2022

ECSA PROJECT

Quality Assurance 354

PETRIE VAN RENSBURG
23985488

Table of Contents

Table of Figures

Introduction

In this project a comprehensive analysis of a dataset will be performed by means of the R-Studio coding platform, and this will enable the user to fully understand all aspects of the dataset. The analysis will include sorting the initial set into different subsets and plotting different results of different aspects to gain further knowledge and understanding. The business and its systems will be monitored as well to provide a full overview of the overall performance.

Data Wrangling (Part One)

The first job was to identify possible issues with the raw data set and deal with them appropriately. It was noticed that there were a few missing values and negative values in the "Price" column of the dataset. The best way to deal with this was to completely remove these instances from the set and to create an invalid dataset to store them in. It was noticed that there were 17 "N/A" values in the price column.

	Primary.key	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
12345	1	12345	18973	93	Gifts	NA	2026	6	11	15.5	Website
16321	2	16321	81959	43	Technology	NA	2029	9	6	22.0	Recommended
19541	3	19541	71169	42	Technology	NA	2025	1	19	20.5	Recommended
19999	4	19999	67228	89	Gifts	NA	2026	2	4	15.0	Recommended
23456	5	23456	88622	71	Food	NA	2027	4	18	2.5	Random
34567	6	34567	18748	48	Clothing	NA	2021	4	9	8.0	Recommended
45678	7	45678	89095	65	Sweets	NA	2029	11	6	2.0	Recommended
54321	8	54321	62209	34	Clothing	NA	2021	3	24	9.5	Recommended
56789	9	56789	63849	51	Gifts	NA	2024	5	3	10.5	Website
65432	10	65432	51904	31	Gifts	NA	2027	7	24	14.5	Recommended
76543	11	76543	79732	71	Food	NA	2028	9	24	2.5	Recommended
87654	12	87654	40983	33	Food	NA	2024	8	27	2.0	Recommended
98765	13	98765	64288	25	Clothing	NA	2021	1	24	8.5	Browsing
144444	14	144444	70761	70	Food	NA	2027	9	28	2.5	Recommended
155555	15	155555	33583	56	Gifts	NA	2022	12	9	10.0	Recommended
166666	16	166666	60188	37	Technology	NA	2024	10	9	21.5	Website
177777	17	177777	68698	30	Food	NA	2023	8	14	2.5	Recommended

Figure 1 - Invalid Dataset

5 negative values were noticed in the price column, surprisingly all with the value of -512, and the instances containing these negative values were also removed.

After dealing with the missing values and negative values a new column, named "Primary.key", was added to act as a index starting from one and ending at the last instance of the dataset. The column named "X" now acts as a secondary index for all instances.

Descriptive Statistics (Part Two)

To analyse the data, different plots were created to gain a better understanding of how the data is distributed. It enables the user to identify trends in data or to see how certain features impact the values.

The first plot is the distribution of age over all the sales and it is clear that there is an strong increase in sales from a young age until 40 and then a gradual decrease as the users grow older.

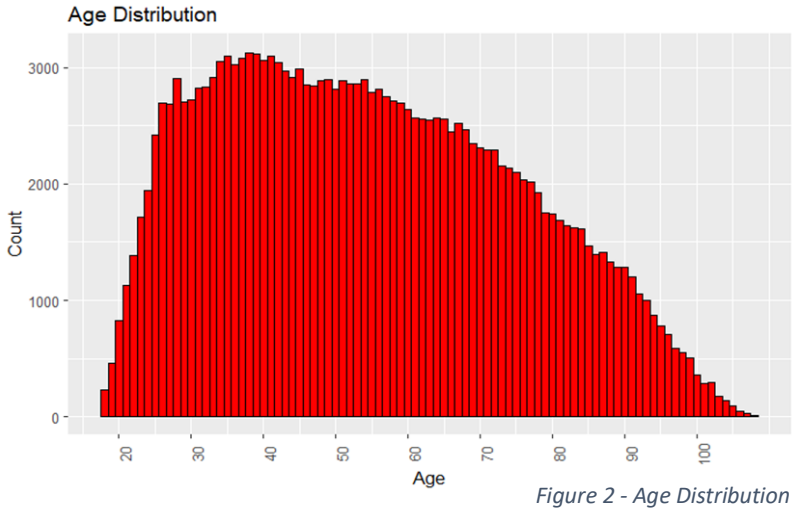


Figure 2 - Age Distribution

Min	1st Q	Median	Mean	3rd Q	Max
18.00	38.00	53.00	54.57	70.00	108.00

The next plot will show how the price of all the sales is distributed. It was noted that the price range vary a lot and that the densest distribution is in the lower price range.

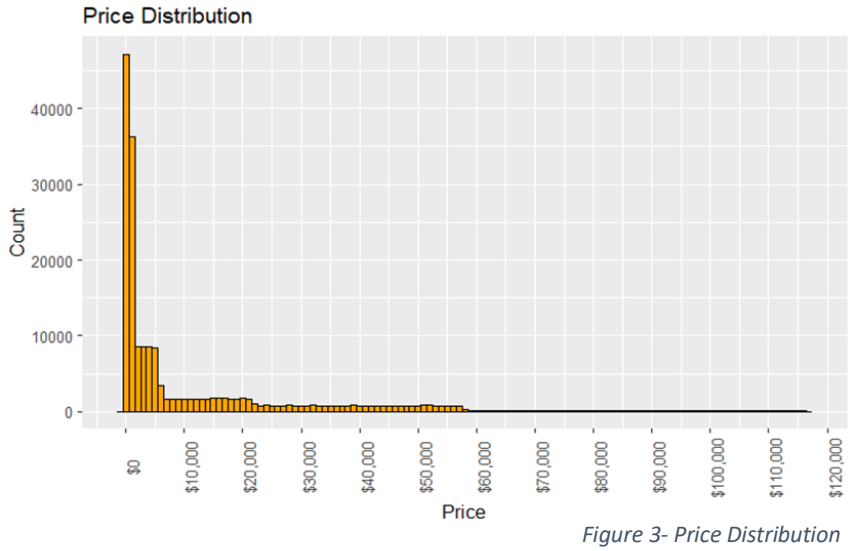


Figure 3- Price Distribution

Min	1st Q	Median	Mean	3rd Q	Max
-588.8	482.3	2259.6	12293.7	15270.7	116619.0

The delivery time distribution is plotted next, and this plot illustrates a bimodal normal curve, and this could possibly be due to the spread of delivery locations and possible difficulties that comes with delivering in certain areas. This could also be due to a split between local and international deliveries.

