5.25	4.99	4.74	4.49	4.23	3.98

Figure 15:Six-Sigma values for x-chart

Six-sigma values for s-chart

	UCL	U2	U1	CL	L1	L2	LCL
Clothing	0.87	0.76	0.66	0.55	0.45	0.34	0.24
Household	7.34	6.45	5.56	4.67	3.78	2.89	2.00
Food	0.44	0.39	0.33	0.28	0.23	0.17	0.12
Technology	5.18	4.55	3.93	3.30	2.67	2.04	1.41
Sweets	0.84	0.74	0.63	0.53	0.43	0.33	0.23
Gifts	2.25	1.98	1.70	1.43	1.16	0.88	0.61
Luxury	1.51	1.33	1.14	0.96	0.78	0.59	0.41

Figure 16:Six-sigma values for s-chart

Control Chart for Luxury

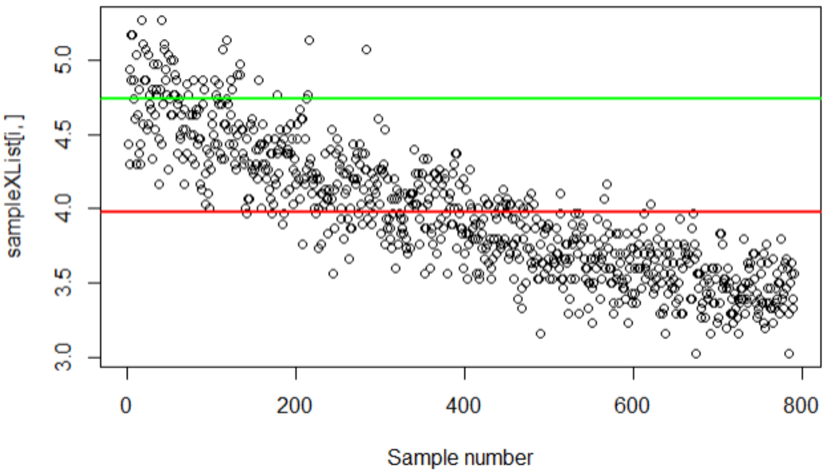

Figure 17:x control chart for luxury

By viewing the constant downward trend in figure 17, it is clear that the process never seemed to be in complete control from start to finish. Although a portion of the data was found within the control limits for a brief period, the mean was constantly decreasing.

5. Part 4: Optimizing delivery time

4.1A: Sample means outside of control limits

Class	Number of Instances outside of Control Limits
<fctr>	<int>
Clothing	20
Household	395
Food	4
Technology	19
Sweets	4
Gifts	2287
Luxury	440

Figure 18:table indicating the total number of samples found outside of the control limits.

Figure 18 represents the total number of samples allocated to each class that falls outside of the 3Sigma control limits. Due to the high number of samples found in clothing, household, technology, gifts and luxury; we will only be utilizing the first and last four samples. As seen, Luxury has a total number of 440 outliers. This corresponds to our control graph above.

Clothing	Household	Food	Technology	Sweets	Gifts	Luxury
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
282	252	75	37	942	213	142
837	387	432	345	1243	216	171
1723	1336	1149	2009	2009	2608	790
1756	1337	1408	2071	2071	2609	791

Figure 19:table indicating the first and last four sample numbers that were found outside the control limits.

Due to both food and sweets having a total number of 4 outliers, I believed using the first and last 2 samples instead of three would be more suitable to avoid any overlapping occurring.

