2022

ESCA Project

PD SWART 23548258

Table of Contents

Introduction

The report will analyse the sales data of a given business. Then the valid data will be extracted and analysed. This process, data wrangling, will be completed to ensure accurate results from the data analysis. The valid data will then be visualized and studied to gain a better understanding of how the data is spread and its distribution. Statistical process control is the next step that will be done to by the setup of s- and x-charts. Then finally process control is done to determine if the process is controlled.

Part 1: Data Wrangling

The original dataset has many invalid instances that need to be removed. The original dataset is split into 2 datasets, valid-data and invalid-data. It was discovered that there is 17 missing values in the Price feature. These missing values was removed from the original dataset and moved to the in-valid dataset. There are also 5 negative values in the original dataset in the Price feature that will lead to an incorrect analysis and therefore moved to the in-valid dataset. After all the in-valid instances were removed, the original dataset was renamed to the valid-dataset and consists of 179 978 instances and the in-valid dataset consists of 22 instances. Both new datasets consists with 11 variables which is the original 10 variables and a new variable of the index in the original dataset.

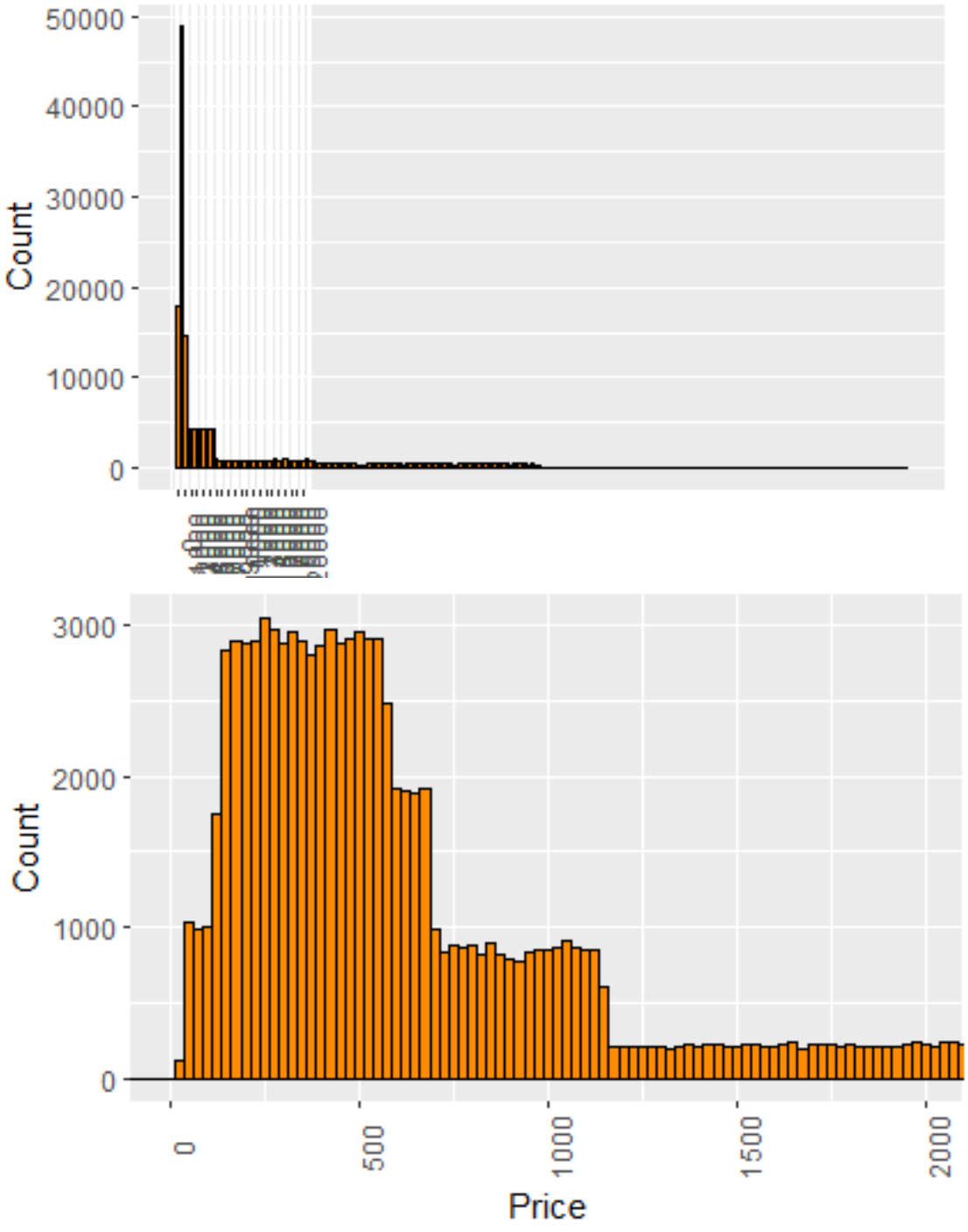
Part 2: Descriptive Statistics

Analysis of Valid Dataset

The valid dataset consists of 179 978 instances and 11 features. There are no missing values in the dataset because it was removed in the previous step. There are nine numeric features and two-character features. There are 3 categorical features which include, Class, Year and Why.Bought, therefore they have a low cardinality. The rest are all continuous features.

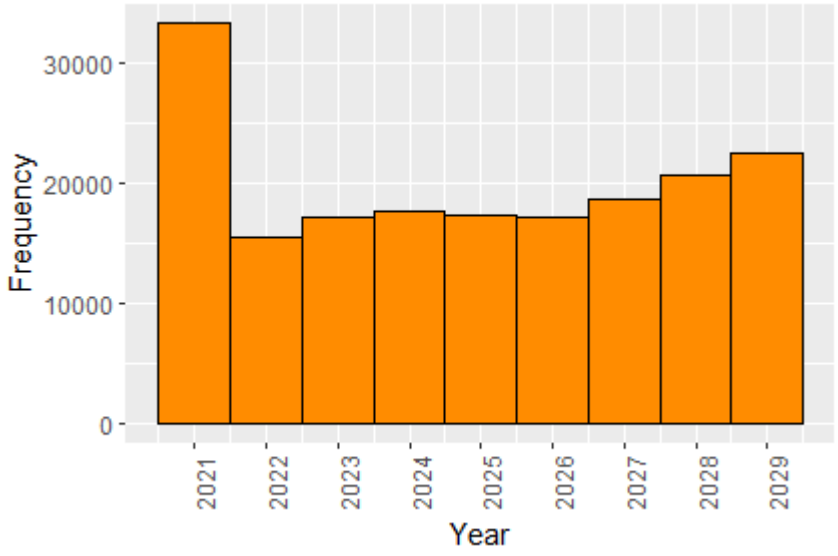
Visualisation of Features

The following graphs indicate the distribution of the Price feature. The code shows that there are 47 122 instances below 500. There are also 75 196 instances below 1000 and 113 332 instances below 5000. With a total of 179 978 instances, it is clear according to the filter code and the histograms generated, that the data is skewed to the right. This indicate that most sales had very low totals, with a few totals adding up to a big sum. The second histogram indicates the distribution of Price between 0 and 2000, indicating this skewed distribution and the left side of the first graph.