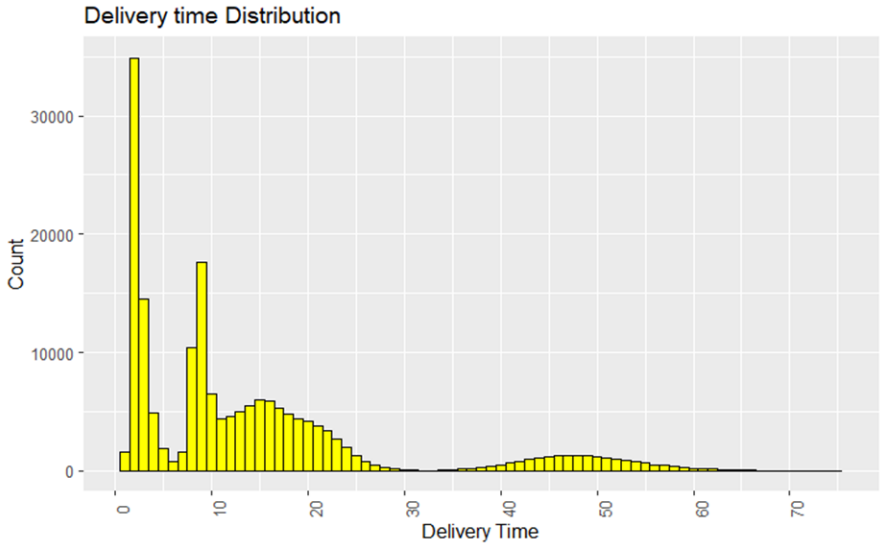


Figure 4 - Delivery time distribution

Min	1st Q	Median	Mean	3rd Q	Max
0.5	3.0	10.0	14.5	18.5	75.0

The following plot will illustrate a histogram of the count of the different reasons why customers bought something from this company. It is obvious that the reason "Recommended" is most popular. This is a very insightful plot, and the company can use it to further gain a better understanding of customer behaviours.

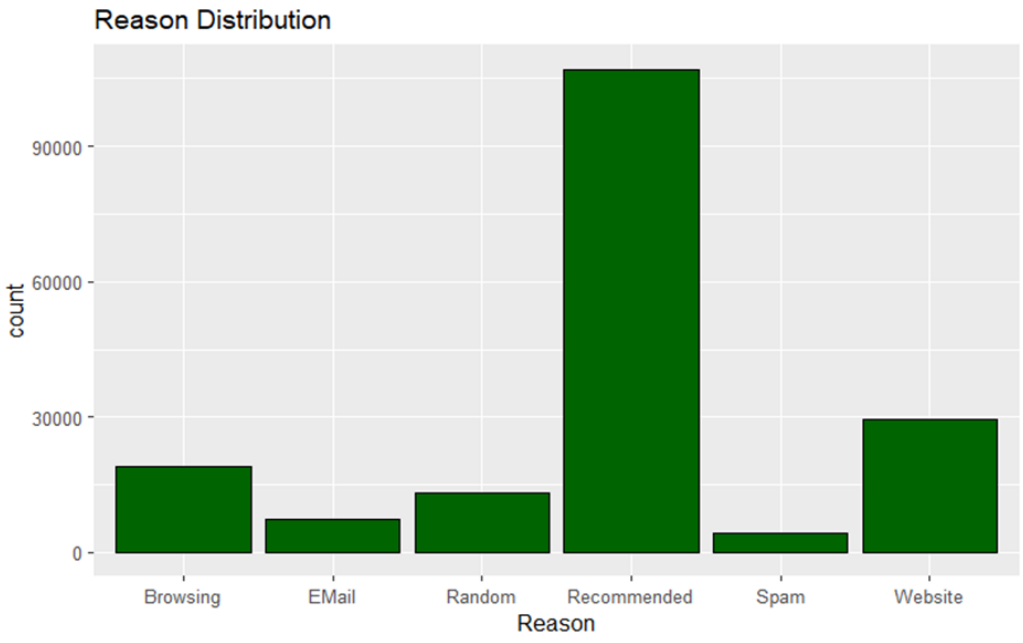


Figure 5 - Reasons for buying

The following plot is also a plot of the age distribution, but this time for each individual class, and the resemblance is apparent. Again, a strong increase until around 30 or 40 years old, followed by a gradual decrease. This is insightful as the company can use this plot to identify their target market and age for each class separately which may increase sales if handled correctly.

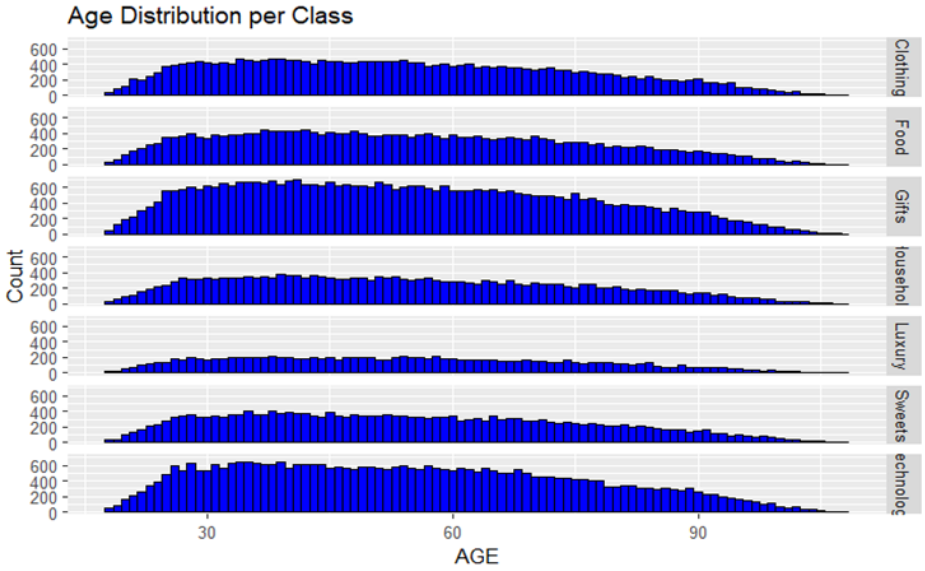


Figure 6 - Age per Class

Now the price per class distribution will be displayed and it is apparent that some classes like clothing, food, and sweets have a more concentrated distribution in comparison with classes like household, luxury, and technology which vary a lot in price.

