ECSA Graduate Attributes Report

by

CF SWART

22791027

ECSA project

Quality Assurance 344

Department of Industrial Engineering

21 October 2022

ABSTRACT

This report explores the Graduate Attribute Report set out to be approved by The Engineering Council of South Africa (ECSA). Industrial Engineering 3rd year students at Stellenbosch University are required to obtain a mark of at least 50% for this project, as a pre-requisite to pass the module Quality Assurance. Quality Assurance is a module focussed on improving processes to best satisfy customers and increase the return generated by a business. It is crucial Industrial Engineering students graduate with adequate knowledge in this area of expertise, as optimization of processes are sacrosanct to the work of an Industrial Engineer in industry. This project ensures competent graduates enter the workforce after graduation and can apply the skills learnt about process optimization in a practical setting.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

Symbols	
H0	Null Hypothesis
HA	Alternative Hypothesis
sd	Standard Deviation
Acronyms	
CL	Centre Limit
ECSA	Engineering Council of South Africa
LCL	Lower Control Limit
UCL	Upper Control Limit

1 INTRODUCTION

The Engineering Council of South Africa requires Industrial Engineering students to have completed this report to expand their knowledge on process control and optimization. Stellenbosch University has incorporated this graduate attribute report into the module Quality Assurance, which focuses on process control.

This report allows students to deal with a large dataset containing several features over a long time period. This dataset is used to gain understanding into the relationships between features and how to analyse these features and their relationships. Using the knowledge learnt on the relationships business decisions can be made in order to increase reliability, service and revenue.

In this report the data is firstly filtered to discard all invalid data that is unusable. The dataset containing only valid entries is utilized along with descriptive statistics tools to better understand the features in the provided dataset and their interconnected relationships. The Process Capability Indices for the valid data are also determined. Next the statistical process control is performed. This is accomplished by means of sampling the data and constructing respective X and S Charts that are then analysed. Optimizing the process occurs next within the boundaries specified and the errors associated within the given rules are calculated. The function MANOVA in R is then used to confirm or reject the hypothesis based on discovered relationships between features. Lastly the reliability of services and products are determined in order to understand the effect of reliability on a process.

2 DATA WRANGLING

Data wrangling was performed on the salesTable2022 dataset provided. The purpose of data wrangling was to remove all data quality issues from the dataset that made it challenging to work with and interpret. Once the data has been wrangled, it can be used to obtain accurate and comprehensive graphs and information. Initially the dataset was not ordered, this can be seen in the snippet of the first 11 rows of the salesTable2022 in Figure 1 below.

	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	1	19966	54	Sweets	246.21	2021	7	3	1.5	Recommended
2	2	34006	36	Household	1708.21	2026	4	1	58.5	Website
3	3	62566	41	Gifts	4050.53	2027	8	10	15.5	Recommended
4	4	70731	48	Technology	41843.21	2029	10	22	27.0	Recommended
5	5	92178	76	Household	19215.01	2027	11	26	61.5	Recommended
6	6	50586	78	Gifts	4929.82	2027	4	24	14.5	Random
7	7	73419	35	Luxury	108953.53	2029	11	13	4.0	Recommended
8	8	32624	58	Sweets	389.62	2025	7	2	2.0	Recommended
9	9	51401	82	Gifts	3312.11	2025	12	18	12.0	Recommended
10	10	96430	24	Sweets	176.52	2027	11	4	3.0	Recommended
11	11	87530	33	Technology	8515.63	2026	7	15	21.0	Browsing

Figure 1: First few unordered rows of salesTable2022

The dataset was ordered in ascending order by Year, Month, Day and X for consistency and to better understand the salesTable2022 dataset. The ordering of the dataset is illustrated by the first 11 rows of the dataset seen in Figure 2 below.

	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	463	47101	50	Clothing	1030.86	2021	1	1	9.0	Recommended
2	2627	88087	21	Clothing	428.03	2021	1	1	10.0	Recommended
3	3374	25418	68	Household	13184.41	2021	1	1	48.5	Website
4	5288	13566	94	Household	7021.90	2021	1	1	42.0	Recommended
5	8182	84692	35	Clothing	475.18	2021	1	1	9.0	Recommended
6	9272	46305	72	Clothing	580.98	2021	1	1	8.5	Random
7	9712	92105	45	Household	6877.00	2021	1	1	43.0	Recommended
8	12163	21614	27	Clothing	513.13	2021	1	1	9.5	Recommended
9	12195	12174	56	Household	14538.64	2021	1	1	41.5	EMail
10	20004	84558	74	Food	255.41	2021	1	1	2.0	Recommended
11	20509	15630	32	Clothing	164.56	2021	1	1	9.0	Recommended

Figure 2: Ordered entire salesTable2022

The data in salesTable2022 was then separated into two separate datasets, one Valid and one Invalid.

2.1 Invalid dataset

The data in the salesTable2022 with invalid data needed to be removed and can be viewed in a separate dataset. Invalid data includes data instances that are negative or contain the value ‘NA’. First all the negative values were removed and placed into a separate dataset, there were 5 instances found containing negative values as seen in Figure 3.

	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

Figure 3: Instances containing negative values

Secondly all the instances containing 'NA' values were removed and separated into a separate database. There were 17 instances that contained 'NA' values, the first few rows of this dataset are illustrated below in Figure 4.

	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

Figure 4: Instances containing 'NA'

The two above datasets containing 'NA' and negative entries were combined to form the Invalid dataset. The 'X' column in the original salesTable2022 was converted to the 'Secondary.Key' column in the Invalid dataset. This column is the original index of the instances found in the salesTable2022 dataset. A 'Primary.Key' column was added to the Invalid dataset and this indicates the new index of the instances in the new dataset. When examining the headings for the columns in salesTable2022 the heading 'AGE' is written in upper casing, this is an inconsistency compared to the other headings. In the Invalid dataset this was corrected to allow for consistent heading titles. The first entries of the Invalid dataset with the aforementioned features can be viewed in Figure 5. This dataset shows the data quality issues were found in the 'Price' column and that there are 22 instances that raise issues in total. Finding that the 'Price' column contains all the invalid entries could indicate there is an issue with the price entering system that causes this random variation. It can be seen as random variation, as these invalid entries do not fall on one day or in the same time period.

	Primary.Key	Secondary.Key	ID	Age	Class	Price	Year	Month	Day	Delivery.Time	Why.Bought
16320	1	16320	44142	82	Household	-588.8	2023	10	2	48.0	EMail
19540	2	19540	65689	96	Sweets	-588.8	2028	4	7	3.0	Random
19998	3	19998	68743	45	Household	-588.8	2024	7	16	45.5	Recommended
144443	4	144443	37737	81	Food	-588.8	2022	12	10	2.5	Recommended
155554	5	155554	36599	29	Luxury	-588.8	2026	4	14	3.5	Recommended
12345	6	12345	18973	93	Gifts	NA	2026	6	11	15.5	Website
16321	7	16321	81959	43	Technology	NA	2029	9	6	22.0	Recommended
19541	8	19541	71169	42	Technology	NA	2025	1	19	20.5	Recommended
19999	9	19999	67228	89	Gifts	NA	2026	2	4	15.0	Recommended
23456	10	23456	88622	71	Food	NA	2027	4	18	2.5	Random
34567	11	34567	18748	48	Clothing	NA	2021	4	9	8.0	Recommended

Figure 5: Invalid dataset

2.2 Valid dataset

All the data instances from the salesTable2022 dataset was extracted and placed in the new dataset named 'Valid'. Valid contains all the instances without data quality issues and is therefore data that can be used to extract useful information. This dataset also contains the aforementioned 'Primary.Key' and 'Secondary.Key' with the same purpose as explained before. The 'AGE' heading has again been corrected to match the other headings in the Valid dataset. Out of the 180 000 instances in the salesTable20222 dataset, 179 978 instances were moved to the Valid dataset. This makes logical sense, because 22 of the instances were already identified as problematic and allocated to the Invalid dataset. An extraction from the Valid dataset can be seen below in Figure 6.

	Primary.Key	Secondary.Key	ID	Age	Class	Price	Year	Month	Day	Delivery.Time	Why.Bought
1	1	463	47101	50	Clothing	1030.86	2021	1	1	9.0	Recommended
2	2	2627	88087	21	Clothing	428.03	2021	1	1	10.0	Recommended
3	3	3374	25418	68	Household	13184.41	2021	1	1	48.5	Website
4	4	5288	13566	94	Household	7021.90	2021	1	1	42.0	Recommended
5	5	8182	84692	35	Clothing	475.18	2021	1	1	9.0	Recommended
6	6	9272	46305	72	Clothing	580.98	2021	1	1	8.5	Random
7	7	9712	92105	45	Household	6877.00	2021	1	1	43.0	Recommended
8	8	12163	21614	27	Clothing	513.13	2021	1	1	9.5	Recommended
9	9	12195	12174	56	Household	14538.64	2021	1	1	41.5	EMail
10	10	20004	84558	74	Food	255.41	2021	1	1	2.0	Recommended
11	11	20509	15630	32	Clothing	164.56	2021	1	1	9.0	Recommended

Figure 6: Valid dataset

3 DESCRIPTIVE STATISTICS

Using the Valid dataset descriptive statistic tools were used to discover useful and detailed information on each feature in the dataset. The information gained on the individual features was then used to analyze the features and their relationships with each other.

3.1 Summary

A 5 number summary was performed on the features Age, Price, Year, Month, Day and Delivery Time. These features were selected, as they were continuous features that were expected to have entries that are repeated. This summary gives a general shape of the distribution of the data, the central tendencies, variability and could indicate possible outliers.

Table 1: 5 Number Summary

	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
Age	18.00	38.00	53.00	54.565519	70.00	108
Price	35.65	482.31	2259.63	12294.098366	15270.97	116619
Year	2021.00	2022.00	2025.00	2024.854643	2027.00	2029
Month	1.00	4.00	7.00	6.521064	10.00	12
Day	1.00	8.00	16.00	15.538949	23.00	30
Delivery Time	0.50	3.00	10.00	14.500311	18.50	75

3.2 Continuous features

To understand the distributions of each feature, they were plotted against the count. In doing so an idea of the skewness of the data was developed to see what features need to be investigated further.

3.2.1 Age

In Figure 7 below the distribution of the amount of sales to each age group can be viewed. It is clear from this distribution is skewed right. This is due to the median being further left than for a normal distribution. It can be further observed that therefore more sales are made to individuals below the age of 60, than above the age of 60.

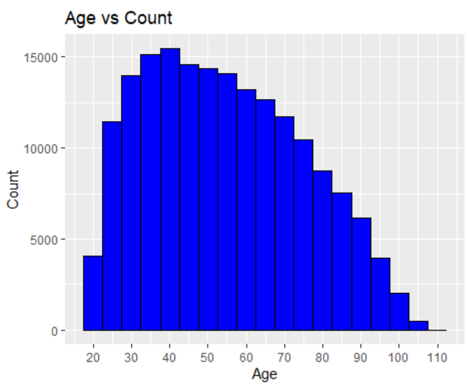

Figure 7: Histogram of Age vs Count

3.2.2 Price

From Figure 8 below it can be observed that the amount of sales decrease drastically when the price increases. There is also a large peak in sales initially then an exponential decline, that later becomes uniformly distributed. This information leads to the understanding that the feature Price will be a feature that needs to be examined further to understand why there is such a rapid decline in sales as the price increases.

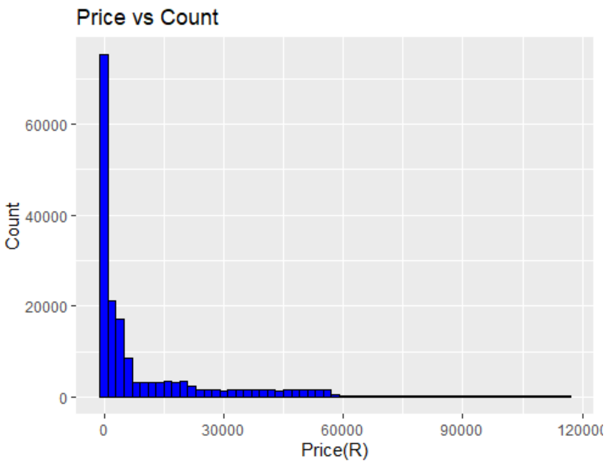

Figure 8: Histogram of Price vs Count

3.2.3 Delivery Time

Below in Figure 9 it can be seen that the distribution is almost bimodal. There is one 'peak' between 0 and 30 as well as another peak between 35 and 65. The reason this distribution is 'almost' bimodal, is due to the fact that there is a large increase in Delivery Times between 0 and 5 followed by a decrease between 5 and 10, slightly altering the normal distribution form this peak should have. From this graph it is clear there are certain Delivery Times that occur more than others. Due to the peak between 35 and 65 being significantly smaller than the first peak, it is a possibility that the company should discontinue such long delivery times and center their process around the 15 hour mark, which is almost center of the initial 'peak'.