4.1 B: Finding 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

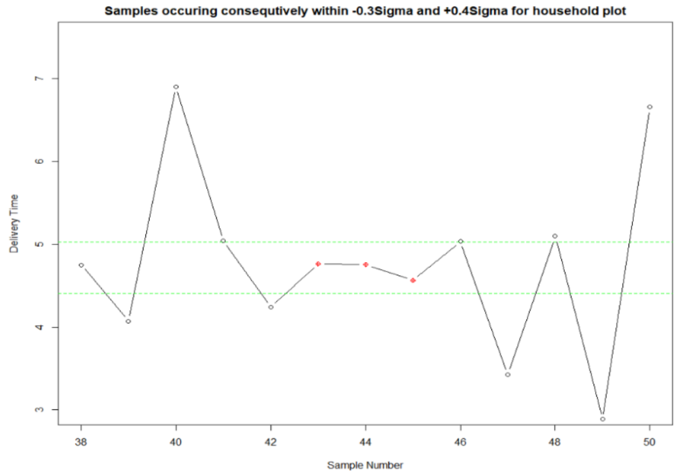

Figure 21:Consecutive household samples

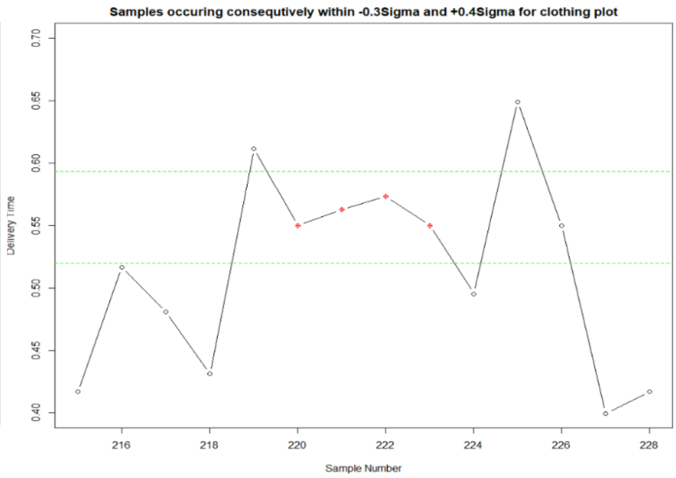

Figure 20:Consecutive clothing samples

As seen from figure 21, the highest order of samples occurring consecutively after one another for household class is 3. This is displayed by the red dots appearing between the 2 green lines, representing the upper and lower control limits. Similarly for the clothing class, 4 samples were found to be the highest number of samples occurring consecutively.

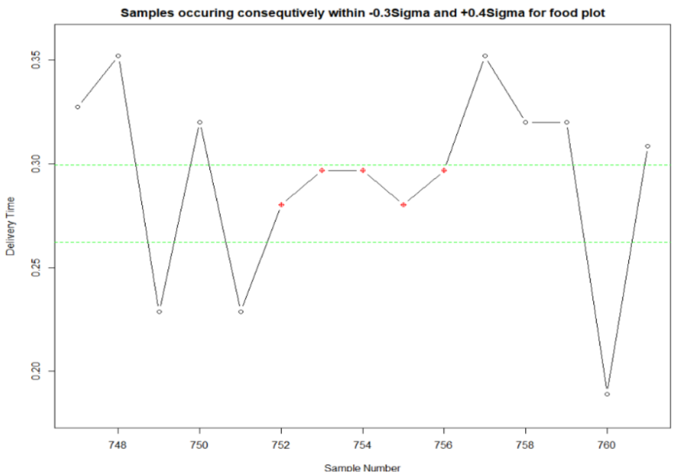

Figure 23:Consecutive food samples

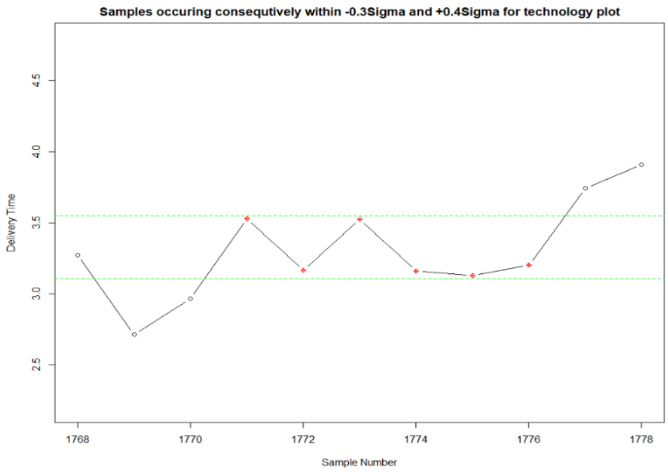

Figure 22:Consecutive technology samples

In figure 23, 5 consecutive red dots occur for the food samples and 6 occur consecutively in figure 22 for our technology plot.