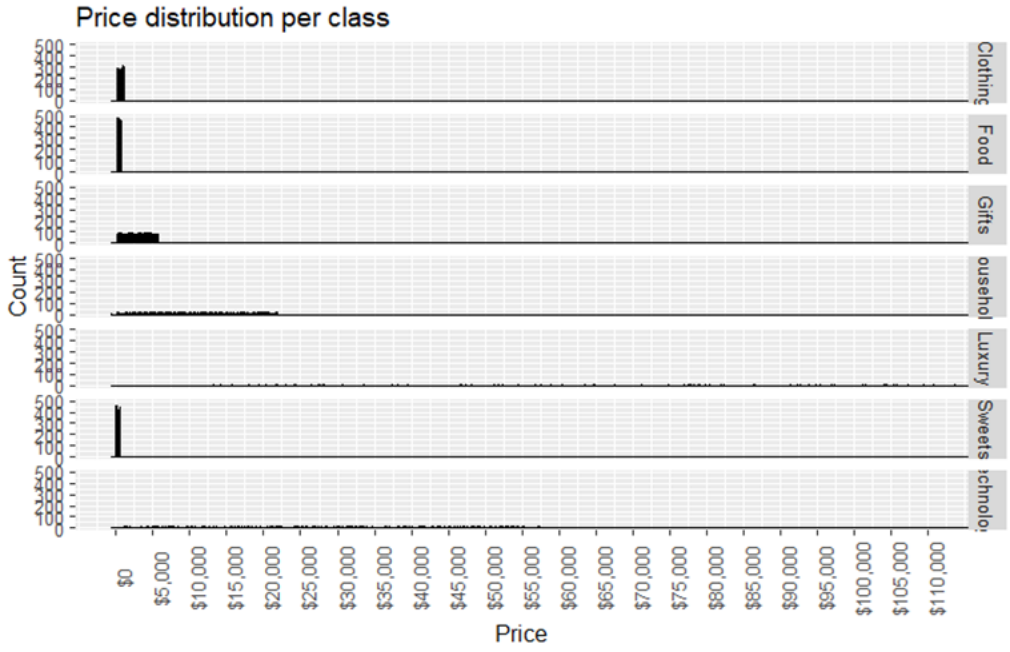


Figure 7 - Price per Class

The delivery time per class will now follow and this is a very helpful plot seeing that the company can now identify where they can increase or decrease the delivery time for certain classes. It is clear that a class like household will take longer to deliver, seeing that this might be larger products, needing larger modes of transport.

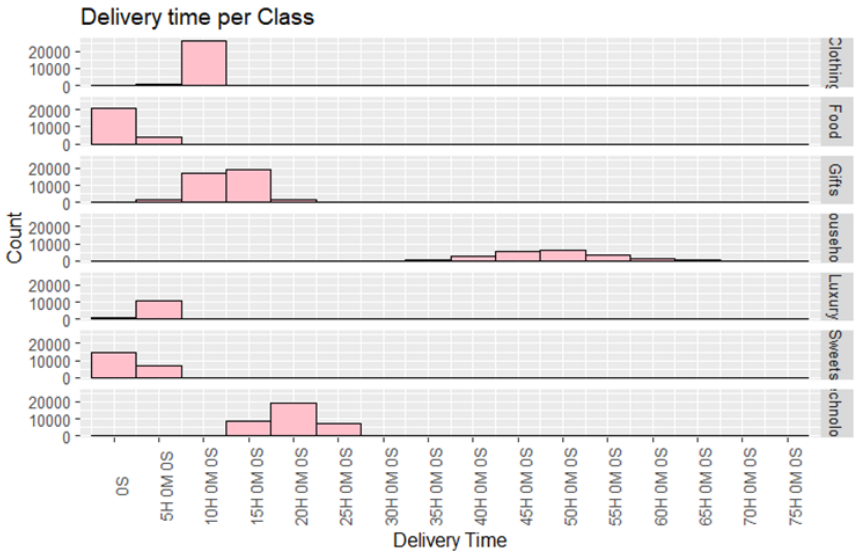


Figure 8 - Delivery time per Class

The final plot is an illustration of the popularity of each class in terms of how many sales occurred in each class.

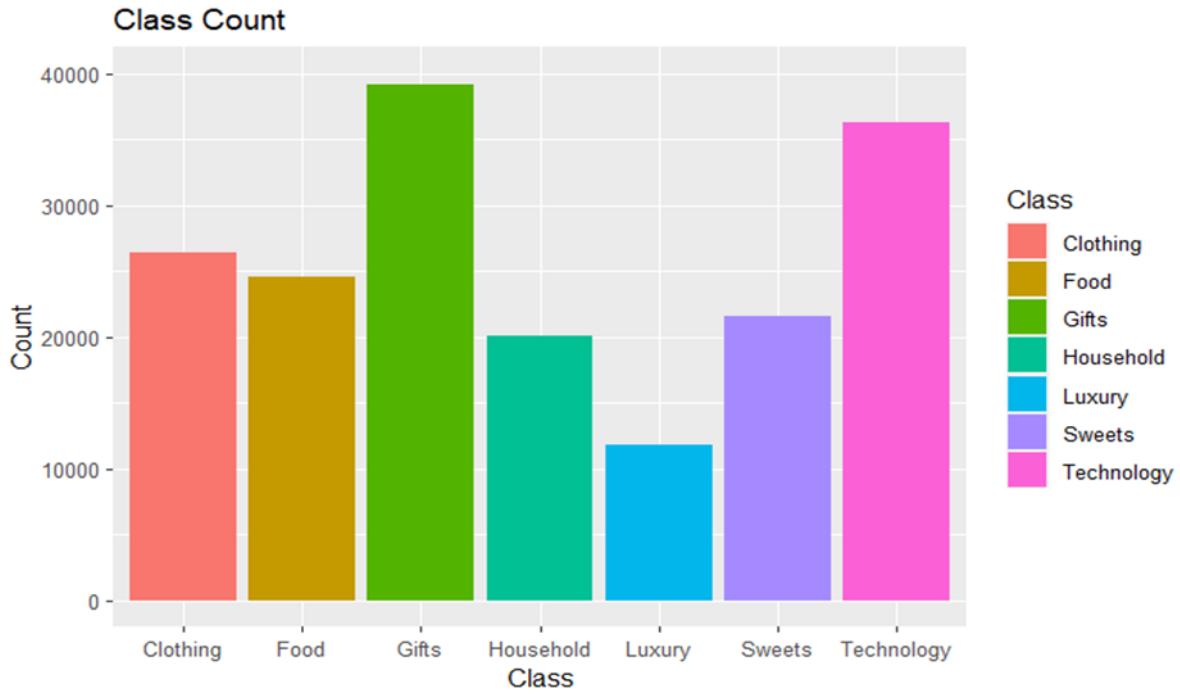


Figure 9 - Class Count

Process Capabilities:

USL	LCL	CP	CPU	CPL	CPK
24	0	1.142	0.379	1.904	0.379

Statistical Process Control (Part Three)

Before constructing the respective X- and – S Charts the valid dataset was reordered in terms of date, thus meaning from oldest to newest date. Then the data of 30 samples of 15 instances each were used to initialize the parameters. Overall, a set of 450 instances was created to calculate the different values of each chart. This was computed according to the different classes of the dataset and the output is as follows:

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	9.390601	9.250401	9.110200	8.970000	8.829800	8.689599	8.549399
Food	2.702226	2.631484	2.560742	2.490000	2.419258	2.348516	2.277774
Gifts	9.451412	9.087978	8.724545	8.361111	7.997678	7.634244	7.270811
Household	50.126859	48.938647	47.750435	46.562222	45.374010	44.185798	42.997585
Luxury	5.468973	5.224501	4.980028	4.735556	4.491083	4.246610	4.002138
Sweets	2.883225	2.748076	2.612927	2.477778	2.342629	2.207479	2.072330
Technology	22.888932	22.050770	21.212607	20.374444	19.536282	18.698119	17.859957

