ECSA Graduate Attribute Report for Data Analysis

T.S. Winter - 24126411

Report for: ECSA

Due date: 21 October 2022

Abstract

Data Analysis plays a very important role today. Most businesses across the globe make use of various forms of data. With increased competition through globalization, companies are constantly trying to get a competitive advantage in the market. One way to achieve this advantage is through data analysis. If a company can master their data, they will experience significant growth and increase in their profits. The following report will analyse the data of an online retailer by looking at the cleanliness, stability, reliability, and the similarities within the data.

Table of Contents

Table of Tables

Table of Figures

Introduction

An online retailer wants to improve their business and have collected various instances of data to analyse their current business processes. To help the online business, data analysis has been conducted. This data analysis included cleaning the data, describing the data, determining the stability of the data, optimizing the data, comparing the relationships between data features, and using statistics to make crucial business decisions.

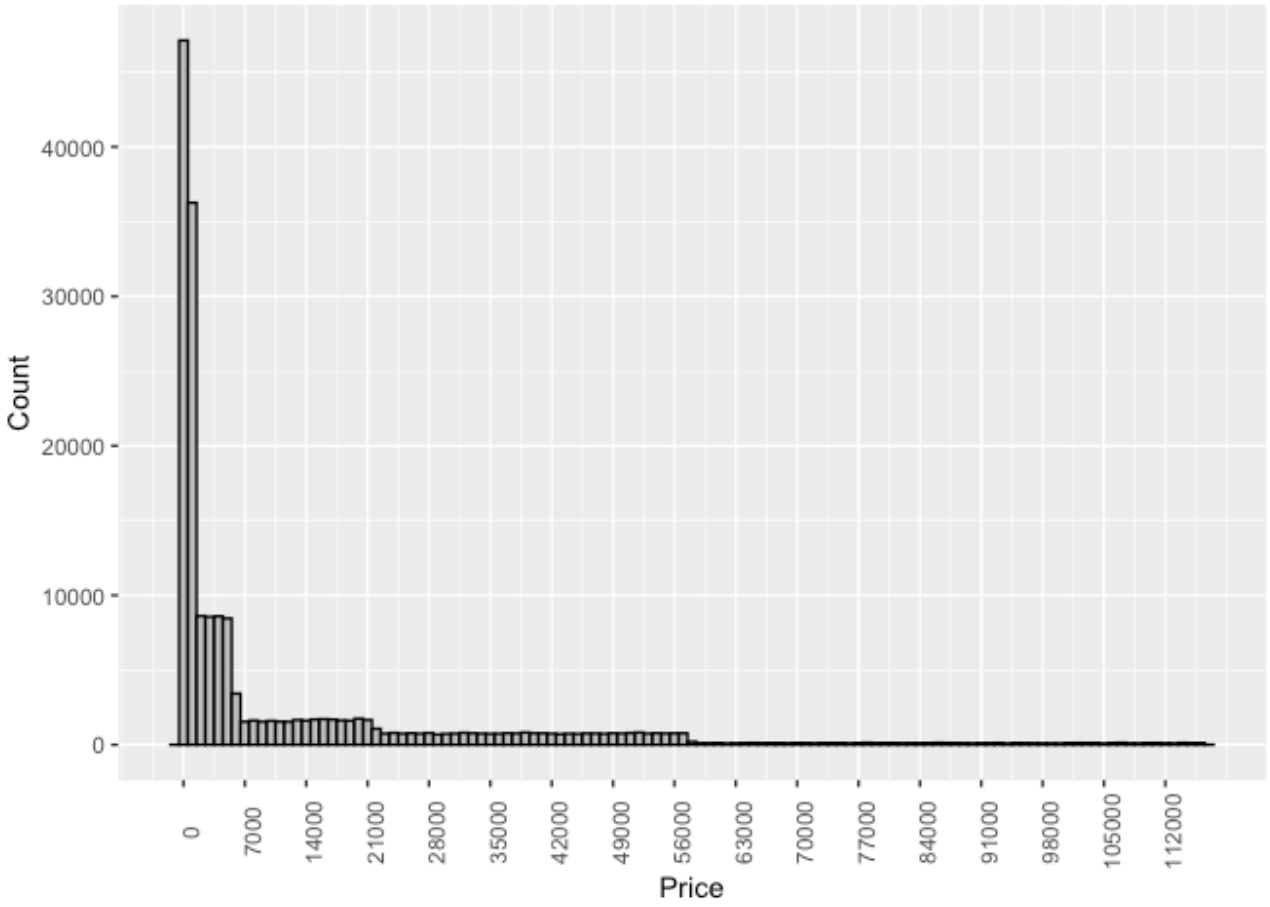
1. Data wrangling

The original sales data set contains numerous missing values indicated by NA. To perform data analysis these missing values must first be removed. To achieve this the complete cases function in R was used to filter the data into valid data (data containing no instances with missing values) and invalid data (data only containing the instances with missing values). After this separation the valid data set consisted of 179983 instances and the invalid data set consisted of 17 instances. Moreover, there are multiple negative values in the price subset. This is also a data quality issue and will be addressed in section 2.

2. Descriptive statistics

After evaluation of the data a few other data quality issues were determined.

A plot of Price vs Count



From the graph the most prices lie in a range between zero and seven thousand and the data is right-skewed. This indicates that lower priced products are sold more frequently at the company. Very notably, there are a few instances where the price is negative. A negative price for a product is not possible and an error was made when the data was accumulated. These negative prices must be removed through means of the subset function in R before further analysis can be conducted.