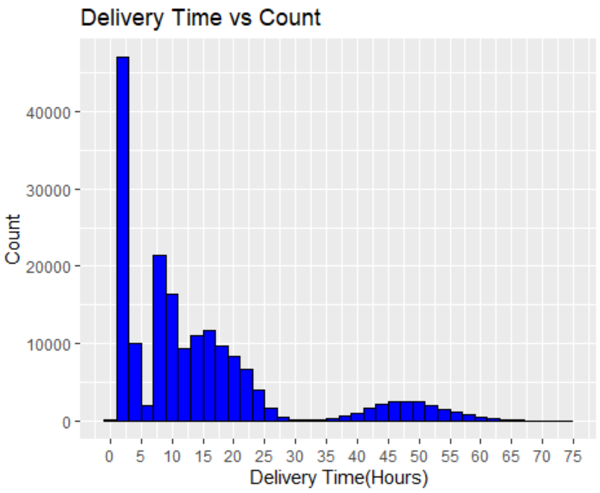

Figure 9: Histogram of Delivery Time vs Count

3.3 Categorical features

To understand the distributions of each feature, they were plotted against the count. In doing so an idea of the skewness of the data was developed to see what features need to be investigated further.

3.3.1 Year

Figure 10 below shows there is a spike in the data in the year 2021, this is also the first year we have data for. A conclusion can be drawn that this is possibly the year the company began, therefore they may have had promotions to draw customers in and this resulted in the spike in sales when compared to the following years. For all the years after 2021 the distribution is slightly increasing, but almost uniform between some adjacent years. A general trend time could be drawn to indicate that after year 2021 sales are increasing linearly, however slowly.

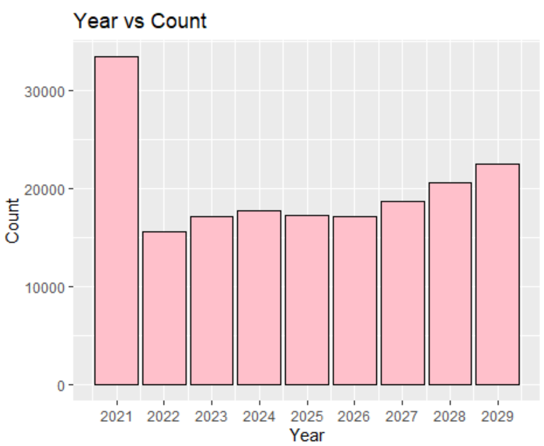

Figure 10: Bar graph of Year vs Count

3.3.2 *Month*

Below in Figure 11 it can be seen that the sales for every month of the year is very uniform. This leads to the conclusion that the company has no seasonality and sales are rather constant throughout the year. From this we understand that the sales of the products do not rely on weather conditions, special holidays or any specific time of year.

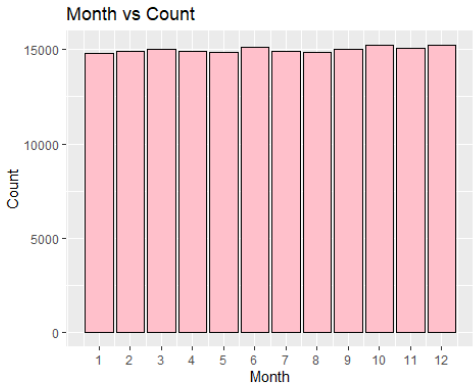

Figure 11: Bar graph of Month vs Count

3.3.3 *Day*

As seen previously with Months, the graph in Figure 12 leads to the conclusion that sales happen regularly and are not dependent on seasons or trends. The uniform distribution seen below shows that sales are constant throughout the 30 days in a month. This trend could be due to the business being an online business. Generally businesses that are not online require people to come to the store to make a purchase and the sale numbers could be significantly more over weekends when less people work. However for an online business sales can be plentiful throughout the week.

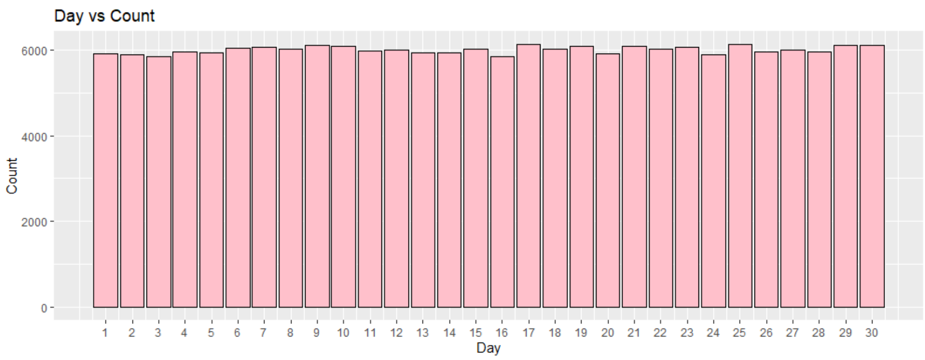

Figure 12: Bar graph of Day vs Count

3.3.4 Class

The below Figure 13 illustrates the sales per Class. From this graph it is clear the Class Gifts generates the most sales, followed by the class Technology and then Clothing. From this the conclusion is draw that the class of sale definitely influences the amount of sales generated. This distribution is not uniform, it is more random and class dependent.

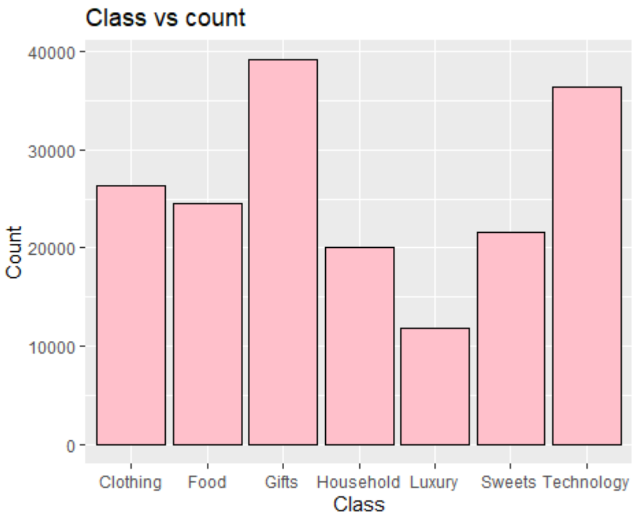

Figure 13: Bar graph of Class vs Count

3.3.5 Why Bought

Figure 14 below details why sales were made and this is classified via the feature 'Why Bought'. From this it is clear that by far the most sales generated are from recommendations. This could mean that the company has a high standard of customer satisfaction and that is why they are readily recommended. The other categories for 'Why Bought' have less sales, but by far the Spam and Email category. This means the company can possibly stop wasting money on emailing potential clients and instead spend money on Website adverts, as this generates the second largest number of sales.

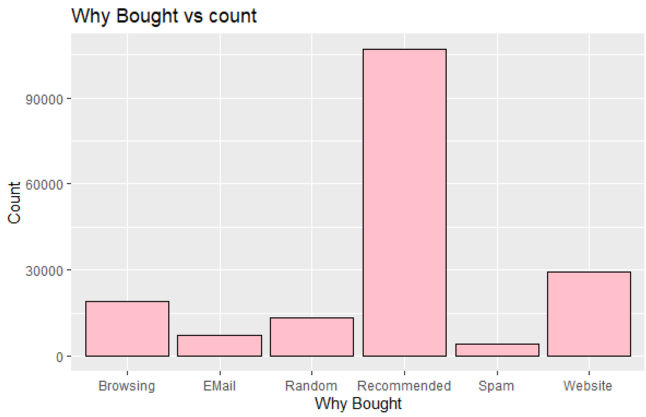

Figure 14: Bar graph of Why Bought vs Count

3.4 Comparisons

3.4.1 Comparing Price, Why Bought and the Count

From Figure 15 below it can be seen that all the data is skewed to the right. This means that no matter the reason customers made the purchase, majority of the customers are paying for cheaper products instead of the most expensive products. There is an exponential decrease in sales, as the price increases for each category of 'Why Bought'. From this it is clear that the reason a purchase is made does not influence the amount of money a customer is willing to spend on a purchase.

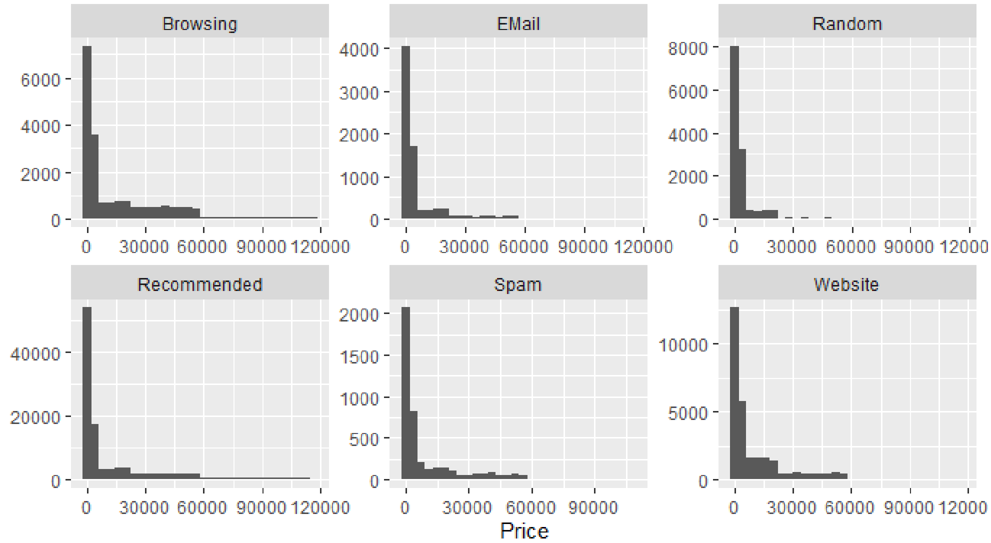

Figure 15: Histograms comparing Price, Why Bought and Count

3.4.2 Comparing Class, Age and the Count

The Figure 16 below shows the distributions of Age by count for each Class respectively. The class Sweets can be viewed as bimodal, while Clothing, Household, Technology and Luxury are skewed right. Food on the other hand is skewed left and Gifts is almost normally distributed. All of these graphs below contain random variation. This random variation means there is not a strict relationship between Classes and Age. The below figure shows that between the ages of 30 and 60 the most sales are made over the majority of Classes. This age group should therefore be the companies target group and their marketing and product innovation should be directed at this age group for a maximum profit. If the company in question is an online business, this age group making more purchases could be due to the fact that younger generations are more savvy with online purchases than older generations.

Figure 16: Histograms comparing Class, Age and Count

3.4.3 Comparing Price, Class and the Count

Below in Figure Figure 17 it can be seen that the distribution for all the graphs are uniform. However when observing the x-axis, it can be seen that some classes generate a much larger income. For example when comparing the graph for the class Technology and the class Clothing. The price range for Technology goes up to 60 000 and the one for clothing goes up to 1000 only. However both at their maximum price rage are achieving 1000 sales. This shows why the Technology class is able to generate a much higher income than the Clothing class. When this occurs in a company, the company should focus in increasing the sales of their products in classes like Technology, as they generate a higher revenue. Therefore the Class is a key feature that must be further analyzed.