Figure 10 - X Chart

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	0.5661217	0.4259213	0.2857210	0.1455206	0.005320229	-0.13488014	-0.2750805
Food	0.2772340	0.2064920	0.1357501	0.0650081	-0.005733868	-0.07647584	-0.1472178
Gifts	1.4721776	1.1087441	0.7453106	0.3818771	0.018443639	-0.34498985	-0.7084233
Household	4.7468725	3.5586602	2.3704478	1.1822355	-0.005976865	-1.19418920	-2.3824015
Luxury	1.0194840	0.7750115	0.5305389	0.2860664	0.041593837	-0.20287870	-0.4473512
Sweets	0.5520133	0.4168641	0.2817149	0.1465656	0.011416392	-0.12373284	-0.2588821
Technology	3.4535377	2.6153751	1.7772125	0.9390500	0.100887383	-0.73727519	-1.5754378

Figure 11 - S Chart

After constructing the X-and-S Charts using the 450 values, the rest of the instances in the dataset was used to construct the samples from 31 and onwards. The control charts for the first 30 samples were plotted to gain further insight of the sample distribution and thereafter the control charts of all the data of each class were plotted. These charts allow us to clearly see the distribution of each class and how many values or how frequent some instances exceed the control limits in each class.

Control charts for samples

Figure 18 - Technology sample chart

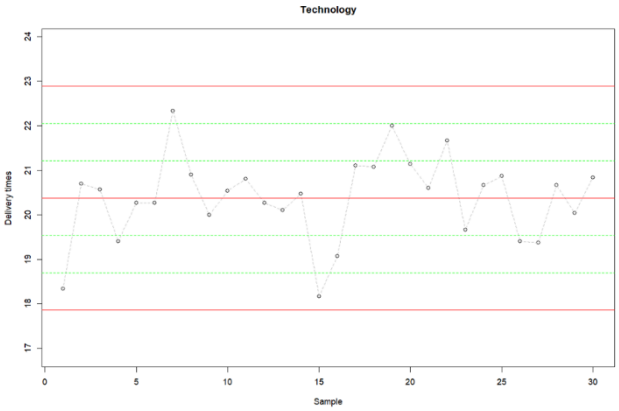


Figure 17- Sweets sample chart

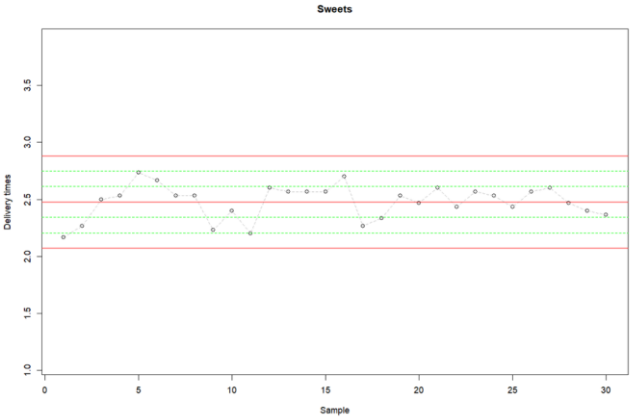


Figure 16 - Luxury sample chart

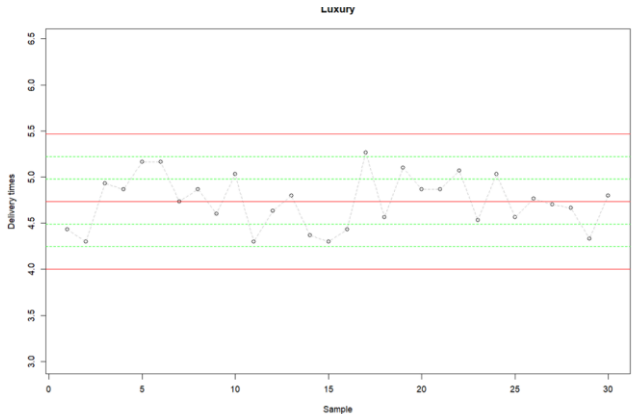


Figure 15 - Household sample chart

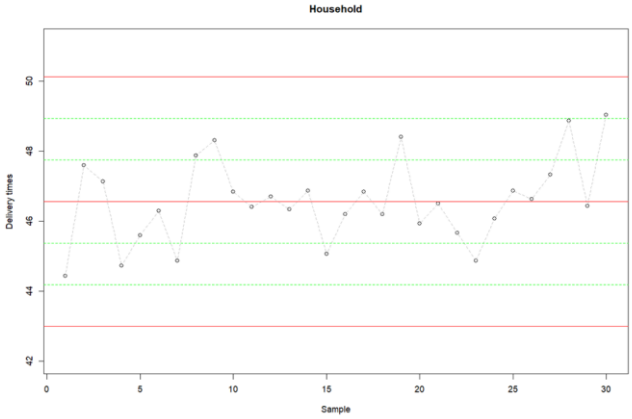


Figure 14- Gifts sample chart

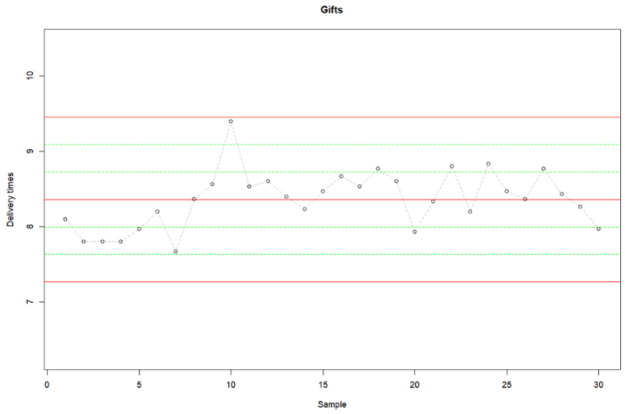


Figure 13- Food sample chart

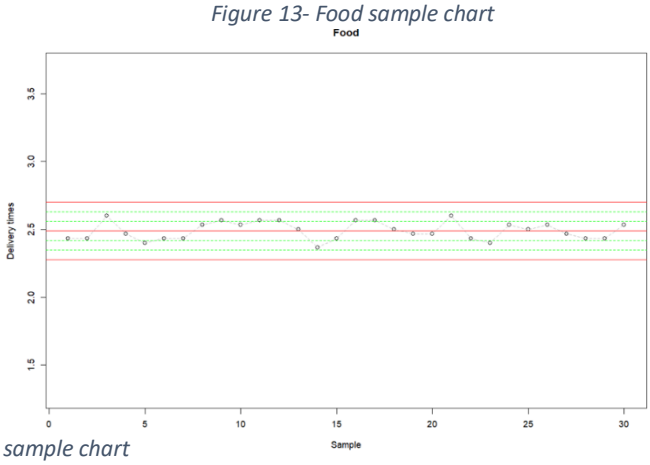
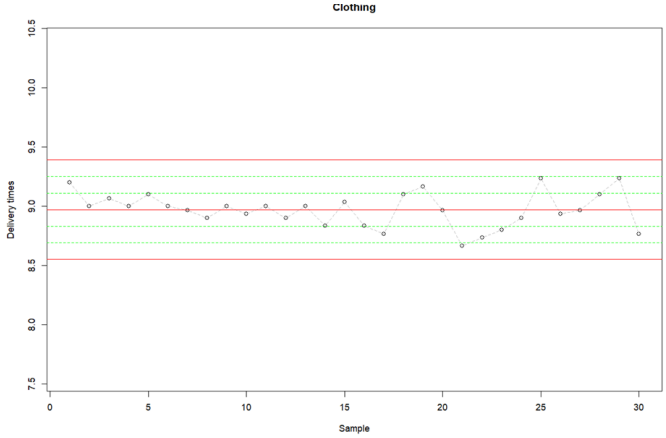