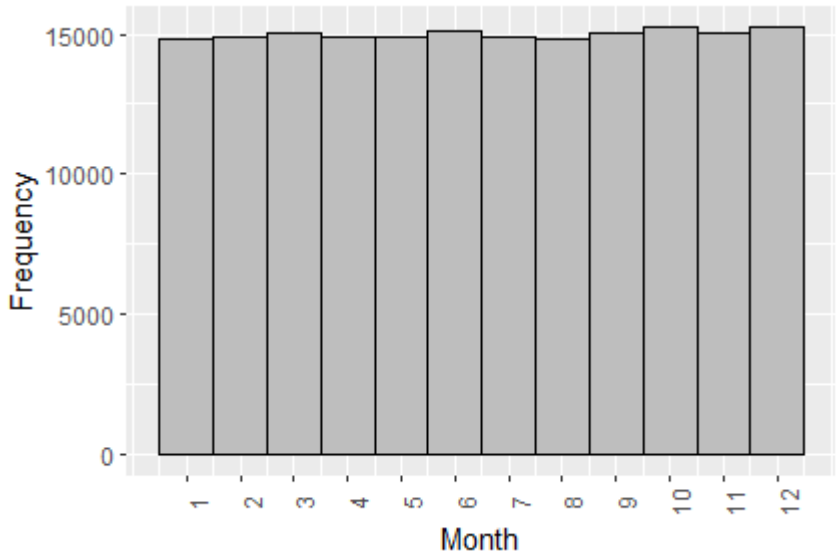


This Histogram indicates that the highest delivery time is between 3-5 days. The distribution is skewed to the right with most delivery time under 25. The is a normal distribution between 30 and 70 in the distribution. The conclusion is drawn that most delivery times are relatively short.

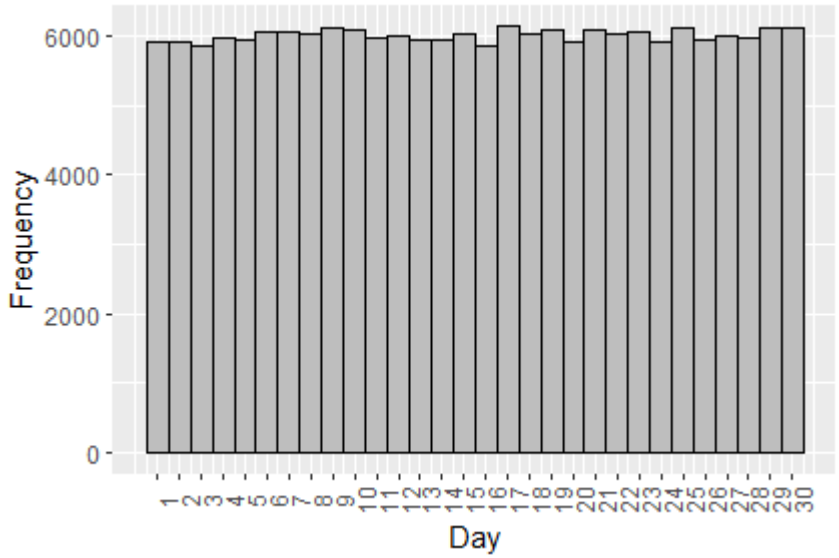
According to the histogram, the distribution is skewed to the right with a mean of 54.56552 and a range of [18,108]. The conclusion is drawn that the sales are generated mostly by the younger generation and steadily decline are age increase.



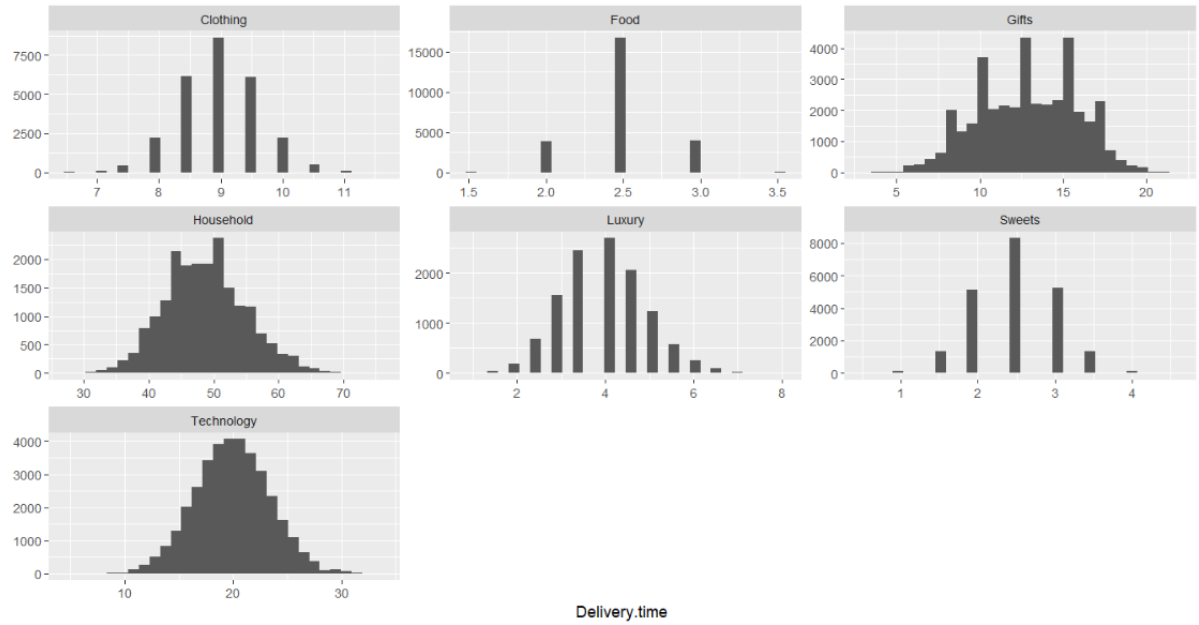
This histogram shows that 2021 had the most sales and the following year, 2022 had the least. Then the sales gradually fluctuated and increased up to 2029.



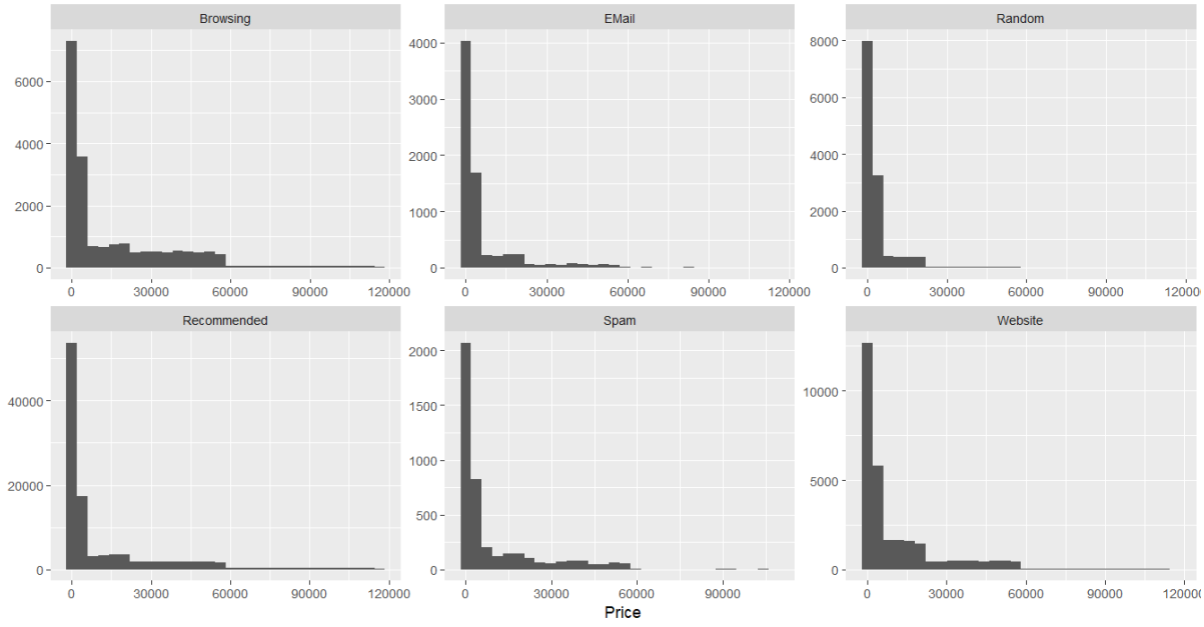
This histogram indicates that there is no real fluctuation in sales per month. A uniform distribution is followed from January to December.