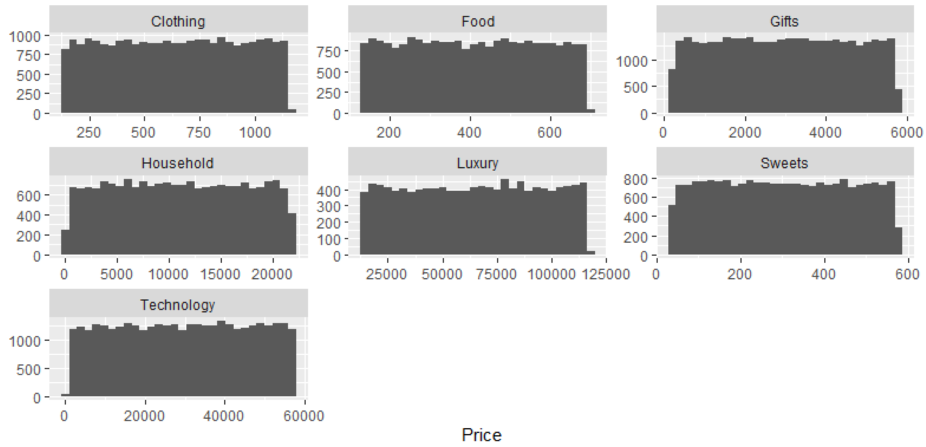

Figure 17: Histograms comparing Price, Class and Count

3.4.4 *Comparing Delivery Time, Class and the Count*

The graphs below in Figure 18 cements the idea that the Class of the product has a big influence on the other features in the dataset. The Delivery Times are mostly, random variation, normally distributed. It can however be observed that each class generally requires a different Delivery Time range. For example the class Household has higher delivery times than the class Technology, as seen on the x-axis. The Clothing, Food, Luxury and Sweets classes are seen not to have continuous Delivery Times. This is noticed due to the bar graph like structure they exhibit. From this it can be deduced that these classes potentially have fixed Delivery Time options that the customer may select from and these Delivery Time are single numbers not a range. Ie. The delivery will take place in 3 working days not between 3-5 working days.

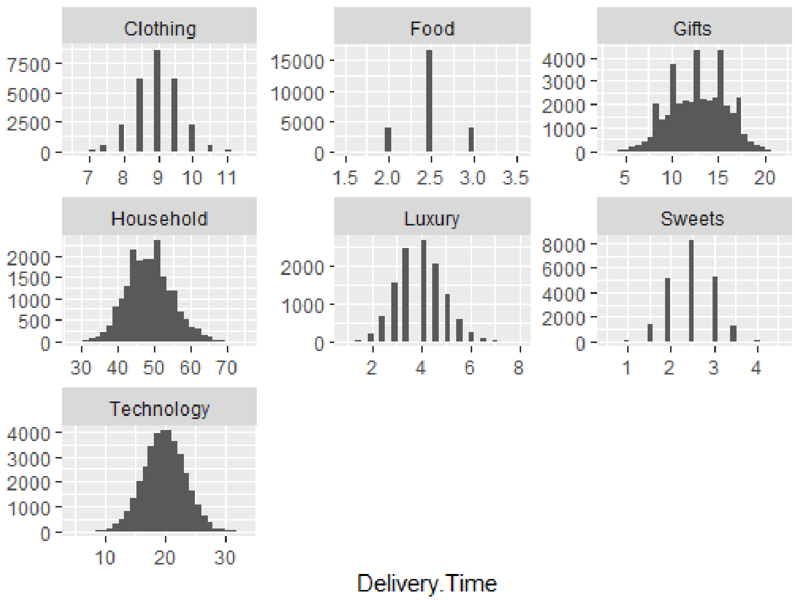

Figure 18: Histogram comparing Delivery Time, Class and Count

3.5 Process Capability Indices

Process capability indices include a Lower Specification Limit (LCL) and an Upper Specification Limit (USL), to determine if a process is performing appropriately it must fall between these two limits. Process Capability Indices are utilized in businesses to shift the process to fall into a target range and in this reduce variation in the process. This will ensure the business conforms to customer specifications (Process Capability Indices, 2022). The process capability indices were calculated for the process delivery times for the class Technology. The USL given was 24 and the LSL given was 0. Having a LSL of 0 is very logical, because we are measuring delivery time. Delivery time cannot be less than zero as it would mean the delivery happened prior to the order being placed, which is logically impossible as time cannot be negative. If a Delivery Time of 0 is achieved this is also still logical, as a rush order could be placed for same day delivery. It is more desirable to have a Delivery Time close to the LSL than one closer to the USL. The shorter the delivery, the happier the customer will be, therefore a good service level was provided and the products can be viewed as a good product. To determine the Process capability indices the standard deviation and mean of the Technology class were determined and utilized in the formulas in Table 2 . The mean was found to be 20.01095 and the standard deviation (sd) was 3.5019927. The resulting capability indices can be found below in Table 3.

Table 2: Process capability indices formulas

Indices	**CP**	**CPU**	**CPL**	**CPK**
FORMULA	$\frac{USL - LSL}{6 \times sd}$	$\frac{USL - mean}{3 \times sd}$	$\frac{mean - LSL}{3 \times sd}$	min(CPL,CPU)

Table 3: Process Capability Indices

Indices	**CP**	**CPU**	**CPL**	**CPK**
	1.142	0.380	1.905	0.380

The results in Table 3 show that CPK is less than CP, this proves that the distribution of the data is not perfectly centered. When there is a deviation on either side of the midpoint, it results in a CPK < CP (Process Capability Indices, 2022). It is expected that the distribution has deviated to the left of the midpoint, as CP>0 and CP>CPK, where the midpoint is the center between the LSL and the USL. The CPK value is generally aimed at 1 or higher, therefore a CP of 1.142 makes the process capable. Having a higher CP will cause the distribution of the data to be less wide which will render the process more accurate and precise, which is optimal and should be aimed for.

4 STATISTICAL PROCESS CONTROL

4.1 First 30 samples X-charts and S-charts

The Valid dataset that has already been ordered in ascending order according to Year, Month, Day and Primary.Key (previously called X) was used to determine the statistical process control charts. Firstly this data was subset into the respective Classes, to be analysed and interpreted separately. The first 450 instances from each Class subset was divided into 30 samples, containing 15 instances each. These samples were used to create control charts using the function qic in R, which labelled the UCL, LCL and CL. The control charts are used to analyse the variation of the process and access its predictability. Points that lie outside the outer control limits are viewed as signs of out-of-control (Control Chart - Statistical Process Control Charts | ASQ, 2022). The first 30 samples are used to create limits for the process that should be adhered to in the future. The X and S control charts per class can be seen in the respective figures below.