Figure 12- clothing sample chart



Control charts for all values

Figure 24 - Technology control chart

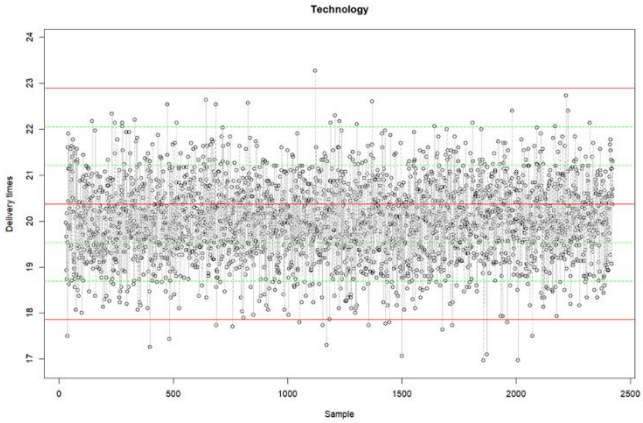


Figure 23- Sweets control chart

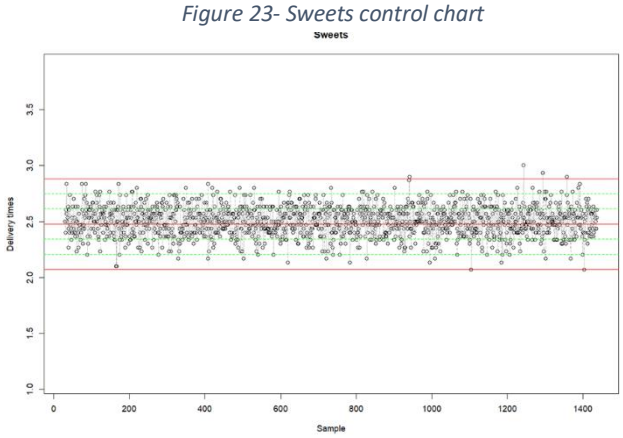


Figure 22 - Luxury control chart

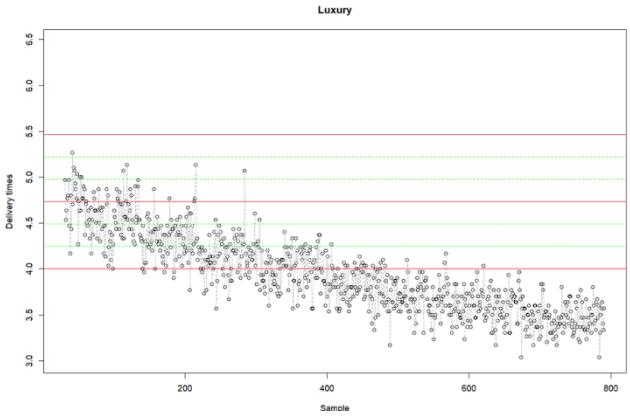


Figure 21 - Food control chart

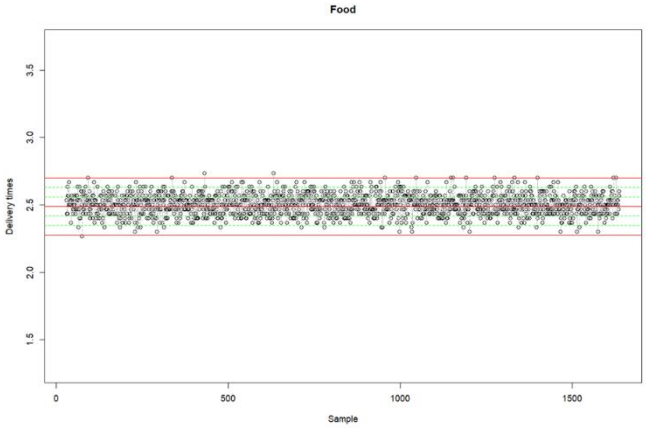


Figure 20 - Gifts control chart

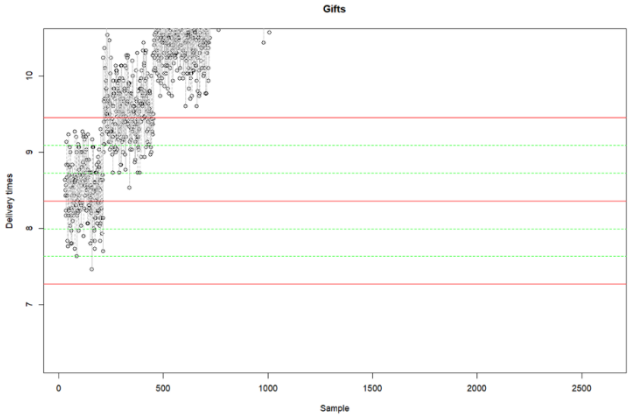


Figure 25 - Household control chart

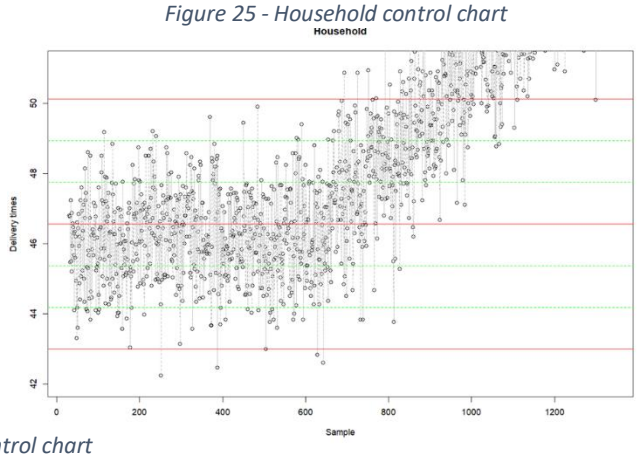
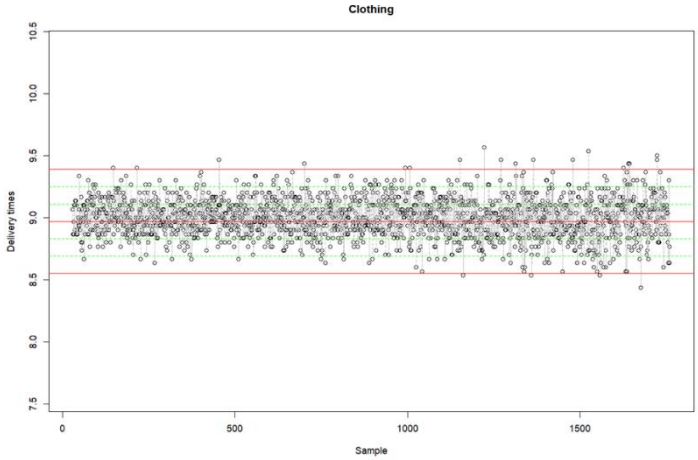


Figure 26 - Clothing control chart



Analysis:

It is very apparent that there is a large difference in the two plots of each class, displayed above. When looking at the sample charts, it is clear that next to none of the values are out of their limits. One may notice these values coming close to the UCL and LCL but never exceed them. This is only the first 30 samples and therefore should expect an result like this.