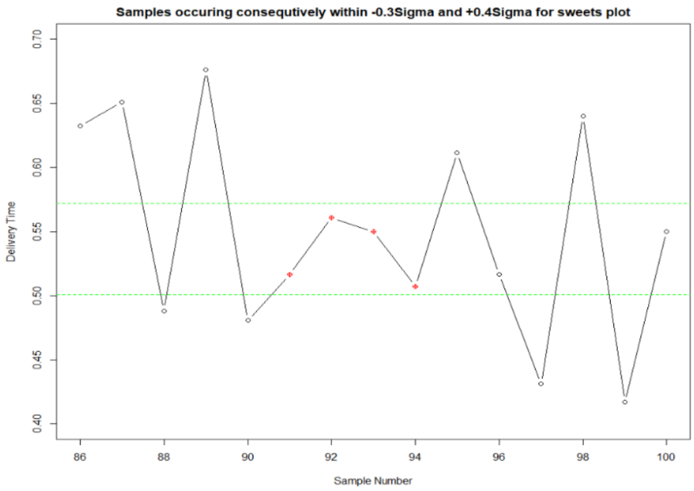

Figure 25:Consecutive sweets samples

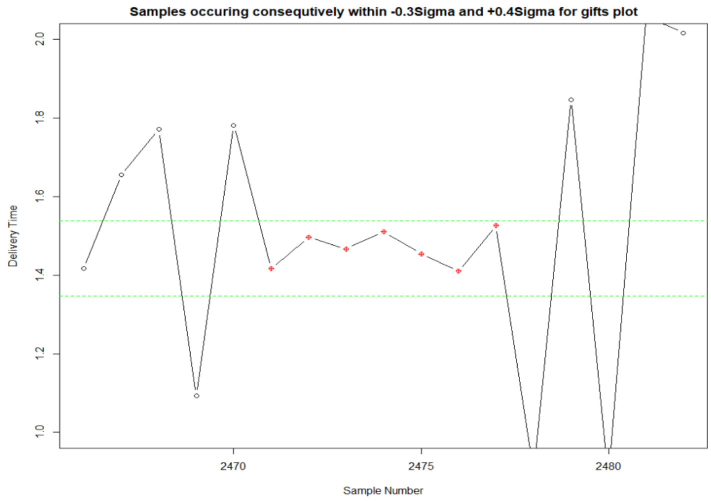

Figure 24:Consecutive gifts samples

Following the method of identifying the consecutive samples, we see that in figure 25, 4 samples resulted in the highest for sweets and 7 for gifts.

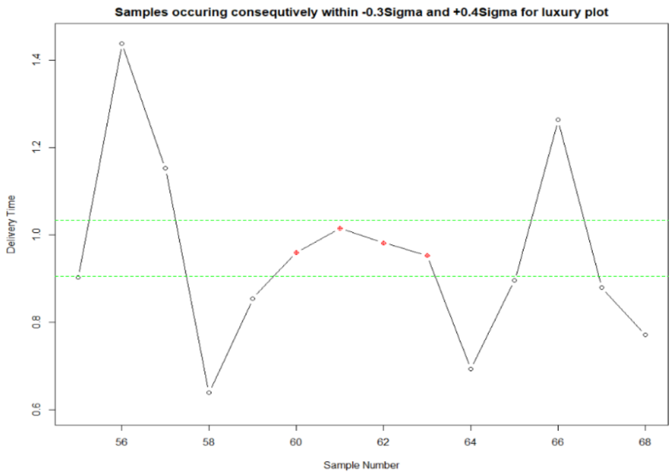

Figure 26:Consecutive luxury samples

Finally, for our luxury plot, we identified 4 consecutive samples occurring within the 2 limits to be the highest.