4.1.1 Technology

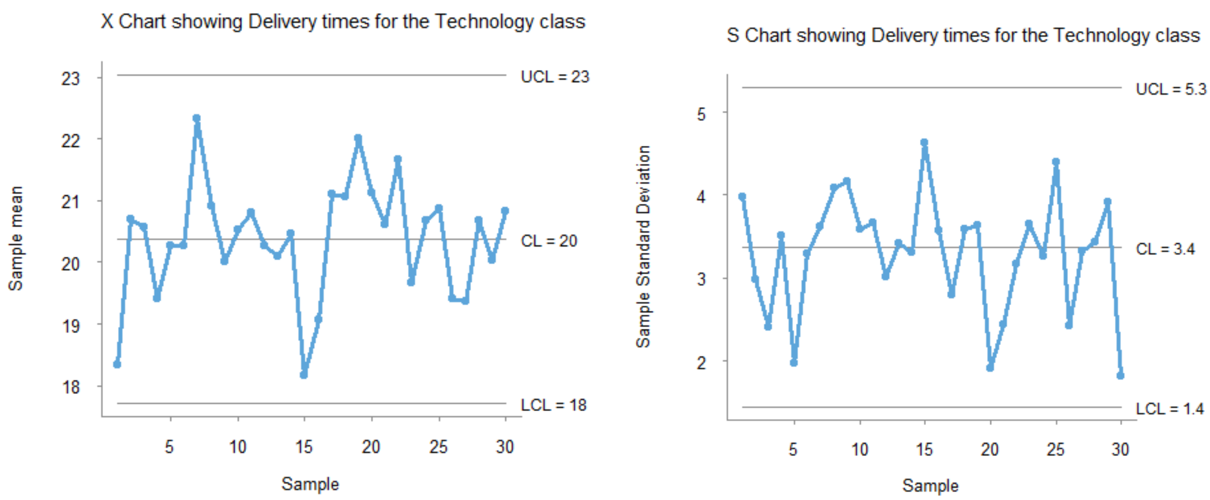

Figure 19: X-Chart and S-Chart for the first 30 samples for the class Technology

4.1.2 Clothing

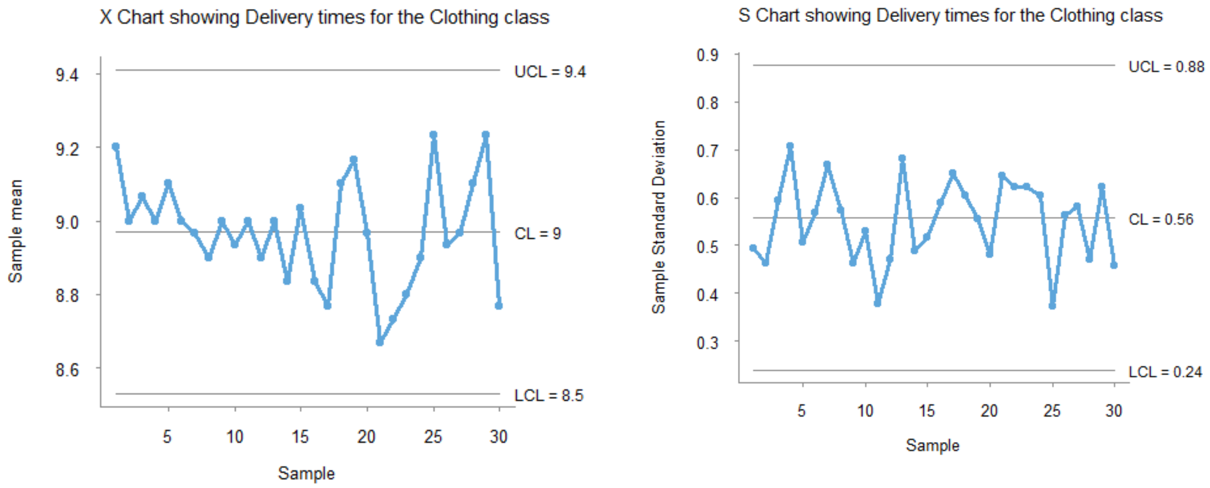

Figure 20: X-Chart and S-Chart for the first 30 samples for the class Clothing

4.1.3 Household

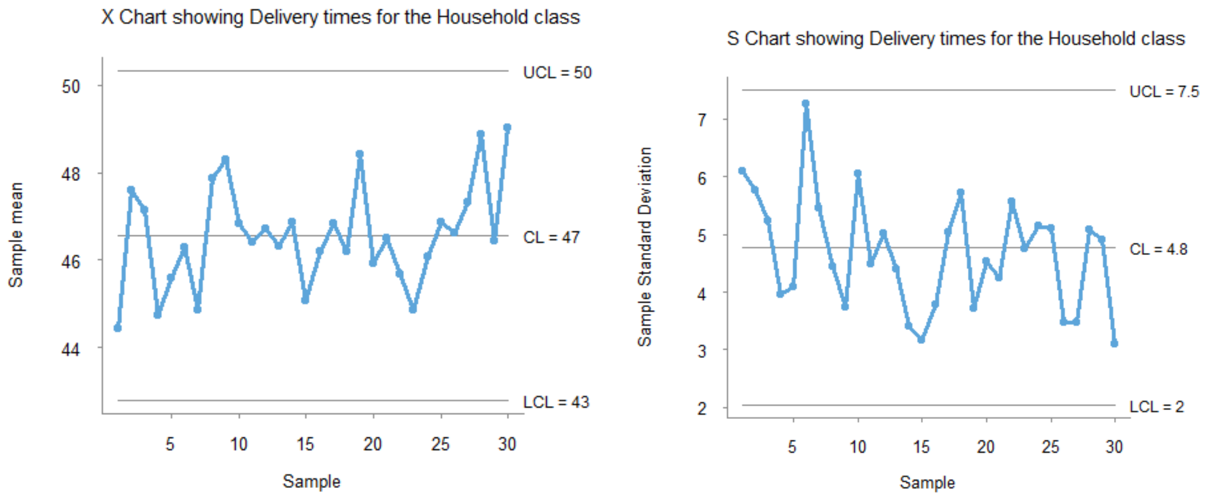

Figure 21: X-Chart and S-Chart for the first 30 samples for the class Household

4.1.4 *Luxury*

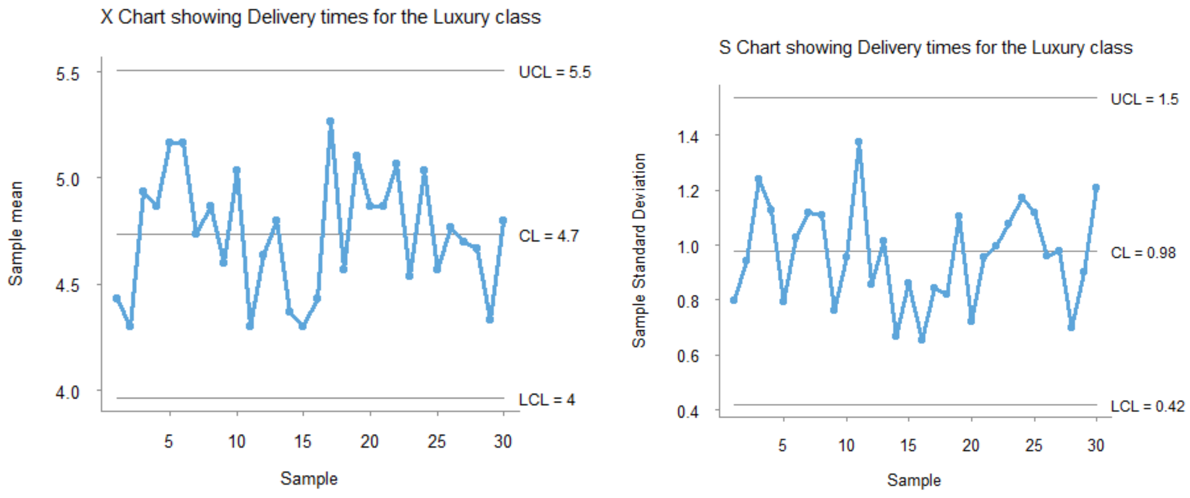

Figure 22: X-Chart and S-Chart for the first 30 samples for the class Luxury

4.1.5 *Food*

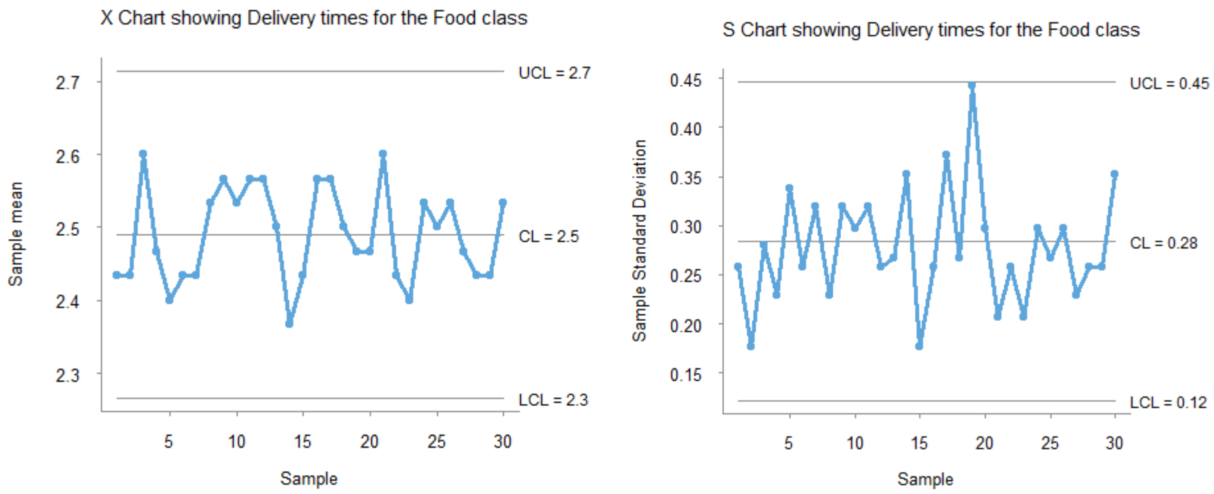

Figure 23: X-Chart and S-Chart for the first 30 samples for the class Food

4.1.6 Gifts

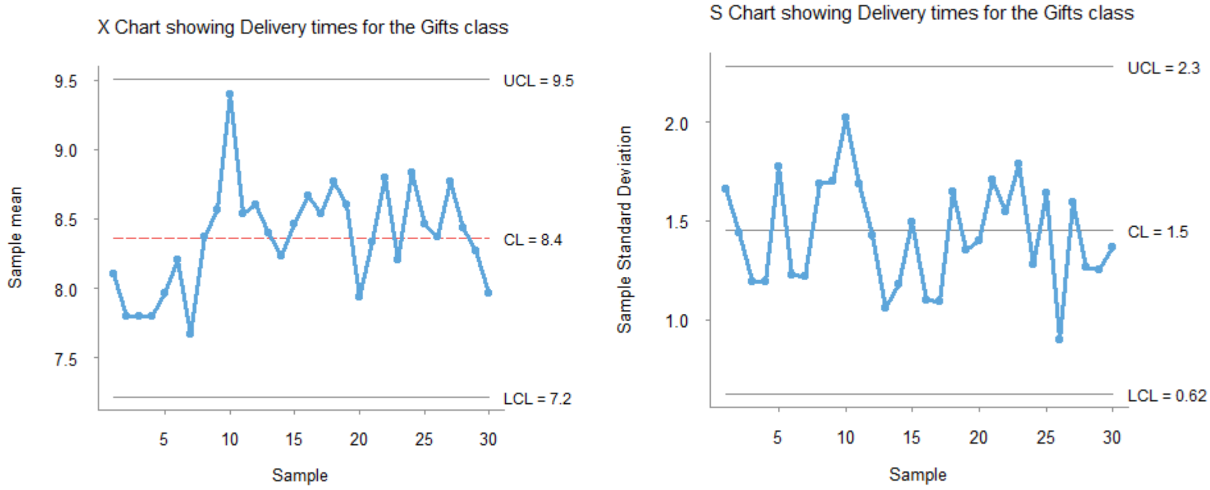

Figure 24: X-Chart and S-Chart for the first 30 samples for the class Gifts

4.1.7 Sweets

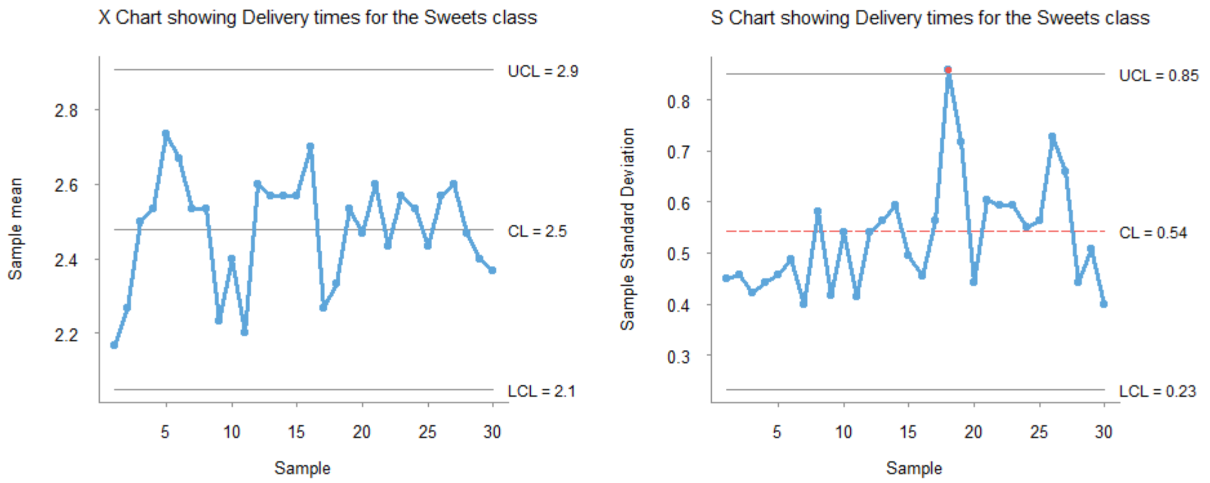

Figure 25: X-Chart and S-Chart for the first 30 samples for the class Sweets

When analyzing the above Figure 19 to Figure 25 it is clear there is only one sample outside the outer bound limits. This sample is located in the S-Chart of the class Sweets and is marked in red in Figure 25. This sample must therefore be removed, as it is a sign of variation and in the determining of the process control limits, we do not want any variation.

4.1.8 Tables

The first 30 samples were then used to determine the centre-lines, outer control limits, 2-sigma-control limits and 1-sigma-control limits. These values were obtained for the X and S-charts using the qic function in R along with some manual calculations. A table in R was created to stores these output values and these values are outlined in Table 4 and Table 5 respectively.

Table 4: X-Charts control limits per class

	Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
1	Technology	23.03203	20.8173716666667	21.2603033333333	20.37444	19.48858	19.93151	17.71686
2	Clothing	9.410027	9.04333783333333	9.11667566666667	8.97	8.82332433333333	8.89666216666667	8.529973
3	Household	50.32763	47.1897883333333	47.8173566666667	46.56222	45.3070833333333	45.9346516666667	42.79681
4	Luxury	5.506919	4.8641165	4.992677	4.735556	4.47843466666667	4.60699533333333	3.964192
5	Food	2.714047	2.52734116666667	2.56468233333333	2.49	2.41531766666667	2.45265883333333	2.265953
6	Gifts	9.50748	8.5521725	8.743234	8.361111	7.978988	8.1700495	7.214742
7	Sweets	2.905507	2.54906616666667	2.62035433333333	2.477778	2.33520166666667	2.40648983333333	2.050049

Table 5: S-Charts control limits per class

	Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
1	Technology	5.294956	3.6894035	4.010514	3.368293	2.72607166666667	3.04718233333333	1.441629
2	Clothing	0.8767079	0.610869816666667	0.664037433333333	0.5577022	0.451366966666667	0.504534583333333	0.2386965
3	Household	7.502185	5.22735	5.682317	4.772383	3.86244866666667	4.31741583333333	2.04258
4	Luxury	1.536861	1.07084933333333	1.16405166666667	0.977647	0.7912423	0.88444465	0.4184329
5	Food	0.4463897	0.311034033333333	0.338105166666667	0.2839629	0.229820633333333	0.256891766666667	0.1215361
6	Gifts	2.28402	1.5914525	1.729966	1.452939	1.17591196666667	1.31442548333333	0.6218579
7	Sweets	0.8522051	0.59379685	0.6454785	0.5421152	0.4387519	0.49043355	0.2320253

4.2 Samples from instance 31:end of datas' X-charts and S-charts

The instances in the dataset continuing after the 450th entry were then sampled into samples containing 15 instances each, still for each respective class. The procedure used to plot these instances on X-charts and S-charts is identical to the procedure used in Section 4.1 of this report. The limits discovered from the first 30 samples were then used to measure if the process for the following samples were operating within those specified limits. If the process sample delivery times were within the limits found it is said to be in-control, however if the process sample delivery times were outside of the limits it is said to be out-of-control.

4.2.1 Technology

The X-Chart below in Figure 26 and the S-Chart in Figure 27 have very similar distributions. These distributions are evenly spread around the CL (average), meaning the samples in reality follow a normal distribution, which is symmetrical around the average. Both have a handful of outliers marked by red dots. This few number of outliers shows that the process for the class Technology is nearly perfectly under control. The X-Chart contains more outliers that breech the LCL than the UCL, controversially the S-Chart contains more outliers breech the UCL than the LCL.