The histogram indicates no significant fluctuation in days of the month. A uniform distribution is followed from day 1 of each month to day 30.



In these graphs it is evident that all delivery times in each class follows a normal distribution. Clothing, food, luxury and sweets does not follow a continuous distribution as these were only delivery on in certain set out schedules. Gifts and household also have a few values that does not follow a normal distribution. There is a clear distinction that household have the highest average delivery time, technology and gifts follow with half that of household. The rest is distributed at a lower range. This makes sense as household is probably bigger items requiring more admin and bigger distributers to be delivered.



The graphs show the distribution of prices of sales vs the reason it was purchased. All the reasons why items were bought follows the same distribution. This shows that most reasons contribute the same general average amount per sale.

Process Capabilities Calculations

USL = 24

LSL = 0

It is logical that LSL is zero because the lower specification cannot be negative since delivery time cannot be negative and zero is the lowest limit.

Standard deviation = 3.501873

Mean = 20.01106

Cp = (USL – LSL) / (6 x standard deviation) = 1.14224574

Cpu = (USL – Mean) / (3 x standard deviation) = 0.37969589

Cpl = (Mean – LSL) / (3 x standard deviation) = 1.90479559

Cpk = Min (Cpu, Cpl) = 0.37969589

The Cp value of over 1 means that the process is capable of meeting the specifications. It means that the delivery times of technology will be able to be met and products delivered on time, within the specifications required. A Cpk value of 0.3797 is not desirable within the reliability process as the process is not within the targets and not centred.

Part 3: Statistical Process Control

3.1 First 30 Samples

X-Chart Limits

The x-chart indicates how the average or mean change over time.

	Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
1	Technology	22.737174	21.799968	20.862762	19.925556	18.988349	18.051143	17.113937
2	Clothing	9.479812	9.326912	9.174011	9.021111	8.868211	8.715310	8.562410
3	Household	54.126083	52.486648	50.847213	49.207778	47.568343	45.928908	44.289473
4	Luxury	4.589988	4.369621	4.149255	3.928889	3.708523	3.488156	3.267790
5	Food	2.712546	2.642068	2.571589	2.501111	2.430633	2.360154	2.289676
6	Gifts	15.281586	14.533650	13.785714	13.037778	12.289842	11.541906	10.793970
7	Sweets	2.921584	2.788093	2.654602	2.521111	2.387620	2.254129	2.120638

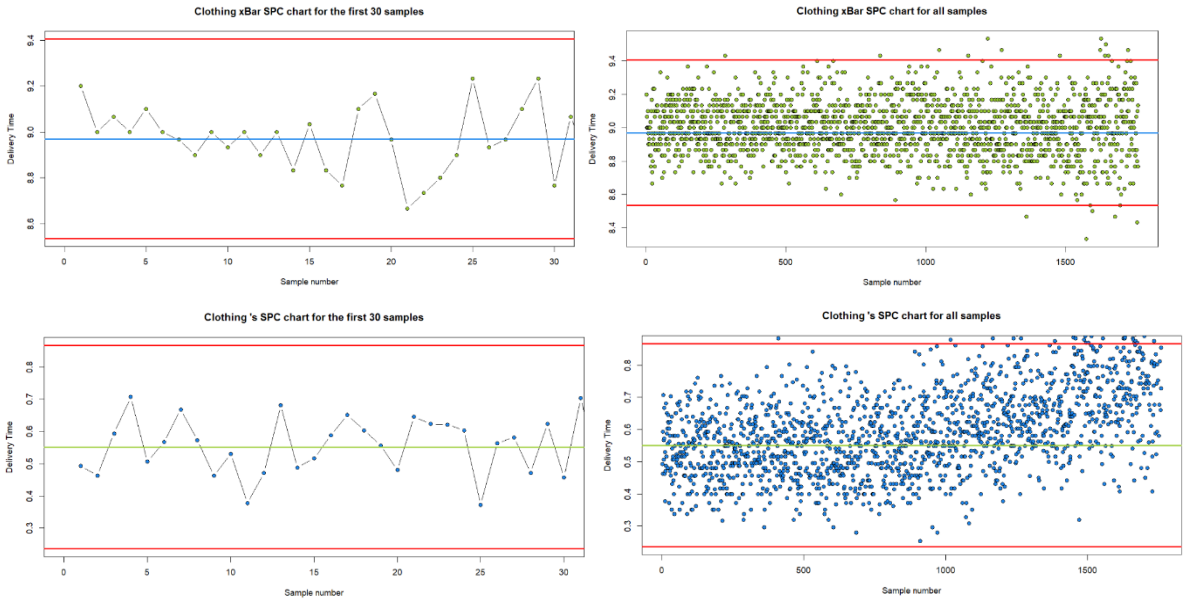
S-Chart Limits

The s-chart indicates the standard deviation of a process over time.

	Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
1	Technology	5.6018567	4.9224118	4.2429670	3.5635221	2.8840772	2.2046323	1.5251874
2	Clothing	0.9139138	0.8030659	0.6922180	0.5813701	0.4705222	0.3596743	0.2488264
3	Household	9.7992085	8.6106700	7.4221316	6.2335932	5.0450547	3.8565163	2.6679779
4	Luxury	1.3171699	1.1574113	0.9976528	0.8378943	0.6781358	0.5183773	0.3586188
5	Food	0.4212624	0.3701678	0.3190732	0.2679786	0.2168840	0.1657895	0.1146949
6	Gifts	4.4705527	3.9283228	3.3860929	2.8438630	2.3016331	1.7594033	1.2171734
7	Sweets	0.7979005	0.7011238	0.6043470	0.5075703	0.4107936	0.3140168	0.2172401