The valid dataset, consisting of 179978 instances and ten features, contains no more data quality issues and is ready for analysis. To determine the validity of the features and their importance a data quality report for the continuous features and categorical features will be constructed. This will be done by using the skimr package in R.

2.1 Data quality report: Continuous features

Feature	Count	% Missing	Cardi-nality	Min	1st Quartile	Mean	3rd Quartile	Max	Std. Dev.
X	179978	0	179978	1	45004	90003	135000	180000	51961
ID	179978	0	15000	11126	32700	55235	77637	99992	25740
AGE	179978	0	91	18	38	54.6	70	108	20.4
Price	179978	0	78832	35.6	482	12294	15271	116619	20889
Year	179978	0	9	2021	2022	2025	2027	2029	2.78
Month	179978	0	12	1	4	6.52	10	12	3.45
Day	179978	0	30	1	8	15.5	23	30	8.65
Delivery .time	179978	0	148	0.5	3	14.5	18.5	75	14.0

Table 1: Data quality report: Continuous features

As seen in the above data quality report, there are 8 continuous features. The feature X has a cardinality value equal to its count value. This feature is therefore not useful when analysing the data and is only used to distinguish between the different instances. The same can be said for feature ID, which has a high cardinality of 15000.

2.2 Data quality report: Categorical features

Feature	Count	% Missing	Cardi-nality	Mode	Mode Freq-uency	Mode %	2nd Mode	2nd Mode Freq-uency	2nd Mode %
Class	179978	0	7	Gifts	39149	21.752	Tech-nology	36347	20.195
Why. Bought	179978	0	6	Recom-mended	106985	59.443	Website	29447	16.36

Table 2: Data quality report: Categorical features

The two continuous features in our dataset are Class and Why.Bought. Both these features are free of missing values and data quality issues and are useful in gaining important information on the dataset.

From the table it is important to note the big difference in Mode % versus 2nd Mode % for the Why.Bought feature. Recommendations is by far the most dominant reason why people buy products at the company and the company should take note of this. Customer satisfaction should therefore be a high priority.

2.3 Histogram plots: Continuous features

In addition to the histogram plot made for the Price feature, these plots will also be created for other useful continuous features in the dataset to analyse the data and determine their distributions.