When looking at the control charts of all the samples we see a more realistic representation of the data and now trends in the data is more visible. When looking at the Gifts, Household, and Luxury classes one can clearly notice a trend in each where the mean values start increasing and exceeding the UCL (Gifts and Household) and decreasing and exceeding the LCL like seen in the Luxury class. The other four classes show a more stable distribution and will have less out of bounds values.

Statistical Calculations (Part Four)

4.1

These tables provide insight into the performance of the control charts and tracks the performance of each separate class.

Class	Number of samples out of control	First 3 and last 3 out of control	Percentage out of control
Technology	20	7, 368, 453, … 1931, 1979, 2041	0.055%
Clothing	22	118, 187, 425, … 1647, 1693, 1694	0.083%
Household	406	222, 357, 599, … 1305, 1306, 1307	2.023%
Luxury	434	112, 141, 154, … 759, 760, 761	3.656%
Food	3	45, 405, 603	0.012%
Gifts	2296	183, 186, 188, … 2577, 2578, 2579	5.865%
Sweets	6	912, 1074, 1213, 1264, 1328, 1373	0.028%

Class	Sequence length	Ending Sample
Technology	15	333
Clothing	23	1668
Household	37	1228
Luxury	79	716
Food	15	405
Gifts	16	2343
Sweets	21	1325

4.2

Error for A = 0.002699796 = 0.2699%

Error for B = 0.5

Meaning that the probability to make a type 1 error is 0.2699%

4.3

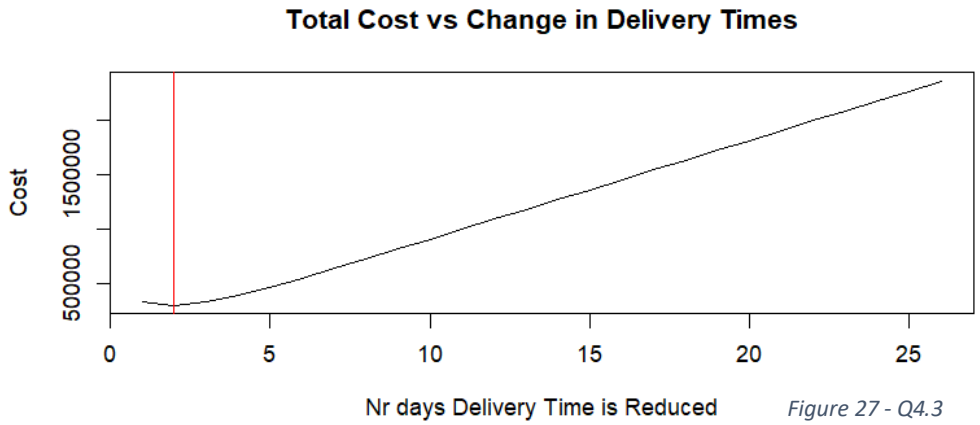


Figure 27 - Q4.3

Average delivery days = 20.01095

Reduce days by 2 days

Minimum Cost = 298201

Optimal number of delivery days = 24 days

4.4

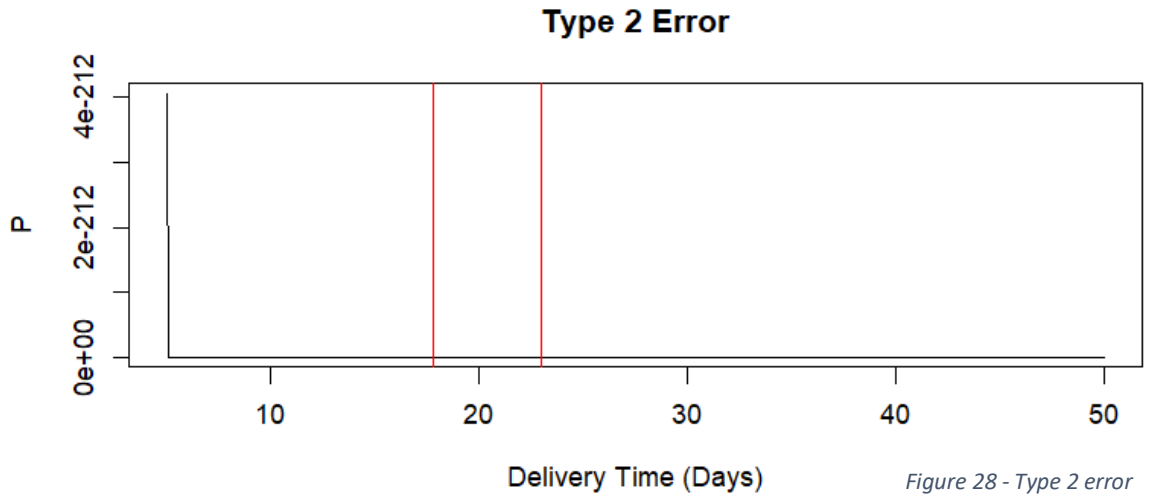


Figure 28 - Type 2 error

Error = 0.4876106

= 48.76% chance of a type 2 error

DOE/MANOVA (Part Five)

Here a MONAVA was set up to determine whether the class of each sale has an influence on the price and delivery time. A Hypothesis test was conducted as follows:

H0, Delivery time: The class of an item have no significant influence on the delivery time of an item.

H1, Delivery time: The class of an item does have a significant influence on the delivery time of an item.

H0, Price: The class of an item have no significant influence on the price of an item.