4.2: Type I error

		A	B
1	Type 1 error probability	0.002699796	0.106991

Figure 27:Type I Errors

We can assume a normal distribution of the control charts when calculating the type I error. A type I error is a false positive conclusion of results. Ho would be, “the process is in control and centered on the centerline calculated using the first 30 samples”. Therefore, for A, there is a 0.27% probability that the process will be assumed out of control when it in fact is in control. This probability is very small which can be expected, however there is still a slight chance of this happening. By this happening it would mean that the sample would need to be found outside of the +-3Sigma control limits, which in other words is known as the Upper and Lower control limits; if this is the case, it would be known as the type I error.

For B, our control limits are -0,3Sigma to 0,4Sigma. We use a reference of 7 consecutive samples as this is the highest, seen above in figure 23. Therefore, the probability of 10,7% calculated for B displayed in figure 25 indicates the probability of 7 samples consecutively occurring within the specified control limits.

4.3: Optimizing delivery time for Technology to maximize profit

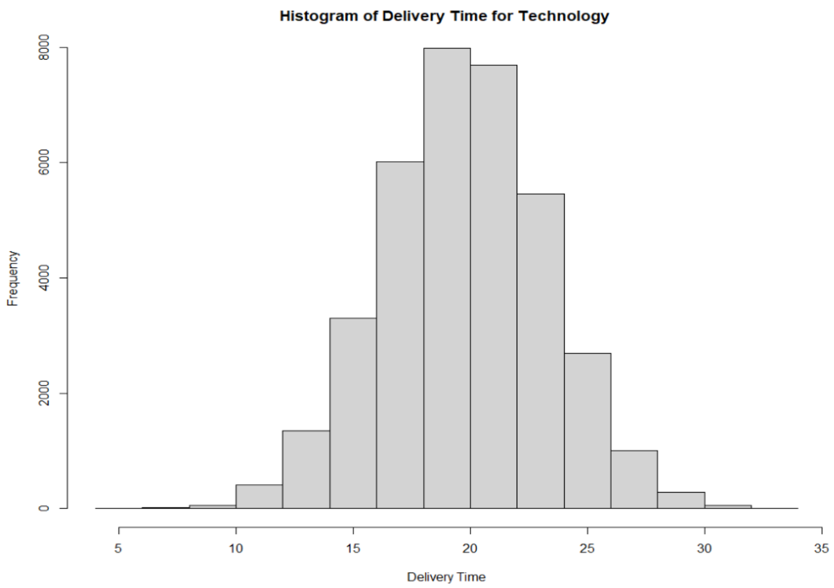

Figure 28:Delivery times for technology

The graph in figure 28 displays the current delivery times for Technology. The aim is to optimize this by reducing or increasing the overall time it takes to deliver the technology; given the stated limitations of costs in the problem statement.

Calculating the number of hours to shift the delivery times for the highest economic benefit:

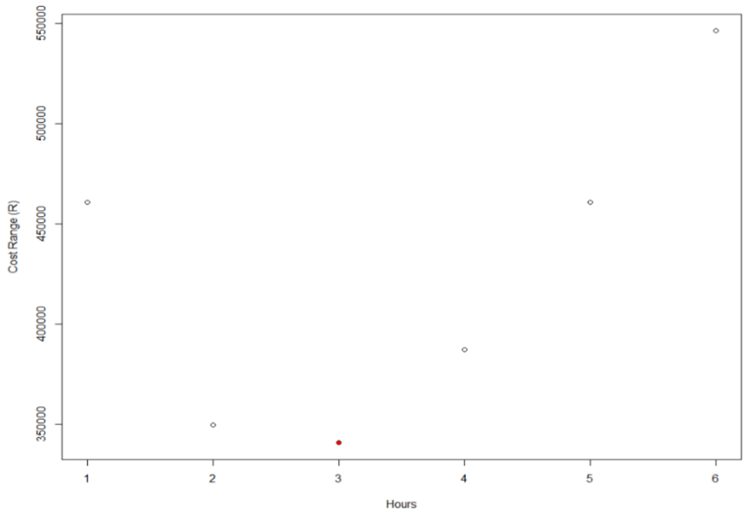

Figure 29:Cost difference when decreasing delivery hours

Figure 29 indicates the change in costs when increasing the number of hours the delivery time is decreased by. As can be seen, the cost is at its lowest point at 3 hours and therefore the optimal amount of hours to decrease the deliver time by is 3.