A plot of AGE vs Count

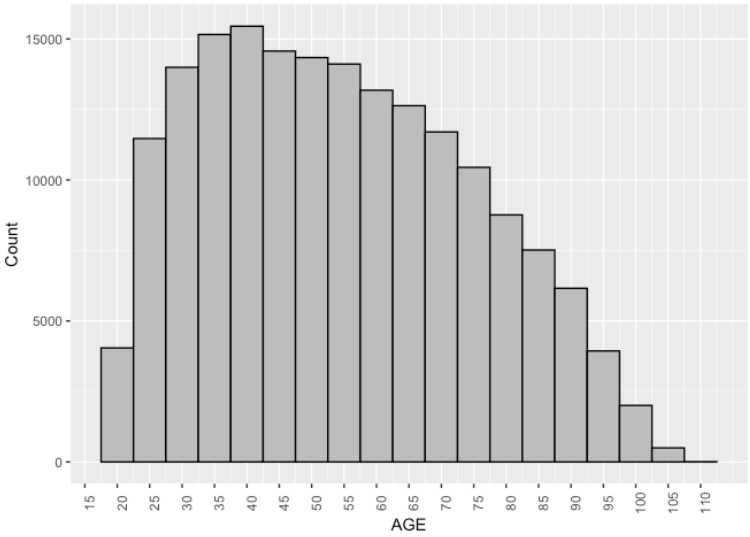


Figure 1: A plot of AGE vs Count

The plot for AGE vs Count indicates a unimodal distribution that is skewed to the right. The maximum age is a hundred and eight years. This is a very high age and it is likely due to an error when the data was collected or inserted by the user. The graph also indicates that younger users, between eighteen and twenty-three, have relatively low sales. This is potentially because they are not yet earning salaries and depend on their parents for money and payments. The peak sales are at an age of thirty-eight to forty-three and the business should therefore keep these customers satisfied and interested. From the age of forty-three onwards, sales start to decrease. The reason for this could be that older users are not as technologically inclined and would rather visit a store than buy a product online.

A plot of Delivery.time vs Count

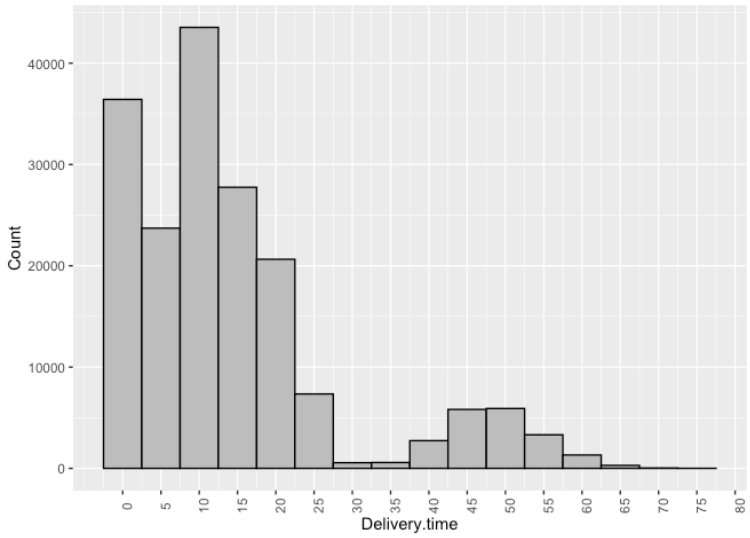


Figure 2: A plot of Delivery.time vs Count

The Delivery.time vs Count plot indicates a unimodal distribution that is skewed to the right. This is good for the company, since it indicates that their delivery times are usually low. From the graph there are various delivery times greater than forty. These values are greater than 1.5 times the inter quartile range from the third quartile and indicate potential outliers. The business should investigate these values and determine what cause these longer delivery times.

2.4 Sales versus time plots

Sales versus time plots give the business an indication of when to increase or decrease their stock levels, when to run promotions or a general understanding of how the business is performing.

The following plot shows the number of sales the business received in each year. From the graph the company made many sales in 2021. This value dropped with a large amount in 2022 but increased gradually and is the highest it has been over the past seven years in 2029. This indicates that the company is moving in the right direction and that they are making the correct business decisions.

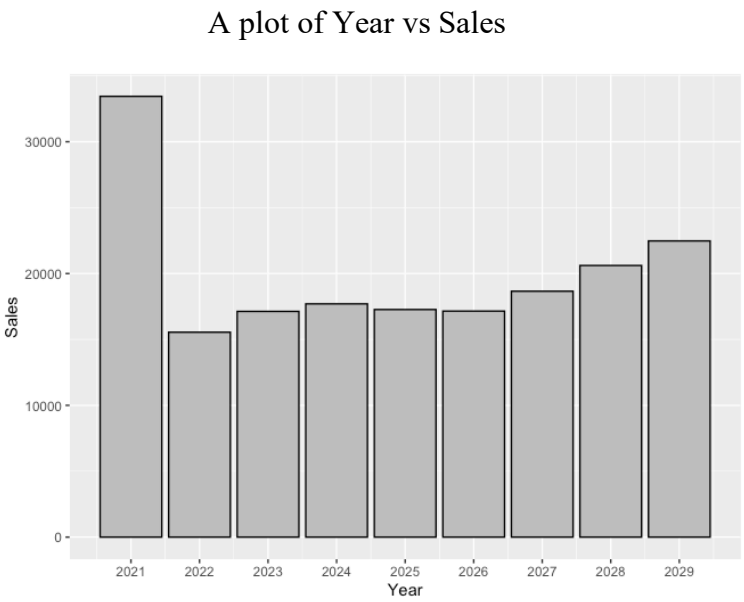


Figure 3: A plot of Year vs Sales

The plot for month versus sales is a uniform distribution. This plot shows that the number of sales the business generates is not determined by the month of the year, in other words there is no cyclical behavior in the data. This also shows that the company should keep equal amounts of stock for each month, since the sales per month are similar.