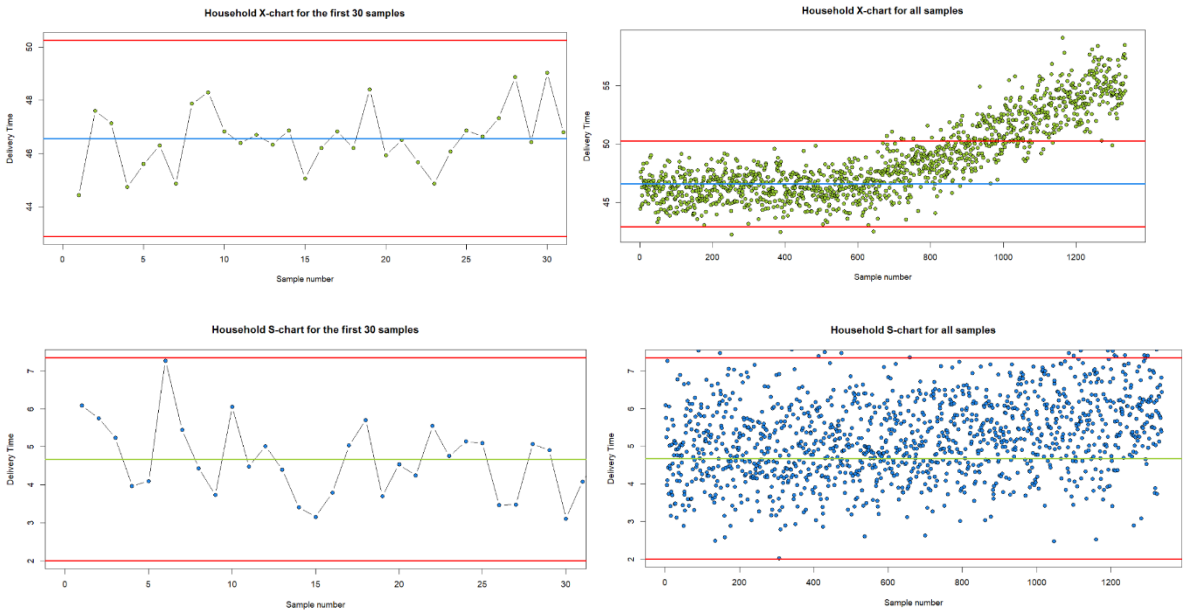
Process Control Charts

Clothing:



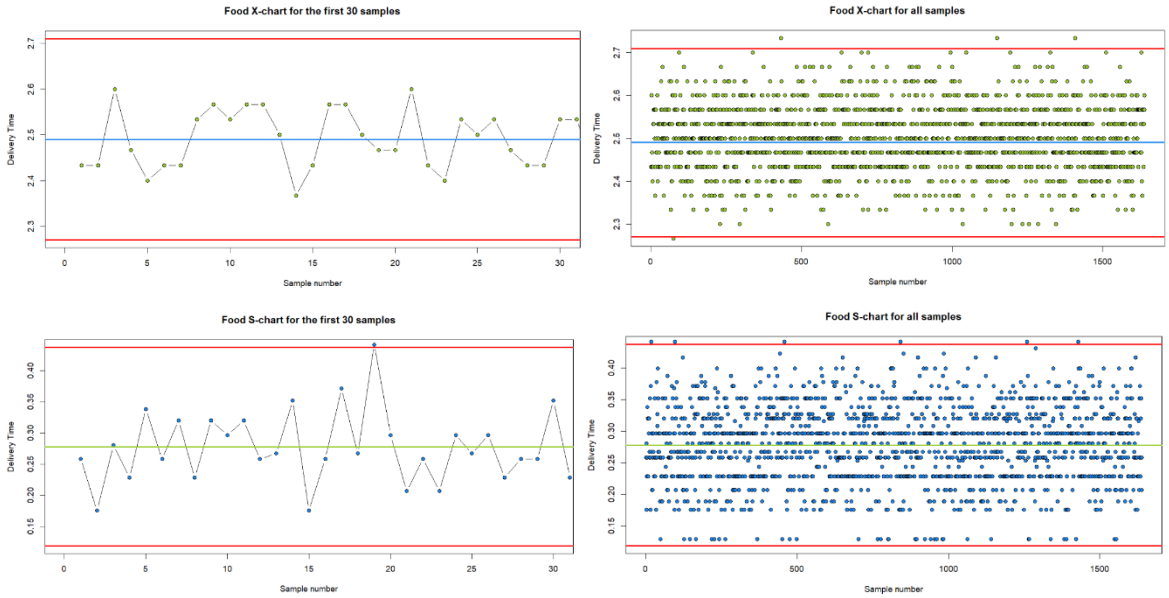
No instances are outside the limits on the x-chart and s-chart in the first 30 samples. There are instances outside the limits on both x-chart and s-chart for all instances. This in indicated by the dots outside of the red lines on all sample charts.

Household:



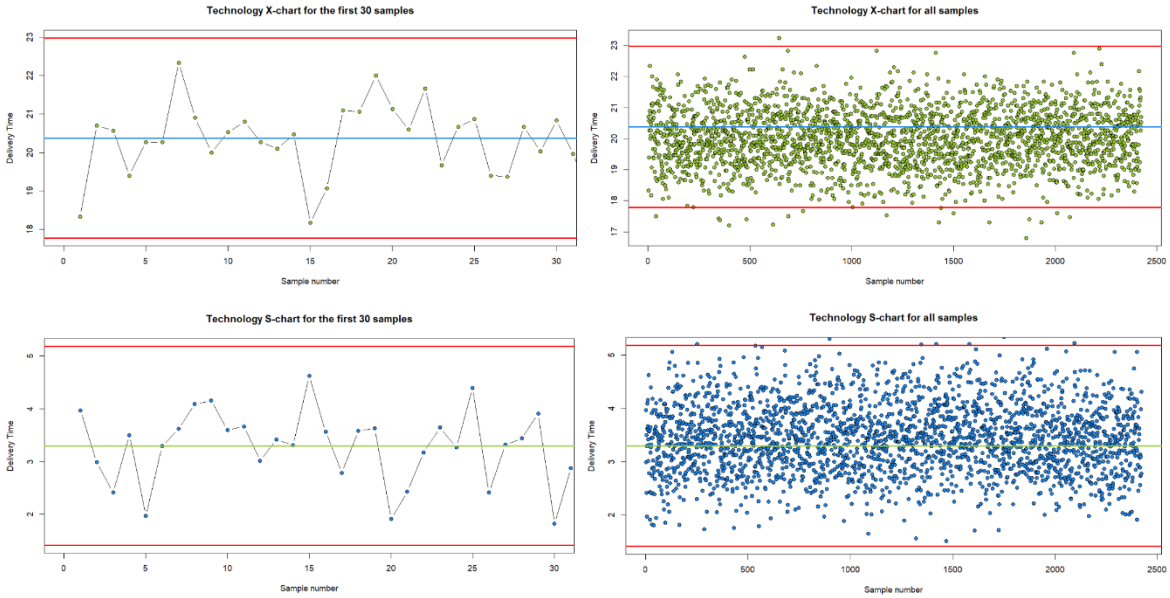
No instances are outside the limits on the x-chart for the first 30 samples but the trend is steadily increasing and breaking the limits in all the samples. The s-chart is on the border of the limits with the first 30 samples and there are many instances crossing the limits in all the samples.