H1, Price: The class of an item does have a significant influence on the price of an item.

```
             Df Pillai approx F num Df den Df    Pr(>F)
Class         6 1.6796   157265     12 359952 < 2.2e-16 ***
Residuals 179976
---
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 33461034 5576839  629489 < 2.2e-16 ***
Residuals   179976  1594464       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 5.7165e+13 9.5275e+12   80238 < 2.2e-16 ***
Residuals   179976 2.1370e+13 1.1874e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 29 - MANOVA

It is apparent from the output displayed that for both tests the p value is 2.2e-16, which is very low, meaning that we reject both null hypothesis and the conclusion is made that class does have a large influence on the price and delivery times of the sale.

Below the mean values of each class is displayed as well as the overall distribution of each class. It is very clear that there are large variety in the price and delivery times of the classes and that classes like luxury, technology, and household have the widest spread of values.

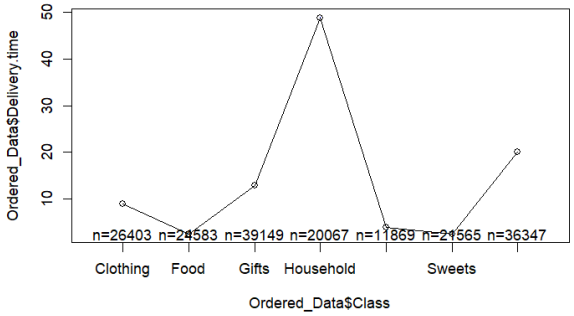


Figure 31 - Mean Delivery time per class

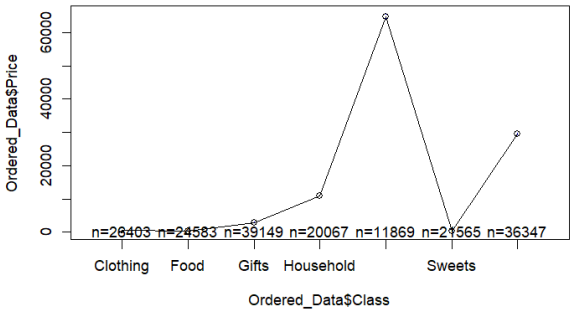


Figure 30 - Mean Price per Class

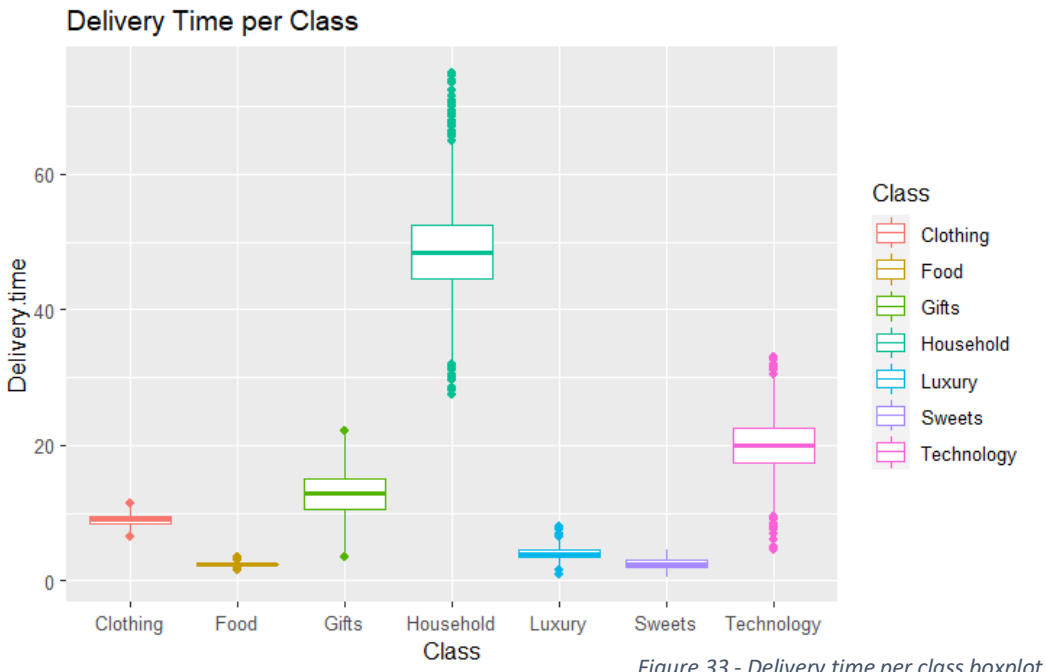


Figure 33 - Delivery time per class boxplot

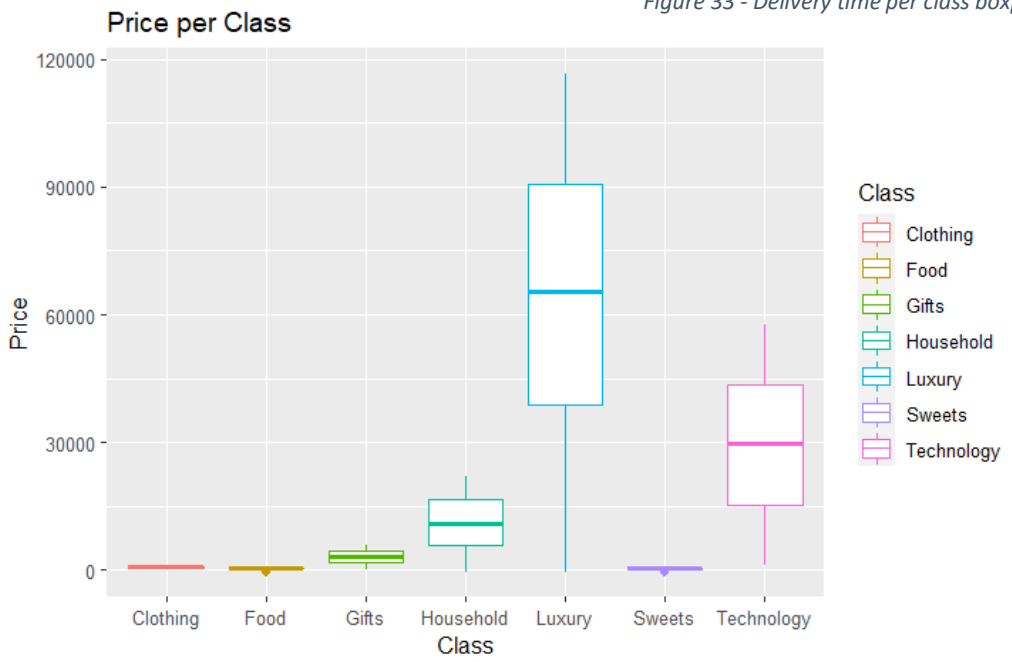


Figure 32 - Price per class boxplot

Reliability of the service and products (Part Six)

Problem 6:

Taguchi Loss Function: $L = k(y-m)^2$

Constant K: $45 = k(0.04)^2$

$$k = 28125$$

Thus: $L = 28125(y-0.06)^2$

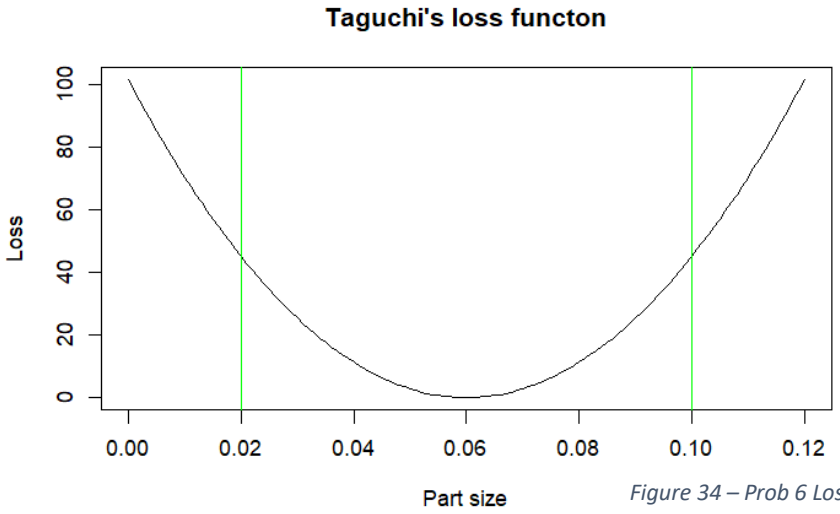


Figure 34 – Prob 6 Loss Function

Problem 7 (a):

Constant k: $35 = k(0.04)^2$

$$k = 21875$$

Thus: $L = 21875(y-0.06)^2$

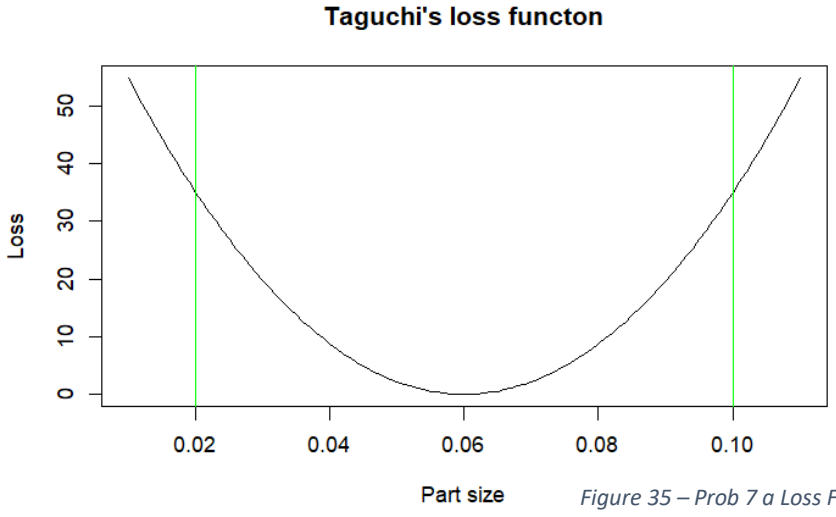


Figure 35 – Prob 7 a Loss Function

Problem 7 (b):

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

$$= 15.946875$$

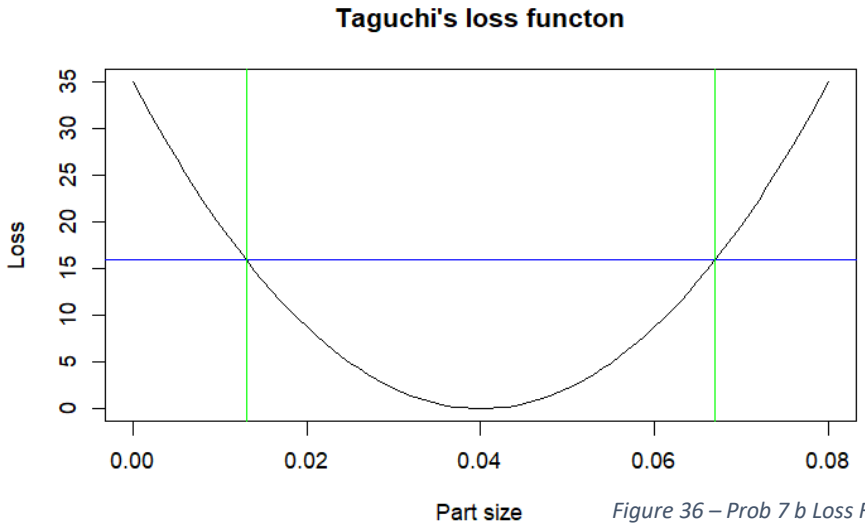


Figure 36 – Prob 7 b Loss Function

6.2) Problem 27

1 Machine at each station:

$$reliability = r(A) * r(B) * r(C)$$

$$reliability = 0.85 * 0.92 * 0.9 = 0.7038$$

$$= 70.38\%$$

2 Machines at each station:

$$reliability(A) = (1 - (1 - r(A))^2$$

$$reliability = \ (1 - (1 - r(0.85))^2) * (1 - (1 - r(0.92))^2) * (1 - (1 - r(0.9))^2)$$

$$= 96.15\%$$

6.3) Delivery Process