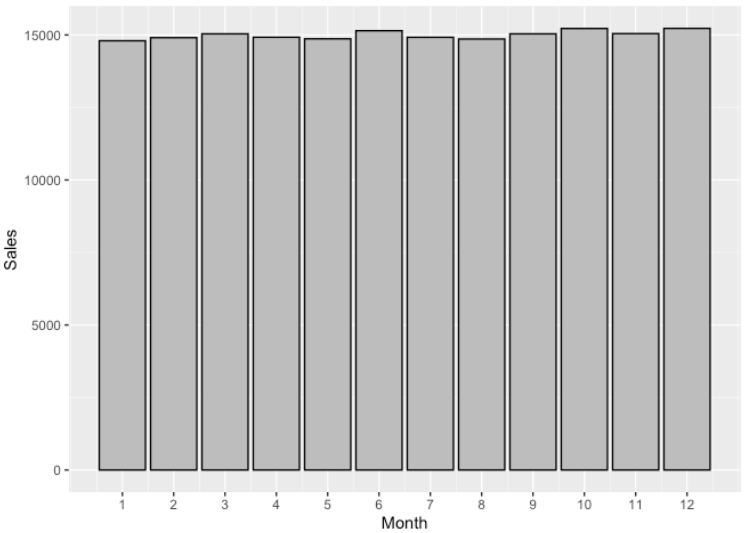


Figure 4: A plot of Month vs Sales

A distribution of sales for each class over 12 months

A further comparison that will help the business determine when to stock certain items is comparing class and month. This will help the business identify peak seasons for each item resulting in better planning and stock keeping. As seen in the graph, clothing and luxury items are uniformly distributed over the 12 months, since the median value is at 6 months. For the other items, more sales occur to the end of the year pulling their medians just above 6 months. This could potentially be because of holidays such as Christmas that happen to the end of the year. The business should keep enough stock during these days.

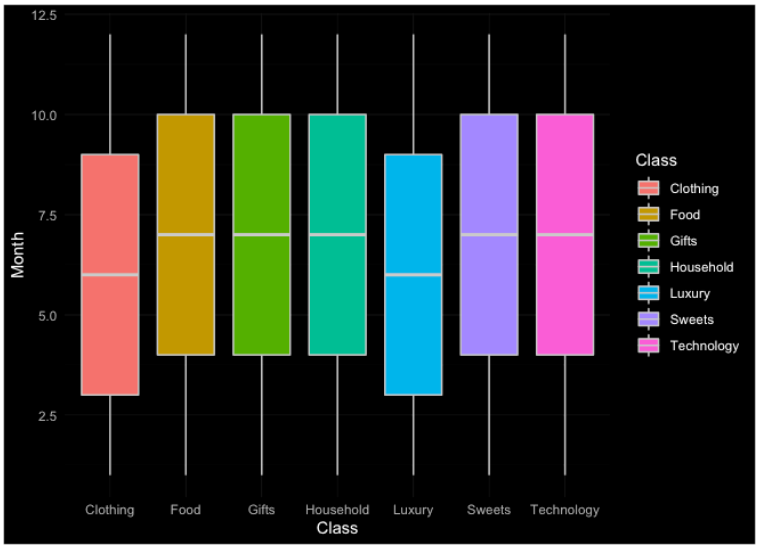


Figure 5: A distribution of sales for each class over 12 months

2.5 Class performances

The performance of each class must be tracked to determine whether the company should keep on selling these products or rather place more focus on specific products.