Food:



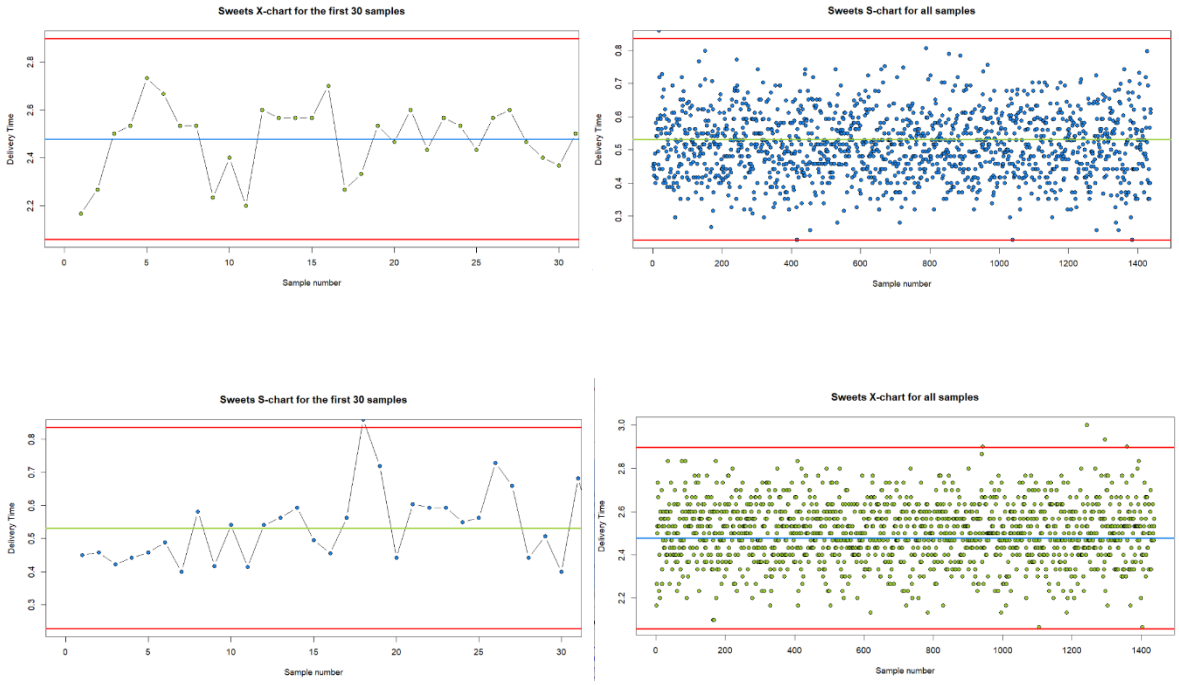
No instances are outside the limits on the x-chart first 30 samples and some instances are outside on the x-chart all samples. There are also a instance outside the limits for s-chart first 30 samples and few outside for the s-chart all samples.

Technology:



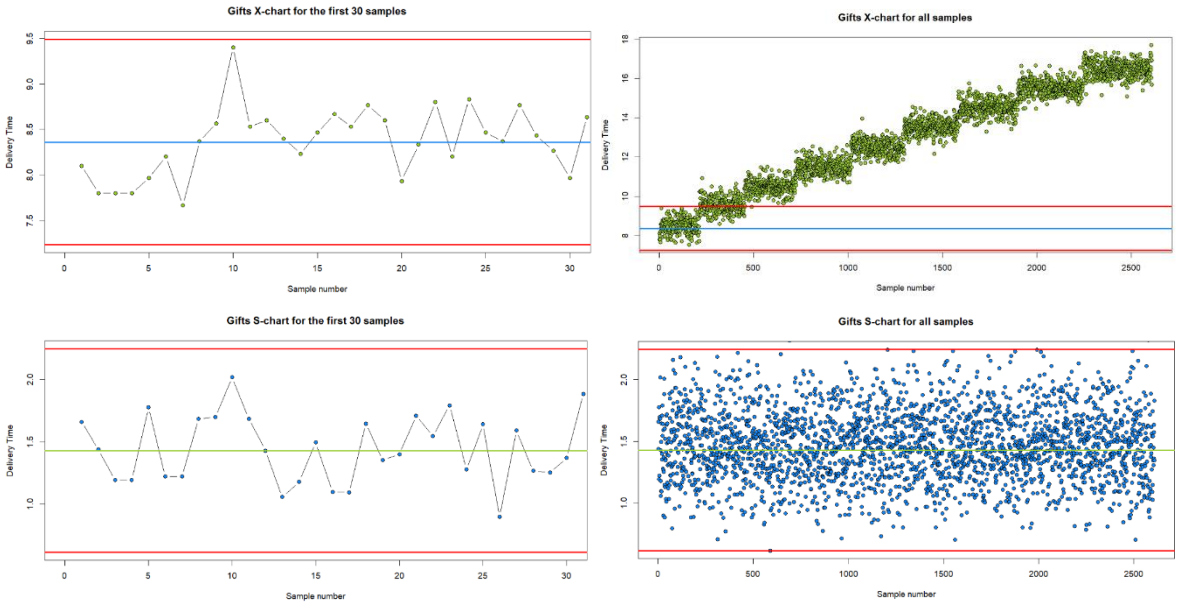
No instances are outside the limits on the x-chart and s-chart for the first 30 samples. Some instances are outside on both x-chart and s-charts for all the samples.

Sweets:



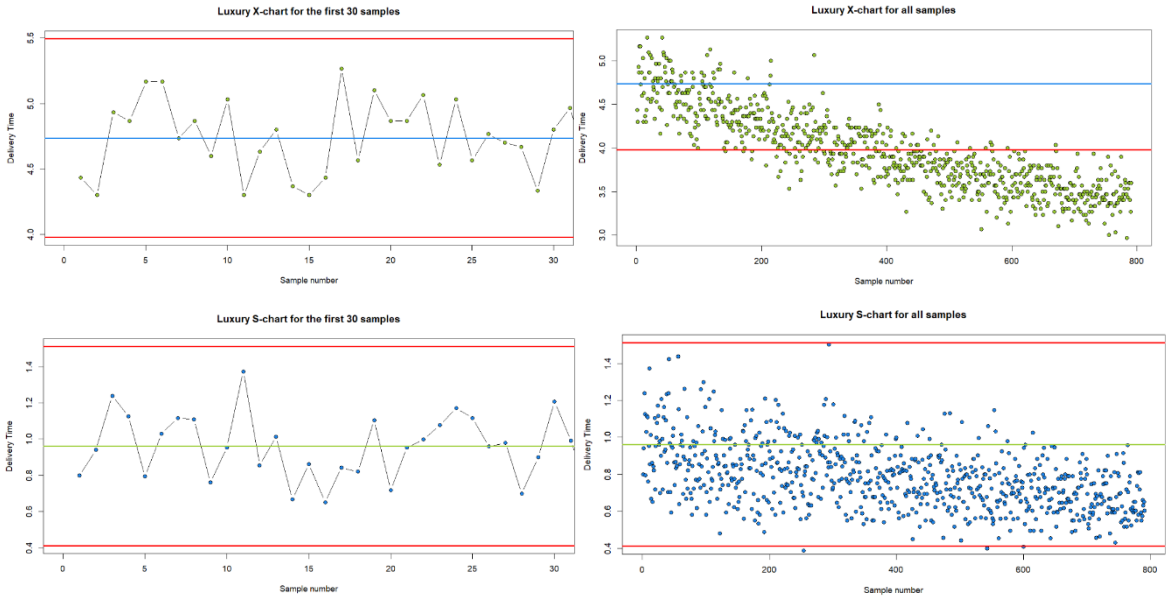
No instances are outside the limits on the x-chart for the first 30 samples and some instances are outside on both s-charts and the x-chart for all samples.

Gifts:



No instances are outside the limits on the x-chart and s-chart for the first 30 samples. Some instances are outside on the s-chart for all samples. There are a few instances outside on the x-chart all samples with a steady increasing trend being evident. This could be caused by seasonality or a purposeful increase by the company.