Given the poor reliability of the business on some days and the good reliability on other days it is difficult to calculate the probabilities and therefore the problem will be handled according to binomial probabilities. Firstly, the number of days more than 16 available vehicles was determined and calculate the binomial probability accordingly, with the help of the brute force method.

Days = 1560 – 190 – 22 – 3 – 1 = 1344

The result of calculating the binomial probability in r of the vehicles is **0.7031868**

The result of calculating the binomial probability of the drivers in r is **0.9553064**

To get the reliable days per year we simply multiply (0.7031868 * 0.9553064) * 365

To get **245.192**, which means that the company will have 245 reliable days with 20 vehicles and 21 drivers.

With 21 Vehicles:

With another vehicle added we calculate the days again and thereafter calculate the binomial probability of the vehicles in r to get **0.8445316**

Once again to get the number of reliable days we multiply (0.8445316 * 0.9553064) * 365

To get **294.477**, which implies that there will be 294 reliable days with the added vehicle for the company

Conclusion

Now that all the data of the company has been sorted and processed, we can use it to make our own assumptions and to evaluate the performance of the company. Many different plots and charts give insight to the business and allows the user to better understand the sales data provided. The overall reliability of the company is good but has room for improvement. The delivery times and Price for the different classes vary quite a lot and needs attention, especially in classes such as Gifts, Household, and Luxury goods. The distribution of sales between the classes is relatively good and positive trends are mostly noticed throughout the company.

References

Stellenbosch, E. C. (2022). *ProjectDescription2022A.pdf.* Engineering Counsel of South Africa.

Marcos, Q. (2022, 10 18). *Six Sigma & SPC Excel Add-in*. Retrieved from Six Sigma & SPC Excel Add-in: https://www.qimacros.com/free-excel-tips/control-chart-limits/index2.php

Reliability, I. o. (2022, 10 17). *Control Chart Constants and Formulae-1.pdf.* Retrieved from Control Chart Constants and Formulae-1.pdf: https://web.mit.edu/2.810/www/files/readings/ControlChartConstantsAndFormulae.pdf

Unknown. (2022, 10 19). *ISIXSIGMA*. Retrieved from ISIXSIGMA: https://www.isixsigma.com/tools-templates/control-charts/a-guide-to-control-charts/#:~:text=Control%20limits%20are%20calculated%20by,the%20average)%20for%20the%20LCL