Units sold per class

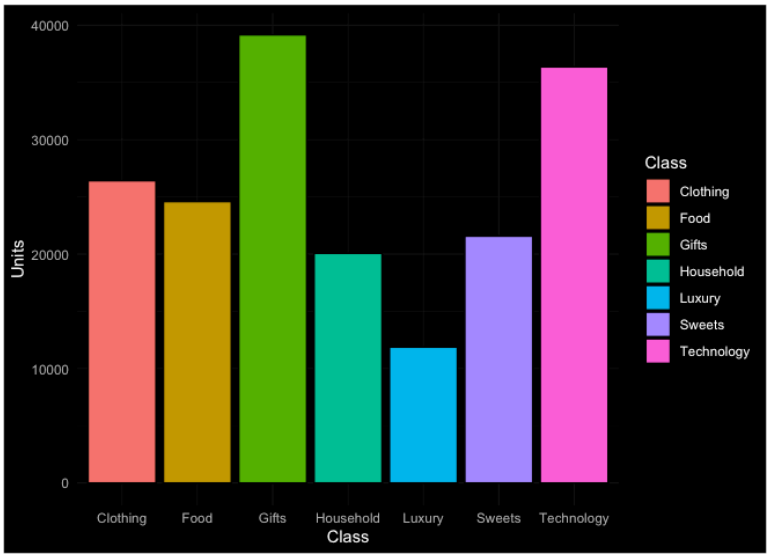


Figure 6: Units sold per class

Total revenue per class

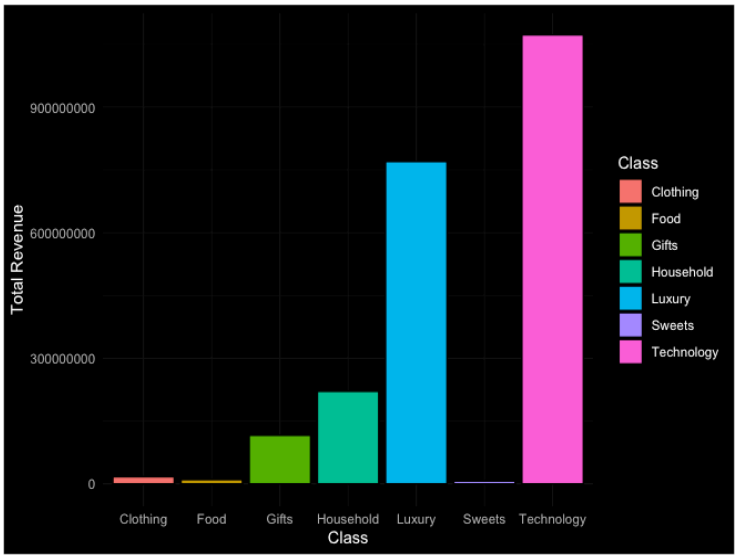


Figure 7: Total revenue per class

Figure 6 indicates that gifts and technology are the most popular items, with luxury being the item that has sold the least. Looking at figure 7 it is however clear that although a low number of sales for luxury items are made, the revenue generated from these items are high due to their expensive prices. Clothing, food, and sweets deliver low revenues and the business should consider to focus more on other items or replace these items with better performing products.

2.6 Further comparisons

A distribution of the delivery times for every class

From the graph, the delivery time for household appliances is the greatest followed by the delivery time for technology. This could be due to bad stock control of these items or slow suppliers. The business should determine the causes for these long delivery times and reduce it. The delivery times for the other items are low, consistent, and well controlled.

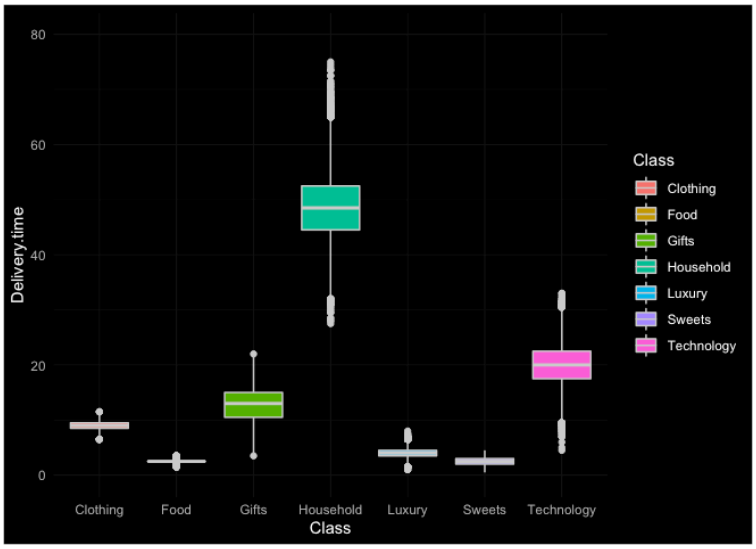


Figure 8: A distribution of the delivery times for every class

It is necessary to determine whether delivery time has an impact on the number of sales made for each product.