Luxury:



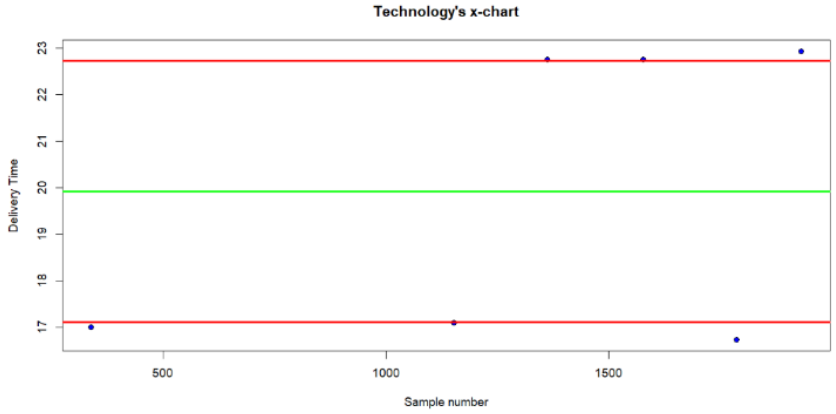
No instances are outside the limits on the x-chart and the s-chart for the first 30 samples. Some instances are outside on the s-chart for all samples. There are a few instances outside the limits on the x-chart for all samples with a steadily decreasing trend being clear.

Part 4: Optimizing the Delivery Process

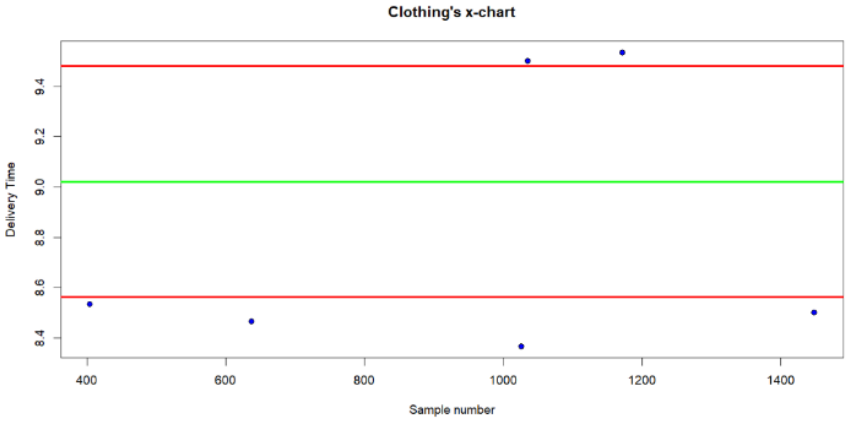
4.1 A. X-Bar Plot sample outside of outer control limit of the sample

A sample size of 15 was chosen from the instances outside of the limits and these graphs indicate to which classes the instances belong.