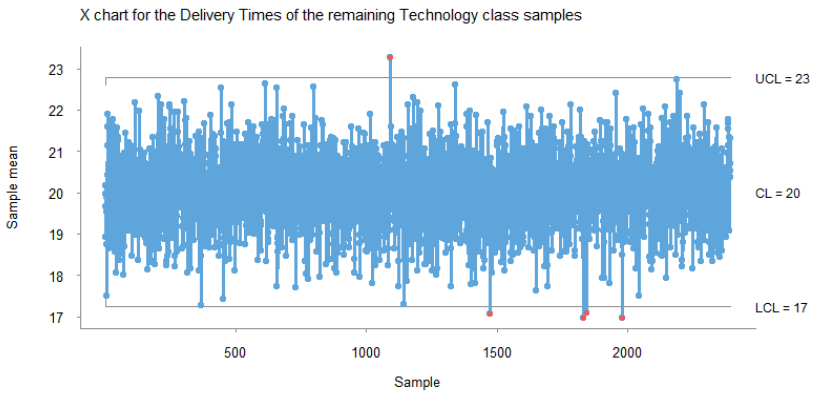

Figure 26: X-chart for the remaining samples in the class Technology

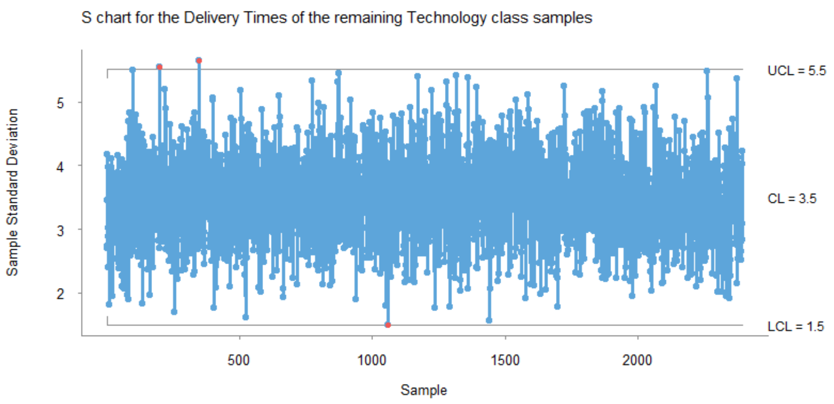

Figure 27: S-chart for the remaining samples in the class Technology

4.2.2 Clothing

In Figure 28 and Figure 29 below the X-Chart and S-Chart respectively for the class Clothing can be observed. The X-Chart follows the same trend as the X-Chart for the class Technology, with very few outliers and the same distribution form discussed. On the other hand the S-Chart contains far more outliers, populated towards the end of the samples. The outliers towards the end of the samples exceed the UCL. This signifies that the process become out of control towards the last samples.

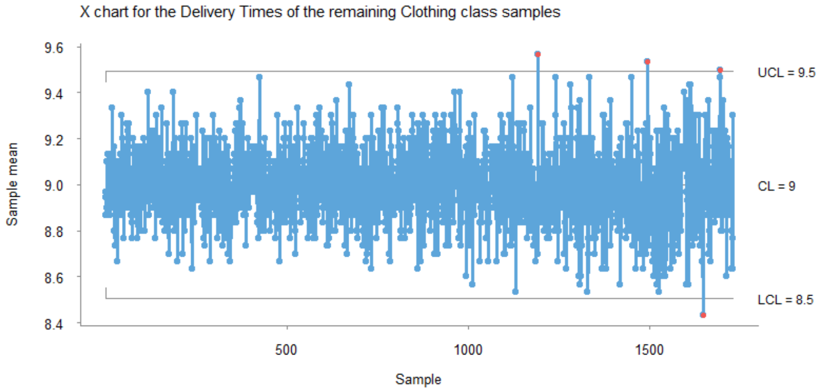

Figure 28: X-chart for the remaining samples in the class Clothing

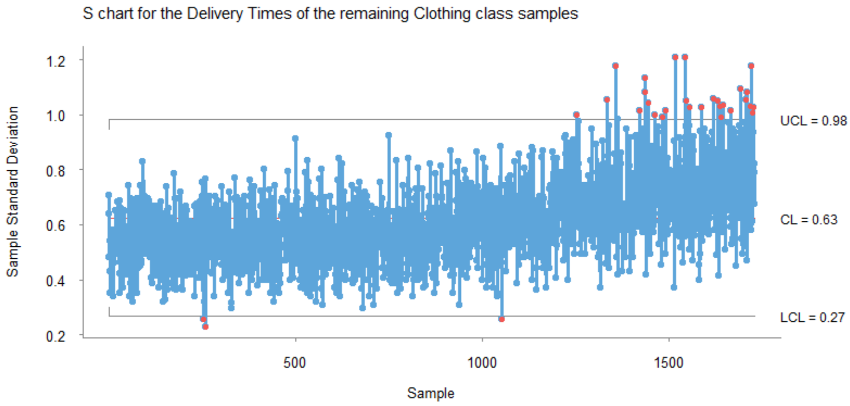

Figure 29: S-chart for the remaining samples in the class Clothing

4.2.3 Household

When reviewing the X-Chart in Figure 30 and the S-Chart in Figure 31 it is found that the class Household had problems remaining within the UCL and the LCL. The X-Chart shows the sample means started by falling below the LCL, then was under control and then deviated above the UCL. This could mean the process was under performing, corrective measures were put into place, but an over correction was done and therefore there process started to exceed the upper limit. Another possible explanation could be that general Households were not willing to pay for fast deliver, therefore their delivery time increased, because they chose a cheaper delivery option that takes longer. The S-Chart shows a handful of outliers majority towards the end of the samples.

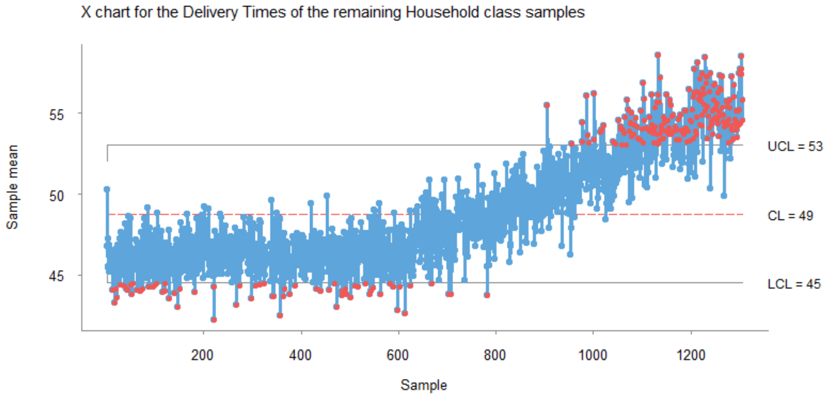

Figure 30: X-chart for the remaining samples in the class Household

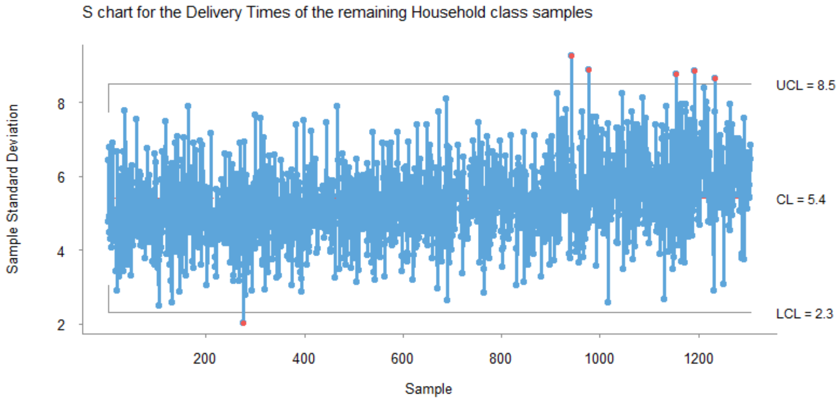

Figure 31: S-chart for the remaining samples in the class Household

4.2.4 Luxury

The X-Chart in Figure 32 indicates that the mean of the samples for the process was only under control for a short period of time. The samples started above the UCL and later towards the end of the samples were below the LCL, this indicates that the delivery times got shorter. This could also mean the process was under performing, corrective measures were put into place, but an over correction was done and therefore there process started to exceed the upper limit. However the S-Chart in Figure 33 has very few outliers beyond the limits, there are only a few above the UCL at the beginning of the samples.

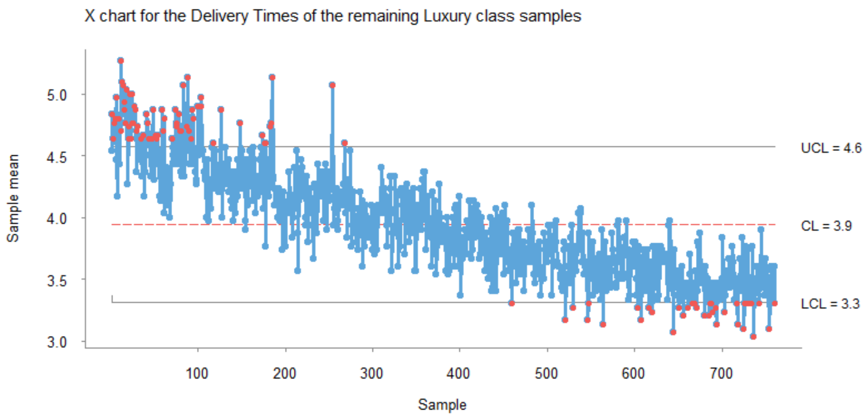

Figure 32: X-chart for the remaining samples in the class Luxury

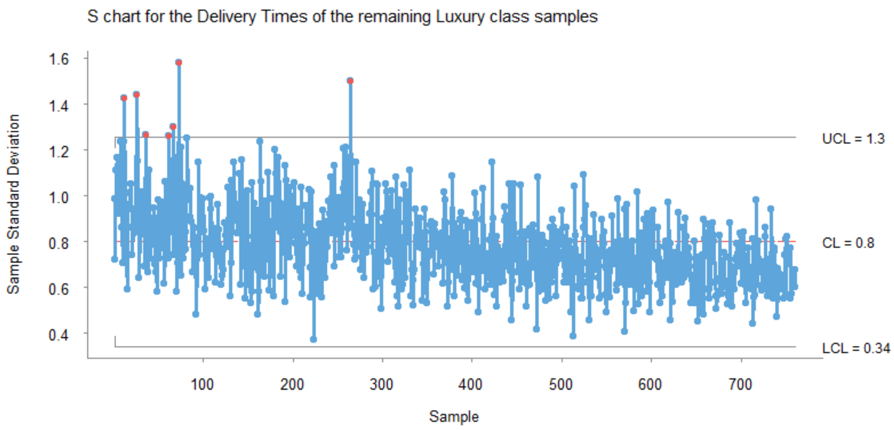

Figure 33: S-chart for the remaining samples in the class Luxury

4.2.5 Food

The X-Chart in Figure 34 and the S-Chart in Figure 35 closely follow the distribution found for the class Technology. Both the classes Technology and Food have a difference of 5 between their UCL and LCL's for their X-Charts and have very few scattered outliers in both their charts. The class Food, along with Technology, therefore has less variation in their process when compared to other classes.

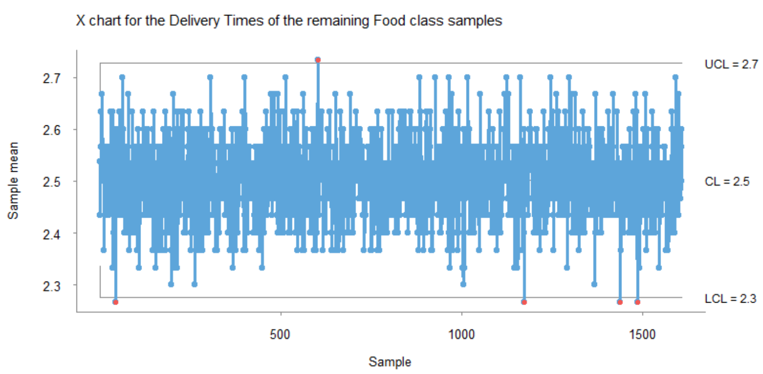

Figure 34: X-chart for the remaining samples in the class Food

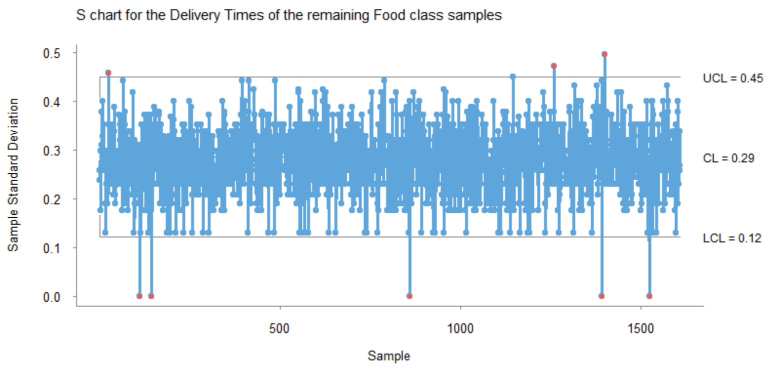

Figure 35: S-chart for the remaining samples in the class Food

4.2.6 Gifts

The X-Chart in Figure 36 shows the first sample to look like an outlier, because it is actually within the control limits and doesn't follow the linear trend witnessed by the other samples. The samples follow a linear pattern with a positive gradient, starting below the LCL, then a very small portion of samples is within control, followed by a large portion of samples beyond the UCL. This means the process was not adhering to the set limits and as time went by the delivery times increased significantly. The S-Chart in Figure 37 shows very few outliers, one to be noted is the first sample, which according to the X-Chart was within the bounds almost uncharacteristically according to the witnessed trend. The rest of the S-Chart shows very little variation form the trend witnessed in the X-Chart.