Sales for each class per year

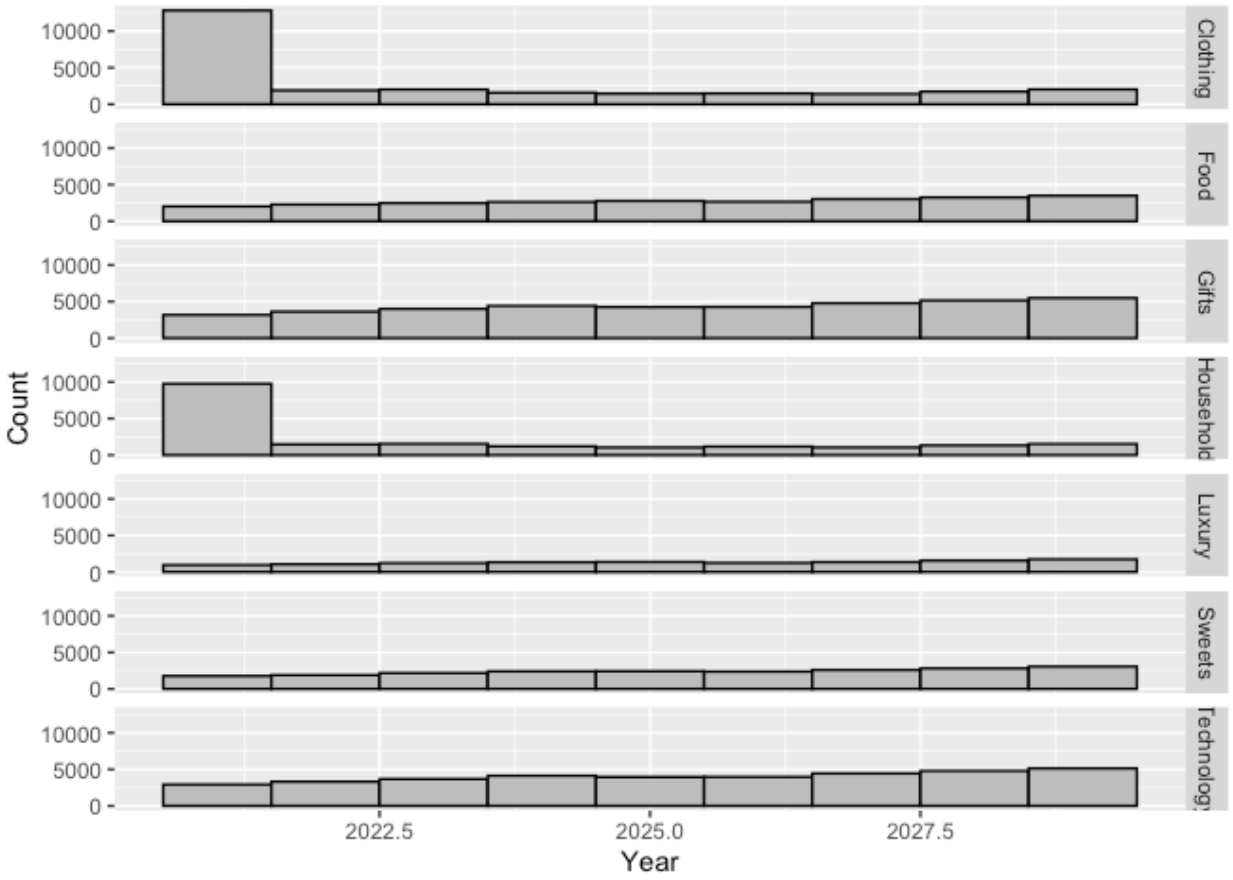


Figure 9: Sales for each class per year

The graph indicates the number of sales per class per year. As seen for the household items, the sales decreased over the years. This could potentially be due to the bad delivery times for this item. Users have realized that delivery for this item takes long and have decided to stop purchasing household items from this business. Looking at food items (items with good delivery times), the sales increased over the years. This indicates that short delivery times potentially result in more sales for a product. The company should therefore reduce the delivery times on their household and technology items to increase their sales.

When comparing the number of sales for each class with respect to age, the most graphs follow a uniform distribution that is right-skewed. This is because the most users of the online business tend to be younger, as determined previously in figure 1. This is however not the case for food and gift items. These items follow more of a normal distribution structure since parents are responsible for buying food and gifts for households. This data can help the business identify how to market certain products and to which users to suggest different items.