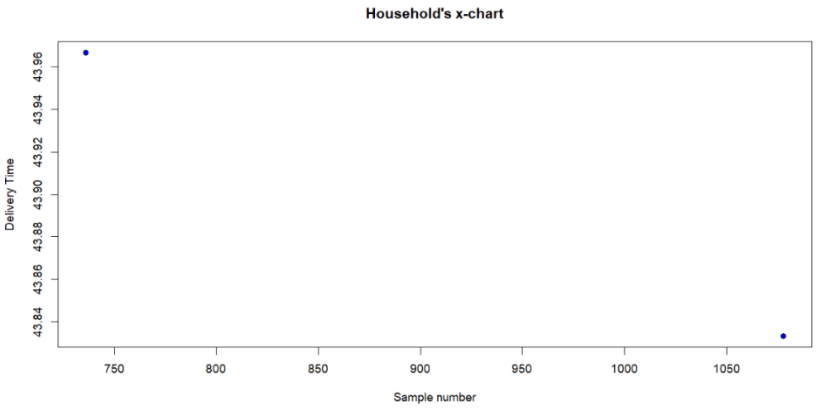
Technology: 6 instances



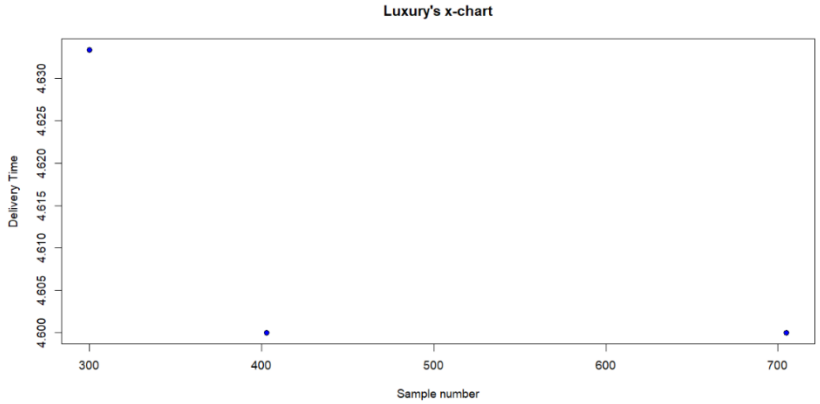
Clothing: 6 instances



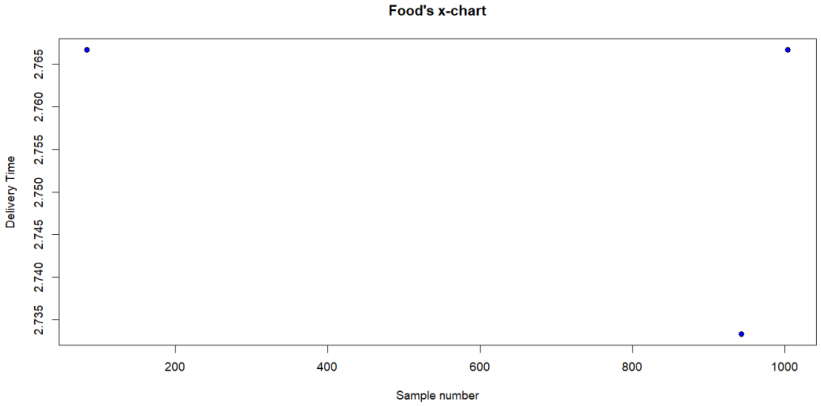
Household: 2 instances



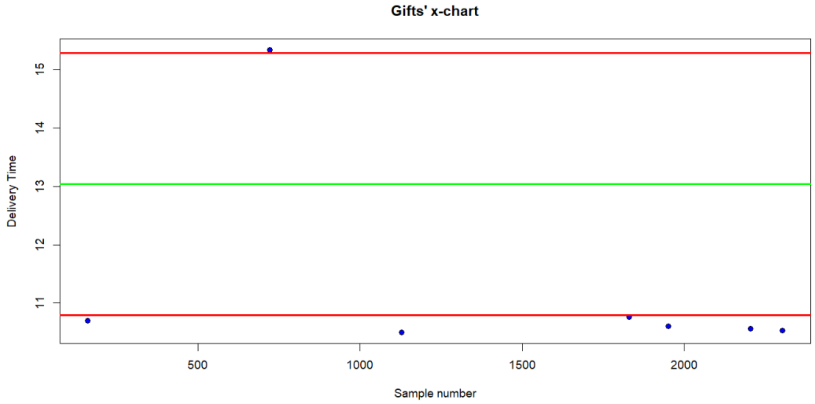
Luxury: 3 instances



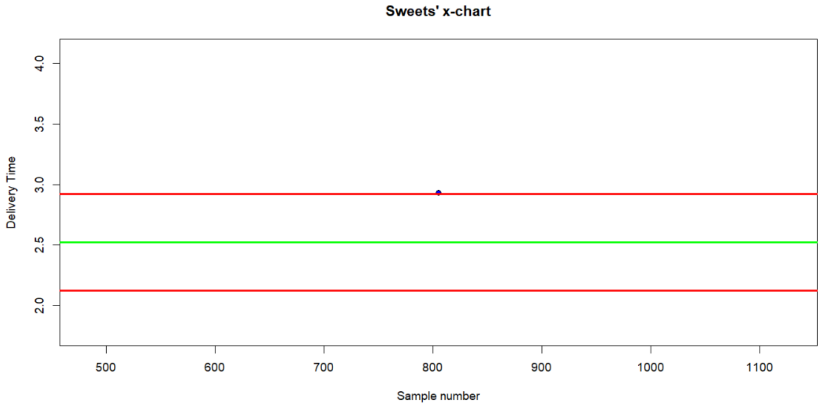
Food: 3 instances



Gifts: 7 instances



Sweets: 1 instance

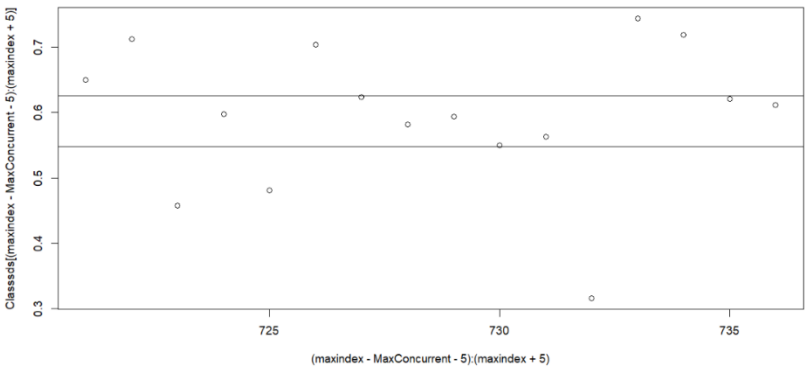


It is evident that Technology, Clothing and Gifts have to most instances outside the limits. This is only a sample of 15 and therefore should not be taken as a precise representation of the data, merely a exploration of the data.

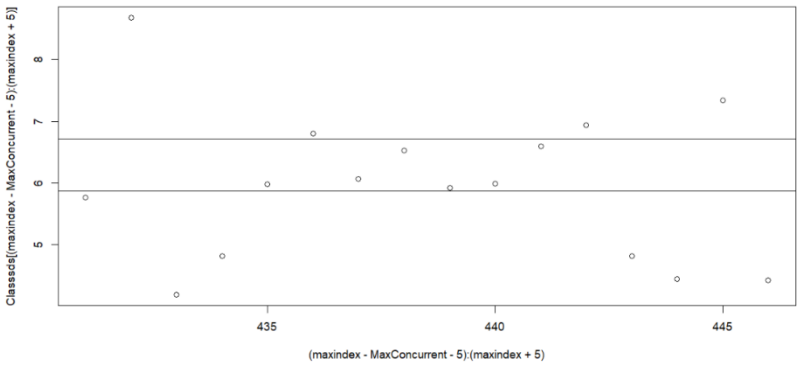
4.1 B. Most Consecutive Samples in Range

The range presented is between -0.3 and +0.4 sigma-control limits.

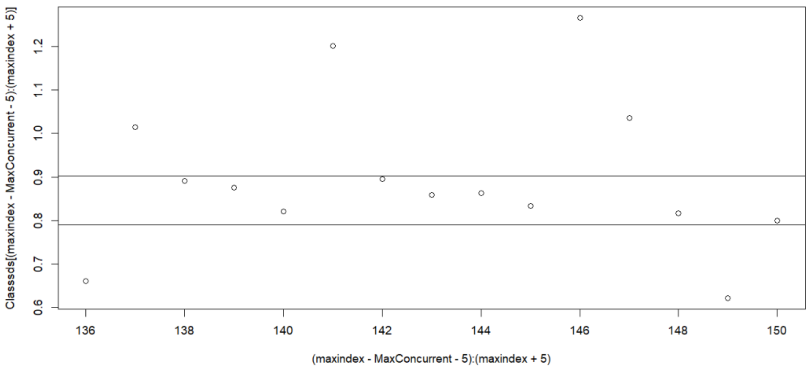
The longest range of instances within the range for clothing is 5731.



The longest range of instances within the range for household is 5441.



The longest rang of instances within the range for luxury is 4145.



4.2 Likelihood of Type l Error

Making a statistical decision always includes uncertainty. A type l error is a false positive conclusion. (Bhandari, 2021)

Assumptions made during calculations:

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

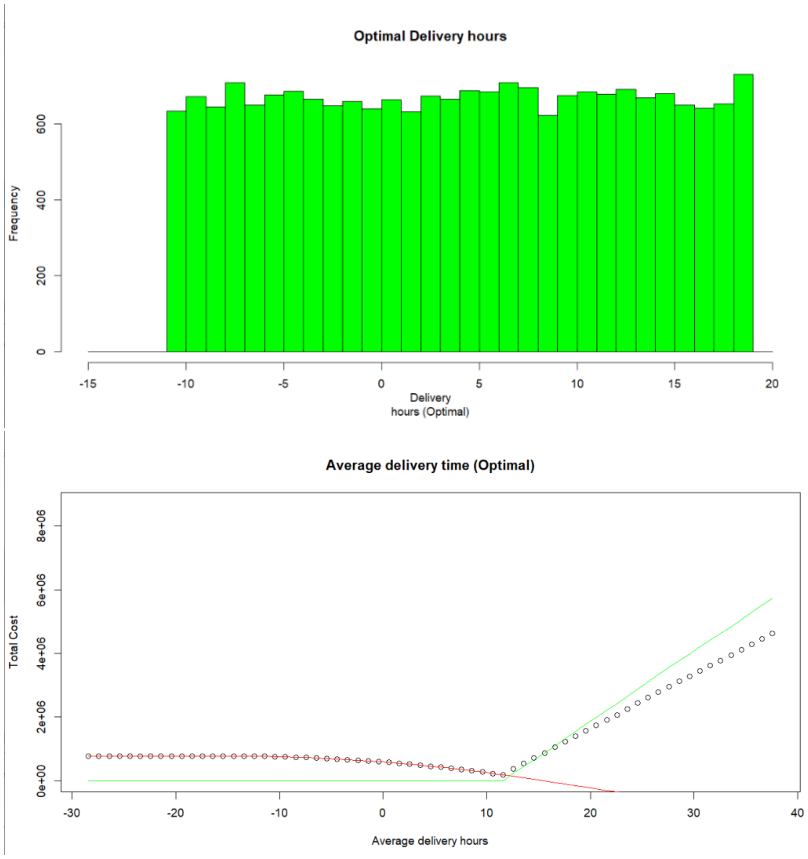
Ha = the process is not in control and has moved from the centreline or has increased or decreased in variation

The probability of a Type l error for A is 0.002699796. The probability of performing a false positive in case A is 0.27%.