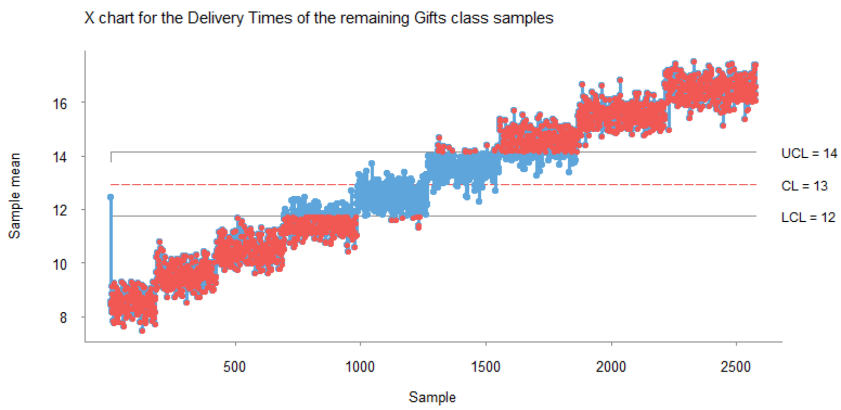

Figure 36: X-chart for the remaining samples in the class Gifts

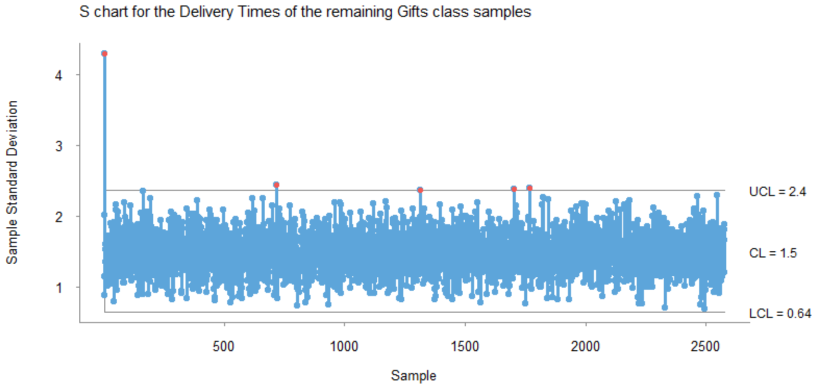

Figure 37: S-chart for the remaining samples in the class Gifts

4.2.7 *Sweets*

The X-Chart in Figure 38 and the S-Chart in Figure 39 below follow similar trend to the classes Technology and Food. These three classes have similar distribution shapes and contain few dispersed outliers. Specifically the S-Chart in Figure 39 is the only graph to not contain any outliers beyond the specified bounds (UCL and LCL). This means these classes are the most in-control.

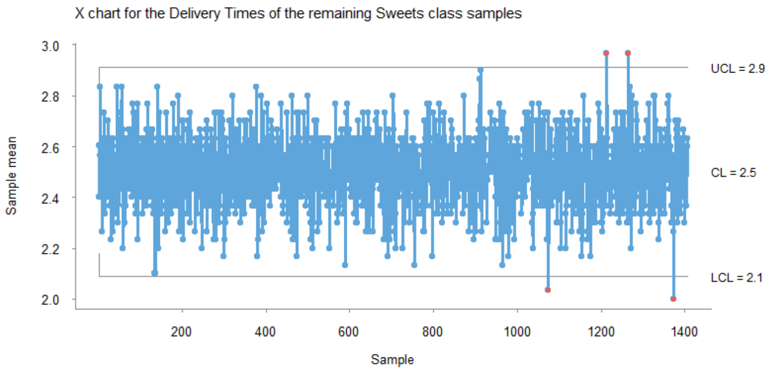

Figure 38: X-chart for the remaining samples in the class Sweets

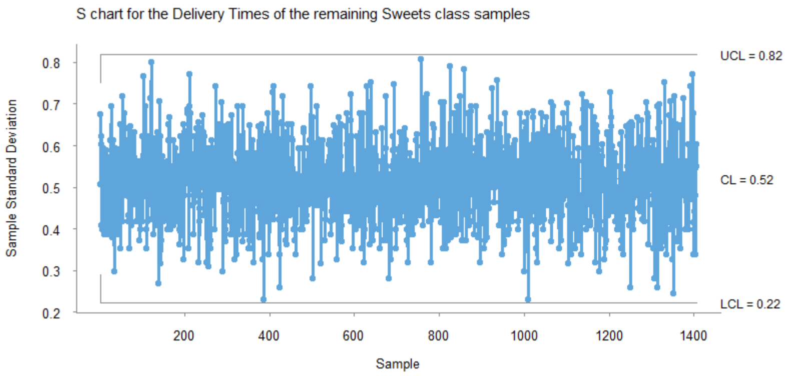

Figure 39: S-chart for the remaining samples in the class Sweets

5 OPTIMIZING THE DELIVERY PROCESS

5.1 Rule A: One sample means outside the control limits

Using the ordered subsets of the original Valid dataset per classes, the entire sub datasets were analyzed. The UCL and LCL obtained from the first 30 samples using the function qic was used to check which samples in the entire sub dataset were out of bounds. Each subset according to their class was first sampled, each sample containing 15 samples and the remaining instances were ignored. The mean of each sample was obtained and compared to the UCL and LCL. If this mean was outside of these bounds, it was flagged as out-of-control. When samples are flagged as out-of-control, it means they are not being delivered in the required delivery time. An issue like this can cost a business loyal customers and therefore profit. The number of samples from each class that were out-of-control, as well as the sample number of the first 3 and the last 3 that were out-of-control can be view below in Table 6.

Table 6: Rule A samples

Class	**First sample**	**Second sample**	**Third sample**	**Third last sample**	**Second last sample**	**Last sample**	**Total number identified**
Technology	37	398	483	1872	2009	2071	12
Clothing	455	702	1152	1677	1723	1724	14
Household	252	387	643	1335	1336	1337	393
Luxury	184	207	227	789	790	791	414
Food	633	-	-	-	-	633	1
Gifts	213	216	218	2607	2608	2609	2282
Sweets	1104	1243	1294	-	-	1403	4

It is clear from Table 6 that the class Gifts by far contained the largest number of samples that were out-of-control. The classes Technology, Clothing, Food and Sweets were the classes that were the most in control as seen in their respective figures below and the table above. The figures below cement the results discovered in the table above.

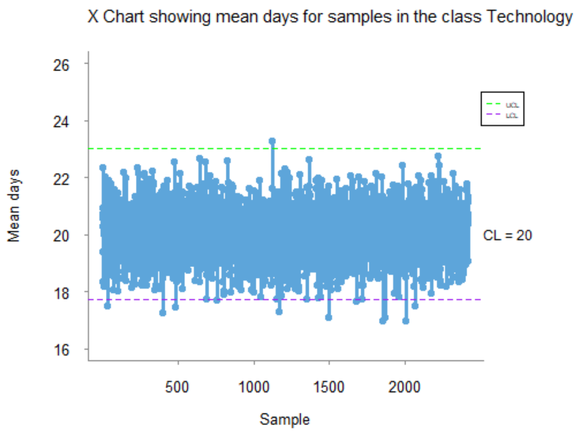

Figure 40: X-Chart for Technology

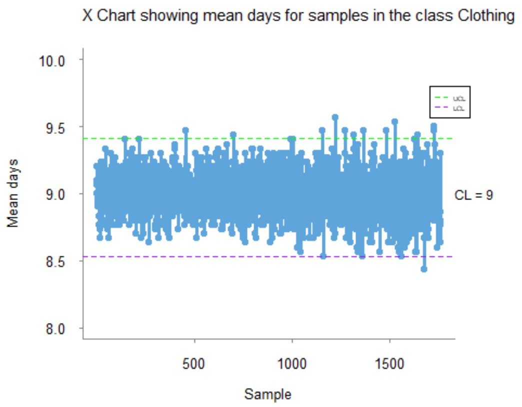

Figure 41: X-Chart for Clothing

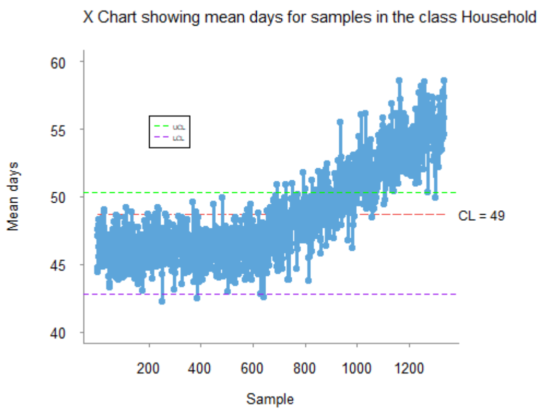

Figure 42: X-Chart for Household

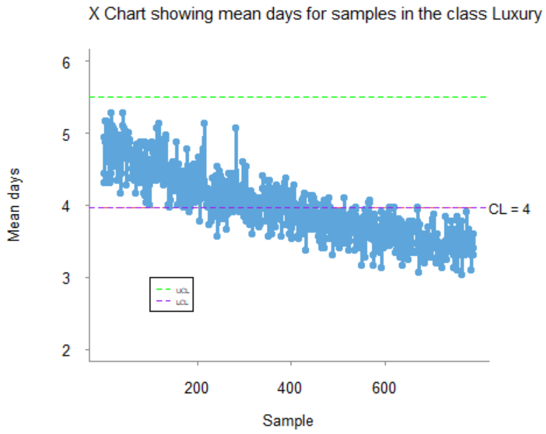

Figure 43: X-Chart for Luxury

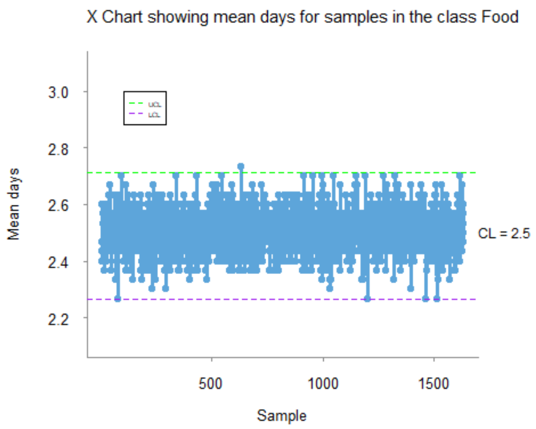

Figure 44: X-Chart for Food

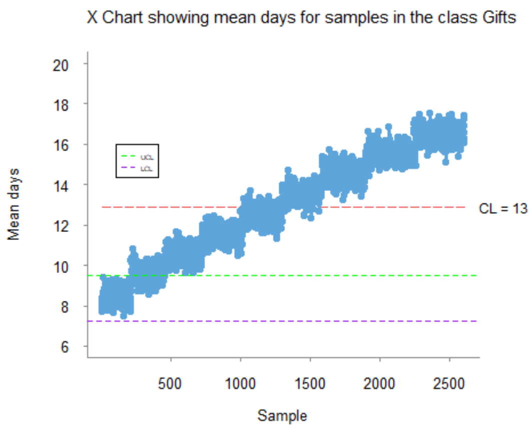

Figure 45: X-Chart for Gifts

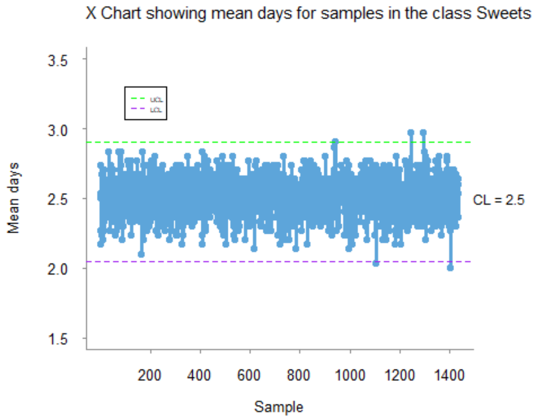

Figure 46: X-Chart for Sweets

5.2 Rule B: Most consecutive samples of standard deviations between -0.3 and +0.4 sigma-control limits and the ending sample number

From Table 7 below it can be seen that Food and Gifts have the most consecutive samples between the specified bounds. This means these two classes are the most under control. All the respective number of consecutive samples for each Class are within a close range, this indicates that the processes are almost within the same control limits.