4.4

A type II error results in a false negative conclusion. This means that the process will be assumed in control when in fact it is out of control. In other words, the sample instances will be outside of the allocated control limits but the process will deem it to be within the control limits.

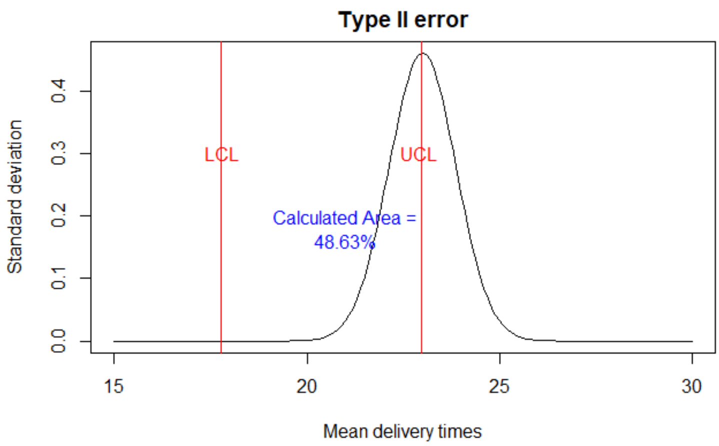

Figure 30:Type II error

Figure 30 displays that the control limits remain the same however the mean delivery time for technology moves to 23 hours. This indicates that there is a 48.63% chance that a type II error will be made, or in other words that the sample will not fall within the original control limits but the null hypothesis will fail to be rejected.