Number of sales for each class per age group

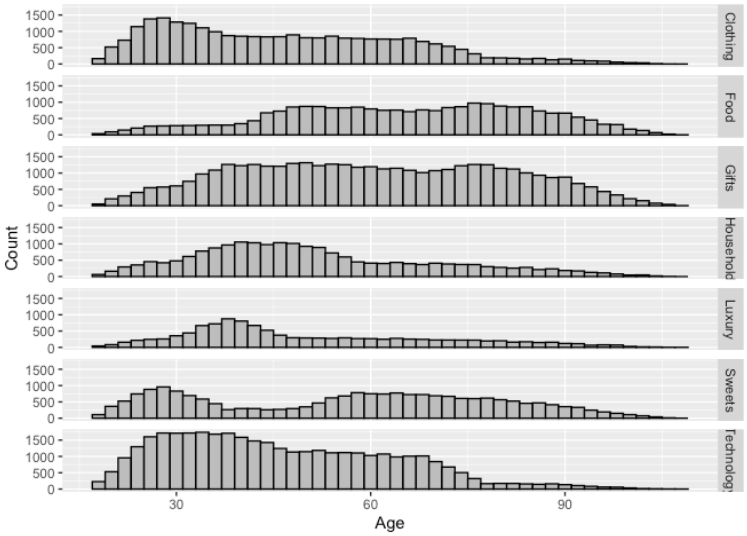


Figure 10: Number of sales for each class per age group

The reason why different ages bought products

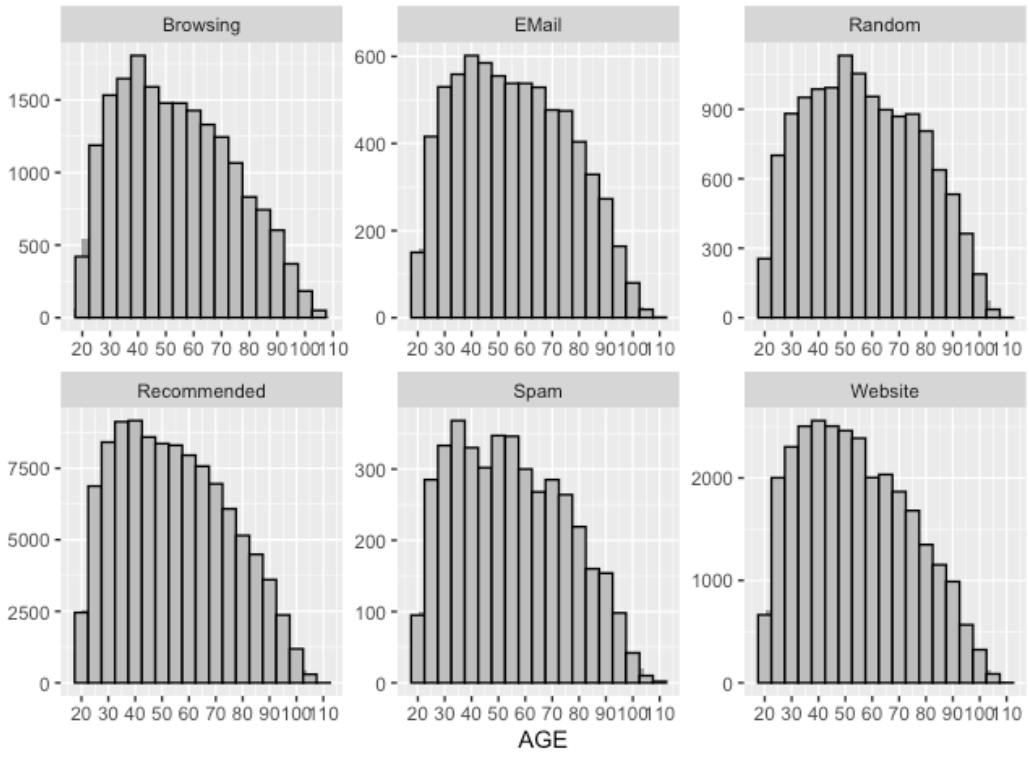


Figure 11: The reason why different ages bought products

The most popular reason why customers purchased products from the online business was because of recommendations. This is a clear indication that the business is satisfying their clients increasing their sales. Furthermore, all the plots have a similar right-skewed structure indicating that age does not have an influence on why customers purchase from the website.