The probability of a Type l error for B is 0.7266668. The probability of performing a false positive is case B is 72.66%.

4.3 Minimize Delivery Cost

The company lose R329/item-late-hour in sales of technology and it costs R2.5/item/hour to reduce the average time by one hour after 26 hours. To determine the minimal delivery cost for technology the costs were compared of the different days and the lowest cost were selected. There was looped through all the possible delivery times and reduction costs related to these items.



The sum was calculated and the optimal time for delivery is 11.56756 days. That is the delivery time the data should be centred around. The second graph shows the loops and where the dots are at its lowest, thus having the lowest total cost is the optimal amount of delivery time.

4.4 Likelihood of Type ll Error

Type ll error is a false negative conclusion. It will occur if the company fail think a parcel is delivered but it is not. (Bhandari, 2021)

Assume:

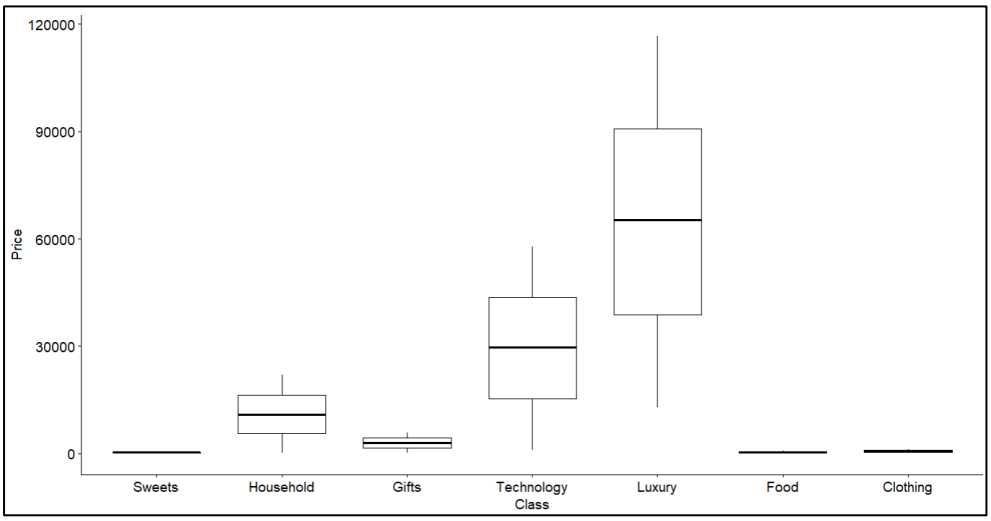
Ha = true

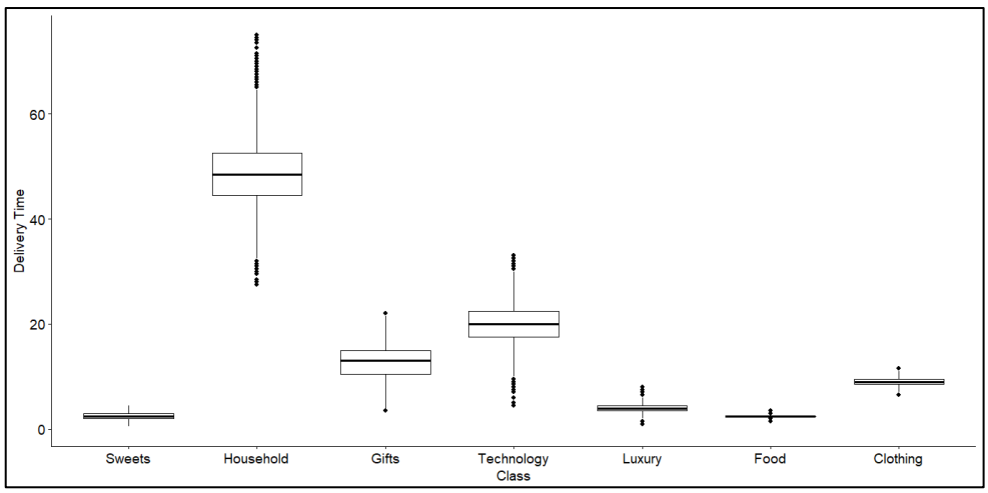
The probability of making a Type ll error in case A is 0.3895719. The probability of performing a false negative conclusion in case A is 38.96%.

Part 5: Doe and MANOVA

MANOVA is the multivariate analysis of variance. It can only be used in certain conditions, if dependant variable should be normally distributed within a group, there should be homogeneity of variances across the range of predictors and linearity across all pairs of dependant variables.

The hypotheses made was that between Price and Delivery time there is no significant difference when looking in Class. The graphs drawn up indicate the relationship between Price and Class in the first graph and the relationship between Delivery time and Class in the second graph.





The graphs indicate that there is no significant correlation between Price and Delivery time, regarding Class and the differences are indicated. The hypotheses are denied and proven to be invalid. This indicates that both Delivery time and Price is important features regarding service and reliability.

Part 6: Reliability of service and products

6.1

Problem 6:

$$L(X) = k(x - T)^2$$

$$L(X) = 45$$

$$T = 0.06$$

$$45 = k(0.06 + 0.04 - 0.06)^2$$

The Taguchi loss function is: $L(X) = 28125(x - 0.06)^2$

Problem 7:

a.

$$L(X) = k(x - T)^2$$

$$35 = k(0.06 + 0.04 - 0.06)^2$$

$$k = 21875$$

The Taguchi loss function is: $L(X) = 21875(x - 0.06)^2$

b.

$$L(X) = k(x - T)^2$$

$$L(0.027) = 21895(0.027 - 0.06)^2 = \$23.82 \text{ per item}$$

6.2

Problem 27:

a.

$$R_{SERIES} = R_A \times R_B \times R_C$$

$$R_{SERIES} = 0.85 \times 0.92 \times 0.9 = 0.7038 = 70.38\% \text{ reliability}$$

b.

$R_{PARALLEL} = [1 - (1 - R_A)(1 - R_A)] \times [1 - (1 - R_B)(1 - R_B)] \times [1 - (1 - R_C)(1 - R_C)] = 0.9015 = 96.15\%$ reliability

Therefore running 2 machines at the same time will increase the reliability with 25.77%. The best decision will be to run 2 machines in parallel, increasing the reliability and effect of breakdowns.

6.3

Using the Binomial function there was determined that the probability of failure is 0.01383834. The code was run using the dbinom function in R and determined that 20 vehicles being available was calculated to be 294.964 days. The same calculation was run while using 21 vehicles would be available 346.7182 days in the whole year.

The expectation was that the reliability would increase with an increase in vehicles, the effect was a difference of 51.7542 days and with only 18.2818 days with reliability issues. This indicates it would be worth it to attain another vehicle to increase reliability.

Conclusion

The dataset given was wrangled into a new dataset to ensure no invalid data was present when the analysis process were started. The new dataset was analysed and relevant information were presented to be used for optimisation and improvements. The x- and s-charts were drawn for the data and the relevant limits obtained to make these required improvements. The probability of the different types of errors that can be made during calculation were calculated. The purpose of the whole process was to determine the optimal delivery time to maximize profit and customer satisfaction.

References

Bhandari, P. (2021). *Type I & Type II Errors | Differences, Examples, Visualizations*. Retrieved from Scribbr: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/