6. Part 5: Doe and Manova

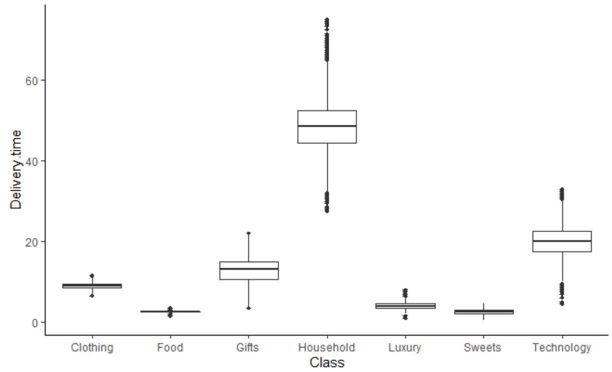

Figure 33:Class vs delivery time boxplot

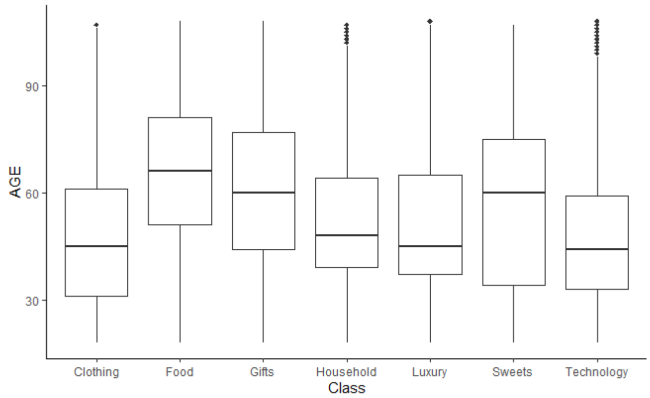

Figure 32: Age vs Class boxplot

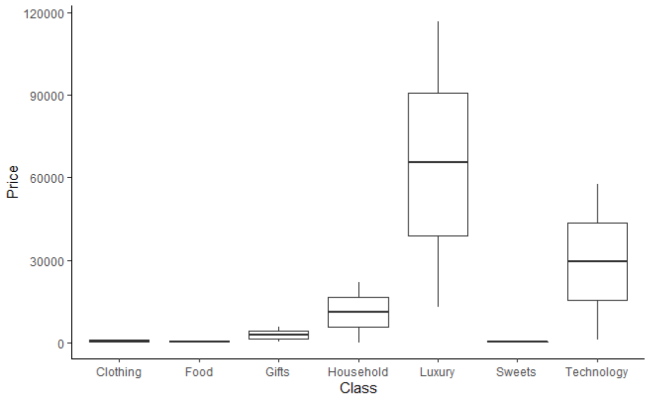

Figure 31:Class vs Price boxplot

```
                 Df Pillai approx F num Df den Df    Pr(>F)    
Class             6 1.7578    16262     30 899855 < 2.2e-16 ***
Residuals 179971                                              
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 Response Day :
                Df   Sum Sq Mean Sq F value Pr(>F)
Class            6      668 111.302   1.488 0.1777
Residuals   179971 13461680  74.799               

 Response Month :
                Df  Sum Sq Mean Sq F value Pr(>F)
Class            6      87  14.576  1.2219 0.2913
Residuals   179971 2146871  11.929               

 Response AGE :
                Df   Sum Sq Mean Sq F value    Pr(>F)    
Class            6  8422401 1403733    3805 < 2.2e-16 ***
Residuals   179971 66394669     369                      
---
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

 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
```

As can be seen from the table in figure 34, the significance codes represent the strength of the relationship between the specified features (Day, Month, Age, Delivery time, and Price) and Class. The day and month a product was bought has no significant relationship towards the class of product bought. Delivery time has the highest correlation with the class of product bought due to its F value being the highest. Price having a lower F value than delivery time could indicate a weaker relationship, however it is still strong. Age also has a high significance but the weakest relationship relative to delivery time and price. These three descriptive features were plotted against class on boxplots for a visual representation of their relationships in terms of the data set. This can be seen in figure 31, 32 and 33.

7. Part 6: Reliability of the service and products

6.1:

The customer experiences a loss of quality the moment product specification deviates from the 'target value'. This 'loss' is depicted by a quality loss function and it follows a parabolic curve mathematically given by L = k(y–m)2, where m is the theoretical 'target value' or 'mean value' and y is the actual size of the product, k is a constant and L is the loss. This means that if the difference between 'actual size' and 'target value' i.e. (y–m) is large, loss would be more, irrespective of tolerance specifications.

Problem 6:

L(x) = k (x - T)

45 = k (0.04) ^2

k = 28125

L(x) = 28125 (x - T) ^2

Problem 7:

a) L(x) = k (x - T)
   35 = k(0.04)2
   K = 21875
   L(x) = 21875(x - T)2
b) L(0.027) = 21875(0.027)2
   = 15.95