Table 7: Showing the most consecutive samples within specified bounds

Class	**Longest number of consecutive samples**	**Ending sample number**
Technology	5	223
Clothing	4	45
Household	5	561
Luxury	4	776
Food	6	326
Gifts	6	477
Sweets	5	63

5.3 Type I error

In R the function pnorm() was utilized to find the Type I errors. Type I errors, also known as Manufacturers Errors, occur when the Null Hypothesis (H0) is rejected when in reality it is actually true. The hypothesis for the testing being done are as follows:

H0: The process is in control and centered on the centerline calculated using the first 30 samples.

HA: The process is not in control and has moved from the centerline or has increased or decreased variation.

5.3.1 Rule A

For Rule A it is required to find the samples that fall between +3 sigma and -3 sigma from the central limit (CL). The probability of a Type I error occurring for this rule can be viewed in Table 8 below. The probability viewed corresponds to 0.27% when rounded off.

5.3.2 Rule B

Using Table 7 above it was found that the longest consecutive sample to fall between -0.3 and +0.4 is of the length 6. This number is used in the calculation of Type I error. The probability for a Type I error for Rule B can be seen in Table 8. This probability corresponds to a 34.77% chance of making a Type I error when rounded off.

Table 8: Probability of Type I error

	Rule A	**Rule B**
Probability of Type I error	0.002699796	0.3476899

5.4 Best profit

The individual delivery times for the class Technology are the focus in this part. In order to maximize profit the delivery process hours must be centered on a specific number of hours. Sales slower than 26 hours result in lost sales and therefore the center of this process must be before 26 hours. In addition to lost sales, a delivery past this number of hours incurs a cost of R329/item-late-hour and it costs R2.5/item/hour to reduce the average hour by 1 hour. Using brut-force the center point for the delivery process needs to be decreased by 14 Days to 36 Days. This center line will incur the least amount of additional cost and maximize profit. This interpretation corresponds to the Taguchi Loss Function, that shows that the more variation from the center of a process, the higher the cost becomes.

5.5 Type II error

Type II error, Consumers errors, occur when a false negative is found or an error of omission. This means the Null Hypothesis is not rejected, when it is actually false and was meant to be rejected. In the case being interpreted, it means the process is thought to be in control, when in reality it is not. The process is actually not centered and corrective steps need to be taken. However due to the false negative the corrective steps will not be taken and therefore costing the company money, time and resources.

6 DOE AND MANOVA

MANOVA is a function in R that stands for multivariate analysis of Variance or Multivariate ANOVA. A MONOVA test must include at least two dependent variables which are utilized to detect differences between the different groups of the independent variable. In the case provided the feature Class and Why Bought will be the independent variables, as we want to see the significant differences between their groups for the selected dependent variables Price and Delivery Time. These dependent features were selected based on the results in the previous sections. In the previous sections it was witnessed that these features varied for each respective class and are therefore candidates to be analyzed.

The null and alternative hypothesis can be viewed as the following respectively:

H0: The respective group means are the same for all groups

HA: At least one group has a mean that differs from the rest

If the null hypothesis is found to be true, it is witnessed that the independent variable has no relationship with the dependent variables. However if the alternative hypothesis is found to be true, there is a relationship between the features (Bevans, 2022). A significance level of 5% or 0.05 is used, this is the chance we are willing to take that we are making an incorrect associated between features.

6.1 Class as the independent variable

Firstly Class was selected as the independent feature and was compared against the dependent variables Price and Delivery time. From Figure 47 we can view the p-value found of 2.2e-16. This is an extremely small p-value. A small p-value that is smaller than the significance level results in a rejection of the H0 (Null hypothesis), because the difference between the means are statistically significant. From this we can conclude that the Price and Delivery Time depends on the Class for at least one group. The boxplots in Figure 48 and Figure 49 confirms the relationship between Class as the independent feature and Price or Delivery Time as the dependent feature. It can clearly be seen the mean of the groups in Class are different, confirming the rejection of the H0. From this it is understood that to deliver a good service and be reliable to customers, each individual group in Class must be tailored according to pricing and delivery times.

```
> summary.aov(MAN1)
 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

 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
```

Figure 47: Summary of the MANOVA results for Class

Figure 48: Boxplot of Class vs Price

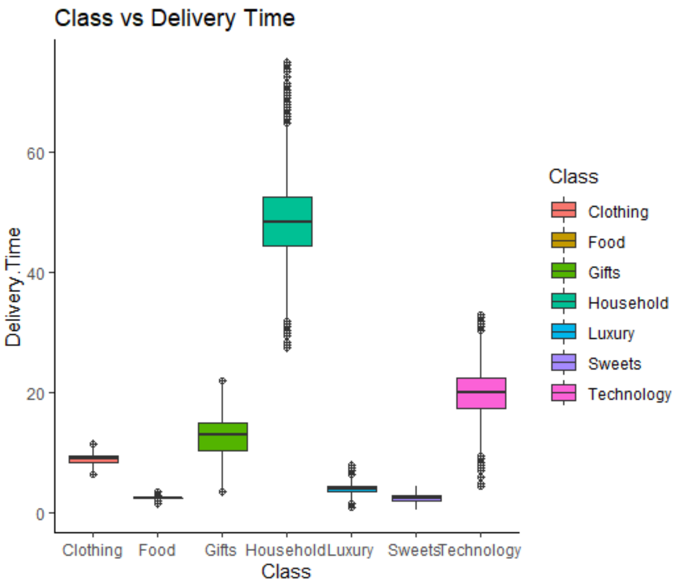

Figure 49: Boxplot of Class vs Delivery Time

From Figure 49 it can be observed that the Classes Luxury, Sweets, Food and Clothing will be the most reliable when referring to delivery time. The distribution of their boxplots are very small and indicate very little variation. The small amount of variation signifies an almost constant delivery time for those classes and this will lead to greater customer satisfaction.

6.2 Why Bought as the independent variable

Now Why Bought was used as the independent variable and compared with Price and Delivery Time respectively as the dependent variables. Below in Figure 50 the p- value of 2.2e-16 can be viewed. This again results in the H0 being rejected, because the p-value is less than the significance level of 0.05 used. These results lead to the conclusion that the Price and Delivery Time does depend on the 'Why Bought' grouping for at least one group. From Figure 51 and Figure 52 we can see that there is clearly a difference of at least one mean in the boxplots per 'Why Bought'. The service delivered regardless of the reason why the customer made the purchase can be seen to be the same in Figure 52. This means there is no bias and the reliability is the same regardless for the reason of the purchase.

```
> summary.aov(MAN2)
 Response Price :
                Df     Sum Sq    Mean Sq F value    Pr(>F)    
Why.Bought       5 1.5742e+12 3.1484e+11  736.26 < 2.2e-16 ***
Residuals   179972 7.6960e+13 4.2762e+08                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Delivery.Time :
                Df   Sum Sq Mean Sq F value    Pr(>F)    
Why.Bought       5   783320  156664  822.74 < 2.2e-16 ***
Residuals   179972 34269697     190                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 50: Summary of the MANOVA results for Why Bought

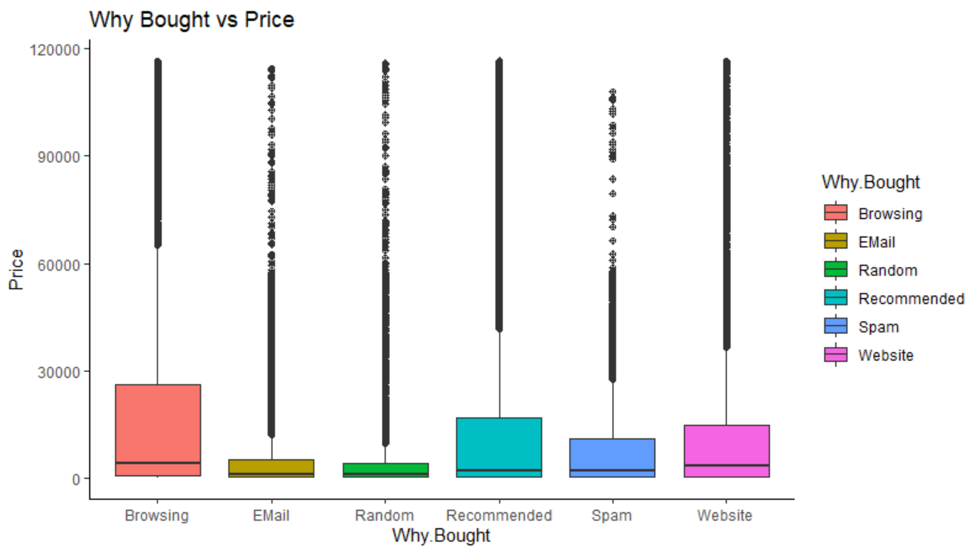

Figure 51: Boxplot of Why Bought vs Price

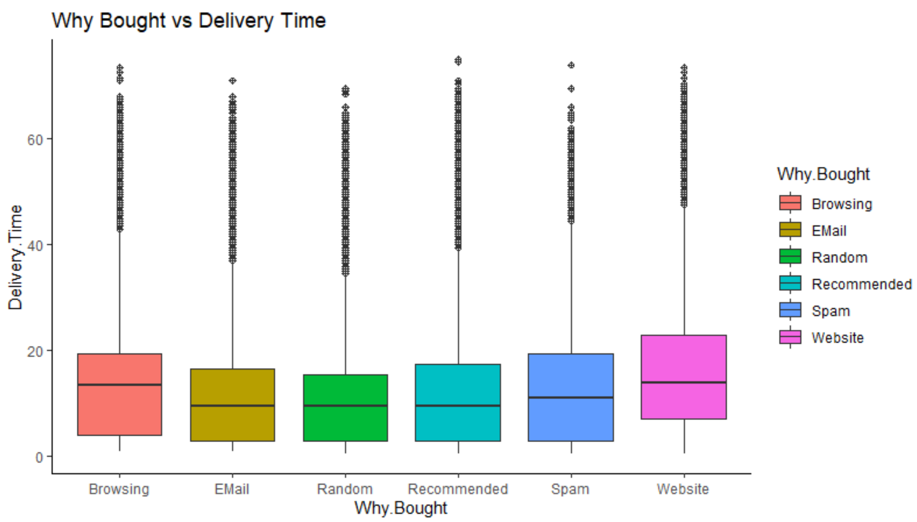

Figure 52: Boxplot of Why Bought vs Delivery Time

7 RELIABILITY OF THE SERVICE AND PRODUCTS

7.1 Problem 6: Cool Food Inc. subsidiary Lafrideradora

The thickness of a refrigerator part should fall between 0.06 ± 0.04 (cm). It costs $45 to scrap the part if it does not fall within these limits.

Where :

$T = 0.06$

$L =$ The Taguchi Loss Function (ie. the associated loss in Dollars ($))

$45 = k(0.04)^2$

Therefore , $k = 45/((0.04)^2)$

Therefore $k = 28125$

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

$L(x) = 28125*(x-0.06)^2$

Using a vector of x ranging from 0.01 to 0.11, which is just beyond the upper bound of 0.1 and the lower bound of 0.02 specified for the process. This will aid the visualization of the Taguchi Loss Function seen in Figure 53. It can be noted that the red dashed line of the left illustrates the lower bound and the red dashed on the right the upper bound, while the blue dashed line in the middle signifies the desired length (cm). When analyzing Figure 53 it is evident that at the desired length, there is no associate cost, however the larger the deviation becomes from the desired length, the greater the loss in the form of cost becomes. This is a clear indicator that businesses need to minimize the variation from the desired specifications in order to save costs. Not only will this save costs, but it will also improve the reliability of the product and in turn improve the customer service. Customers do not want a part that is smaller or bigger than what they require and that the company specifies. If this is the came the business will lose customers and return.

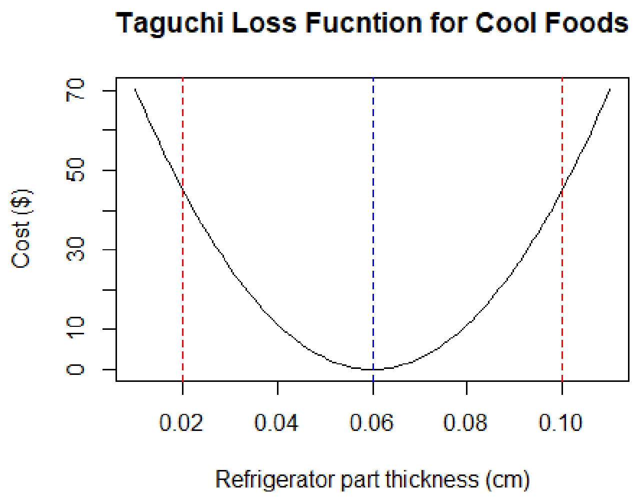

Figure 53: Taguchi Loss Function for problem 6

7.2 Problem 7: Cool Food Inc. subsidiary Lafrideradora

The previous conditions from Problem 6 still apply, however the scrap cost is reduced to $35 per part.

7.2.1 Problem 7a

The Taguchi Loss Function for this situation was determined.

$35 = \mathrm{k}(0.04)^\wedge 2$

$\text{Therefore, k} = 21875$

$\mathrm{L(x)} = \mathrm{k}^*(\mathrm{x}\text{-}\mathrm{T})^\wedge 2$

$\mathrm{L(x)} = 21875^*(\mathrm{x}\text{-}0.06)^\wedge 2$

The same appropriate x co-ordinate values were utilized to analyze the Taguchi Loss Function for this problem. The Taguchi Loss Function for this problem can be viewed below in Figure 54, it is clear that for a lower scrap cost, the parabolistic distribution is less steep. This means as the deviation increases further away from the desired length, the cost increases less dramatically, as expected. The red lines again indicate the desired part length range must fall between, while the blue line indicates the target.

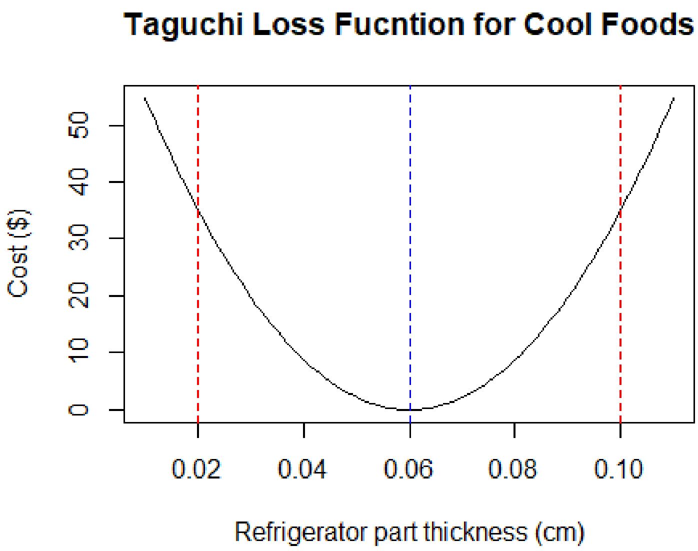

Figure 54: Taguchi Loss Function for problem 7a

7.2.2 Problem 7b

The previous conditions from problem 7a hold for this problem, except the deviation is now reduced to 0.027 cm. The Taguchi Loss Function was determined for these new conditions.

L(0.027) = 21875*(0.027)^2

L(0.027) = $15.946875

The above calculation proves that when the deviation from the target is 0.027cm, the associated cost will be $15.95 per part produced. The decrease in the variation will make the product more reliable and customers will be greatly satisfied.

7.3 Magnaplex

The reliabilities for machines A, B and C are 0.85, 0.92 and 0.90 respectively.

7.3.1 Magnaplex problem a

The system reliability, assuming only one machines is at each stage:

System reliability = reliability A * reliability B * reliability C

= 0.85*0.92*0.90

=0.7038

This results in a system reliability of 70.38%, which is above average but not ideal at all for a manufacturing company.

7.3.2 Magnaplex problem b

The reliability when using two machines at each stage:

System reliability using 2 machines = (1-(1-0.85)^2)* (1-(1-0.92)^2)* (1-(1-0.90)^2)

= 0.96153156

Which translates to having a 96.15% reliability when both machines are operational. This reliability is significantly higher than when only having 1 machine operational. There is an increase of 25.77% reliability, which is significant. For a manufacturing company that depends on their machines running to generate products to make an income, the higher reliability is more optimal. Therefore it is recommended that both machines are operational.

7.4 Delivery Process

<u>Information given:</u>

For the delivery process, there are 21 delivery vehicles available, of which 20 is required to be operating at any time to give reliable service. During the past 1560 days, the number of days that there was only 21 vehicles available was 190 days, only 20 vehicles available was 22 days, only 19 vehicles available was 3 days and 18 vehicles available only once. There are also 21 drivers, who each work an 8 hour shift per day. During the past 1560 days, the number of days that there were only 20 drivers available was 95 days, only 19 drivers available was 6 days and only 18 drivers available, once only.

Estimate on how many days per year we should expect reliable delivery times, given the information above.

This question was solved using the binomial probability distribution and cumulative probability distribution. Using the function dbinom() in R, the probabilities of certain events where calculated. The reliability of the drivers and vehicles were calculated independently to begin with and then later a overall probability was calculated, as a delivery requires both elements to be completed. Finally when stating the total number of days the delivery process can commence, the values calculated were rounded down to the nearest integer.

7.4.1 Starting with 21 vehicles

In order for the delivery process to be completed, at least 20 of the vehicles of the 21 available must be operating. This means there can only be zero or one vehicles that have failed and are not operating.

Vehicles:

Using R the probability that a vehicle is reliable = 0.00707166119021143

For 0 vehicles failed:

The probability is 0.861541147258885

Days 0 vehicles failed = probability 0 vehicles failed * 365

= 0.86154*365

= 314.4625187

Therefore there are 314 days that there are zero vehicle failures.

For 1 vehicle failed:

The probability is 0.128854282822688

Days 1 vehicles failed = probability 1 vehicles failed * 365

= 0.128854*365

= 47.0318132302813

Therefore there are 47 days that there is 1 vehicle failure.

Total:

Total vehicle reliability probability = the probability 0 vehicles failed* the probability 1 vehicle failed

= 0.861541147258885 * 0.128854282822688

=0.990395430081574

Days of total vehicle reliability = total vehicle reliability probability * 365

=0.990395430081574 * 365

= 361.494331979774

Therefore there are 361 days that the vehicles are reliable.

Drivers:

For the delivery process to be completed, it was previously mentioned that at least 20 of the 21 vehicles must be available. This indicates that there must at least be 20 available driver to operate the vehicles for the delivery to take place. Therefore there can only be zero or one driver unavailable.

Using R the probability that a driver is reliable = 0.00322440233541938

<u>For 0 drivers were unavailable:</u>

The probability is 0.934426926565182

Days 0 drivers unavailable = probability 0 drivers unavailable * 365

= 0.9344269*365

= 341.065828196291

Therefore there are 341 days that there are zero drivers unavailable.

<u>For 1 driver unavailable:</u>

The probability is 0.0634770110729551

Days 1 driver is unavailable = probability 1 driver unavailable * 365

= 0.063477*365

= 23.1691090416286

Therefore there are 23 days that there is one driver unavailable.

<u>Total:</u>

Total driver available probability = the probability 0 drivers unavailable* the probability 1 driver unavailable

=0.934426926565182 * 0.0634770110729551

=0.997903937638137

Days of total driver reliability = total driver available probability* 365

=0.9979 * 365

= 364.23493723792

Therefore there are 364 days that the drivers are reliable.

<u>Overall reliability of drivers and vehicles:</u>

Overall total reliability = Total driver available probability* Total vehicle reliability probability

= 0.997903937638137*0.990395430081574

=0.988319499497218

Days of overall total reliability = Overall total reliability*365

=360.736617316485

Therefore there are 360 days per year that the drivers and vehicles are reliable, resulting in total reliability and therefore allowing deliver to commence.

7.4.2 Using 22 vehicles

If we increased our number of vehicles by one to 22, how many days per year we should expect reliable delivery times?

The same aforementioned procedure was utilized in this section to find the overall days the delivery process can commence.

<u>Vehicles:</u>

Since there is an extra vehicle now and all other parameters remain the same, there can be 0, 1 or 2 vehicles failed and the delivery process can still commence.

Using R the probability that a vehicle is reliable = 0.00707166119021143

<u>For 0 vehicles failed:</u>

The probability is 0.855448620164044

Days 0 vehicles failed = probability 0 vehicles failed * 365

= 0.85544*365

= 312.238746359876

Therefore there are 312 days that there are zero vehicle failures.

<u>For 1 vehicle failed:</u>

The probability is 0.1340355960865

Days 1 vehicles failed = probability 1 vehicles failed * 365

= 0.13403*365

= 48.9229925715723

Therefore there are 48 days that there is 1 vehicle failure.

<u>For 2 vehicles failed:</u>

The probability is 0.0100233521413271

Days 2 vehicles failed = probability 2 vehicles failed * 365

= 0.0100233*365

= 3.65852353158438

Therefore there are 3 days that there is 2 vehicle failures.

<u>Total:</u>

Total vehicle reliability probability = the probability 0 vehicles failed* the probability 1 vehicle failed* probability 2 vehicles failed

= 0.855448620164044* 0.1340355960865*0.0100233521413271

=0.999507568391871

Days of total vehicle reliability = total vehicle reliability probability * 365

=0.999507568391871* 365

= 364.820262463033

Therefore there are 364 days that the vehicles are reliable.

<u>Overall reliability of drivers and vehicles:</u>

Assuming to use the availability of the drivers found for the 21 vehicles.

Overall total reliability = Total driver available probability* Total vehicle reliability probability

= 0.997903937638137*0.999507568391871

=0.987832819736735

Days of overall total reliability = Overall total reliability*365

=364.055576442039

Therefore there are 364 days that the drivers and vehicles are reliable, resulting in total reliability and therefore allowing deliver to commence. This means when 1 extra vehicle got added, an extra 4 days got added to the delivery process reliability per year. From this it can be suggested that the company adds the additional driver as it will increase reliability by a substantial amount and therefor increase customer satisfaction. When a business retains satisfied customers, they promote the business to other individuals and this can aid the growth of the business without incurring any additional costs.

8 CONCLUSION

In conclusion when manipulating the valid dataset in R, there was found to be a clear relationship between features in the dataset. The relationships between these features were carefully analysed and lead to solutions to increase the companys' customer satisfaction, use of materials, reliability, service and return. The class in which a product falls greatly influences the Price if is sold at and the Delivery Time required by the customers. Therefore each class should be tailored separately in order to best optimize the functioning of the company and their efficiency.

9 REFERENCES

Control Chart - Statistical Process Control Charts | ASQ (2022) *Asq.org*. Available at: https://asq.org/quality-resources/control-chart#:~:text=The%20control%20chart%20is%20a,are%20determined%20from%20historical%20data. (Accessed: 13 October 2022).

Process Capability Indices (2022) *Six-Sigma-Material.com*. Available at: https://www.six-sigma-material.com/Process-Capability-Indices.html (Accessed: 12 October 2022).

Bevans, R. (2022) *The p-value explained*, *Scribbr*. Available at: https://www.scribbr.com/statistics/p-value/ (Accessed: October 13, 2022).