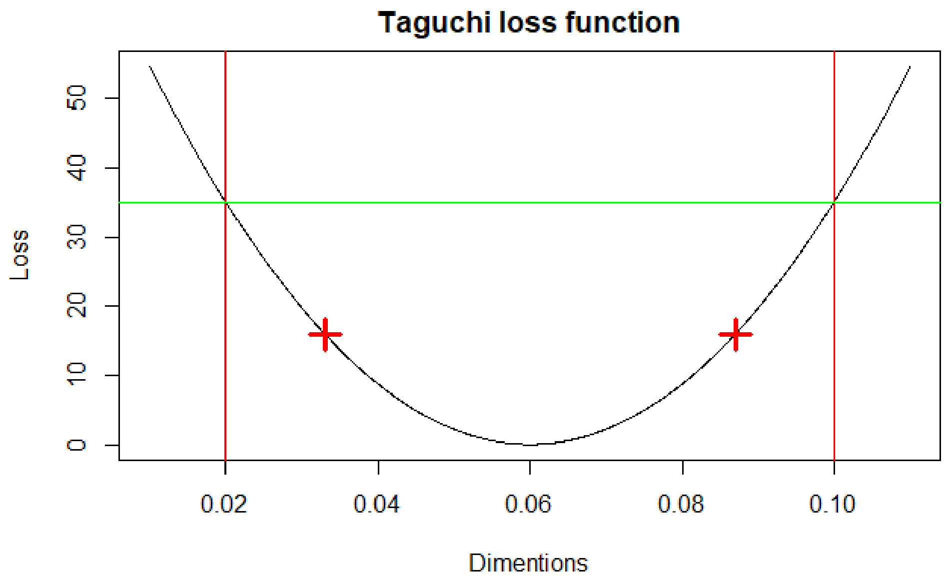

Figure 34:Taguchi loss function

In figure 34 we can see that at 0.06cm, which is the specification, the cost is 0. As the specifications vary from this, the cost increases. The vertical lines are +- 0.04 away from the specification, which is where the scrap cost of $35 is located, which is identified as the horizontal line. The '+' signs found on the parabolic curve is the cost calculated in problem 7b.

6.2

a) The reliability of the system when in series:

RaRbRc= (0.85)(0.92)(0.90)

RaRbRc= 0.7038

b) The reliability when the system in now in parallel:

RaaRbbRcc = [1−(1−Ra)2] * [1−(1−Rb)2] * [1−(1−Rc)2]

RaaRbbRcc = [1−(1−0.85)2] * [1−(1−0.92)2] * [1−(1−0.9)2]

RaaRbbRcc = [1−(0.0225)] * [1−(0.0064)] * [1−(0.01)]

RaaRbbRcc = [0.9775] * [0.9936] * [0.99]

RaaRbbRcc = 0.9615

Reliability is higher when the system is in parallel.

6.3

21 vehicles:

Vehicle availability:

P(all vehicles available) = 0.8615411

Days with all vehicles available = 314.5625

P(1 vehicle not available) = 0.1288543

Days with 1 vehicle not available = 47.03181

Total vehicle reliability = 0.8615411 + 0.1288543 = 0.9903954

Total days with 1 or less vehicles unavailable = 361.4943

Driver reliability:

P(Driver reliability) = 0.003224402

P(Zero drivers off) = 0.9344269

Days with zero drivers off = 0.9344269*365 = 341.0658

P(One driver off) = 0.06347701

Days with one driver off = 0.06347701*365 = 23.16911

Total driver reliability = 0.9344269 + 0.06347701 = 0.997904

Total days with reliable driver = 0.997904*365 = 364.235

This indicates a very high probability that there will be a driver available every day to complete deliveries.

Total delivery reliability = 0.988319

Days with reliable delivery = 0.988319*365 = 360.74

This indicates an area of improvement. Although they only have around 4 or 5 days of the year with unreliable service, one of the main reasons as to why their products are purchased are through recommendations from other customers. This provides good motivation to strive for perfection. If their services are always reliable the business will receive much more customers and thus greater revenue.

Increasing number of vehicles to 22:

P(0 vehicles breaking down) = 0.8554486

Days with 0 vehicles breaking down = 312.2387

P(1 vehicle breakdown) = 0.1340356

Days with 1 vehicle breaking down = 48.923

P(2 vehicle breakdowns) = 0.0100234

Days with 2 vehicle breakdowns = 0.0100234*365 = 3.65852

Total delivery reliability = 0.0100234+0.8554486 + 0.1340356 = 0.9995076

Days with reliable delivery = 0.9995076*365 = 364.056

This indicates that by increasing the total number of available vehicles by 1, the delivery reliability increases to near 100% of the year

Conclusion

This report uses a valid data set of 17978 instances to correlate certain class items belonging to an online sales company to different descriptive features of the business. Features such as delivery time, Age and Price have a strong relationship to the class of items sold while time related features such as months and days these items were sold did not. The delivery time of technology was optimized to reduce the costs involved when taking certain penalties in account. Reducing the delivery times by 3 hours had the greatest economic benefit for the company. We investigated the probability of a type I and type II errors incurring and the consequences both of these errors might lead to.

Bibliography

Jobe, J. M. (2016). *Statistical Methods in Quality Assurance.* New York: Springer Nature.

Scribbr. (2021, January 18). *Scribbr*. Retrieved from Scribbr: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

STHDA. (2020, February 17). *STHDA*. Retrieved from STHDA: http://www.sthda.com/english/wiki/r-plot-pch-symbols-the-different-point-shapes-available-in-r