2.7 Process capability indices

Process capability indices measure the variation in a process given certain process limits, in this case an upper specification limit (USL) of 24 hours and a lower specification limit (LSL) of 0 hours is used. An LSL of 0 hours is logical in this case, since process delivery times can't be lower than 0, these values can only assume positive numbers. Zero is therefore the lowest limit. Process capability indices can further be used as a measurement to compare processes and to determine whether these processes fulfil the necessary expectations.

USL = 24 hours
LSL = 0 hours
μ = 19.76053
σ = 3.501993

The Capability potential (Cp) and Capability performance (Cpk) are used to measure the ability of a process to meet certain requirements (mountz, 2018).

$Cp = (USL - LSL)/6\sigma$
$Cp = (24 - 0)/6*3.501993$
$Cp = 1.142$

This indicates that the spread of the values fit into the limits 1.142 times and the process is within the USL and LSL.

Cp does however not take the location of the process into account and further analysis, by determining the Cpk, should be done (mountz, 2018)
.

$Cpu = (USL - \mu)/3\sigma$
$Cpu = (24 - 19.76053)/3*3.501993$
$Cpu = 0.404$

$Cpl = (\mu - LSL)/3\sigma$
$Cpl = (19.76053 - 0)/3*3.501993$
$Cpl = 1.881$

$Cpk = \min(Cpl, Cpu)$
$Cpk = \min(0.4035293, 1.880884)$
$Cpk = 0.404$

This is a very low Cpk, indicating that the process is not very capable of meeting the necessary requirements. This can be due to the average not being close to the centre of the USL and LSL values and/or the standard deviation of the process is too large (mountz, 2018).

3. Statistical process control (SPC)

Statistical process control enables us to track real time behavior of our processes. In this business the delivery times for each class is tracked. This way, the business can quickly identify whether their processes are stable or not.

3.1 X&s-charts for every class of sale (First 30 samples)

3.1.1 X-chart table

The X-charts illustrate the trend in average delivery times for each class over time. The average delivery time for each sample must remain between an upper control limit and a lower control limit that is calculated for each class, when this is the case, the process is in control and stable.

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

Table 3: X-chart table

3.1.2 s-chart table

The s-charts illustrate the trend in the standard deviation of delivery times for each class over time. The standard deviation of delivery time for each sample must remain between an upper control limit and a lower control limit that is calculated for each class, when this is the case, the process is in control and stable.

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

Table 4: s-chart table

3.1.2 X&s-chart plots

X&s-chart plots were made for the first thirty samples of each class to track the stability of each process. The business can then use these plots do determine whether there are issues in their processes

and can rectify any mistakes immediately. Due to the X bar chart control limits being derived from s-bar values, out-of-control values in the s-chart will cause incorrect X bar chart control limits (Institute, 2022). The s-bar chart should therefore be examined first to identify any unstable behaviour. If any unstable behaviour is present, the cause of this behaviour should be identified and corrected to produce stable s-bar charts. Only once this has been done, the x-bar chart can be created.

3.1.2.1 Technology:

The s-bar chart for technology is stable and no samples pass the upper or lower limits. An accurate x-bar chart of delivery times for technology can therefore be derived.

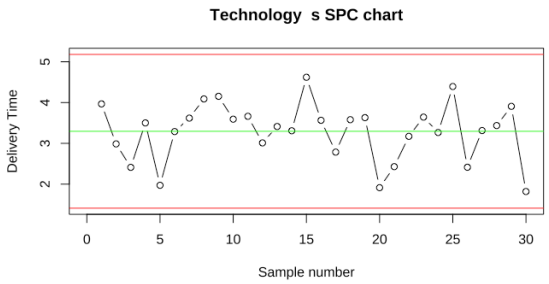


Figure 12: Technology s SPC chart

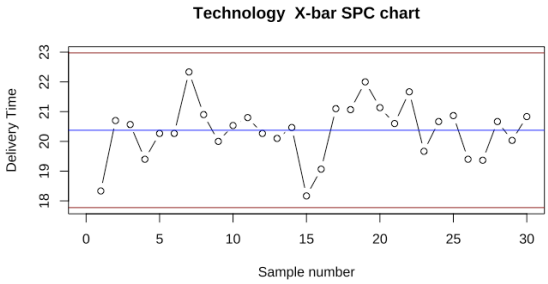


Figure 13: Technology X-bar SPC chart

3.1.2.2 Clothing:

The same can be said for the s-bar chart of clothing and its corresponding x-bar chart.

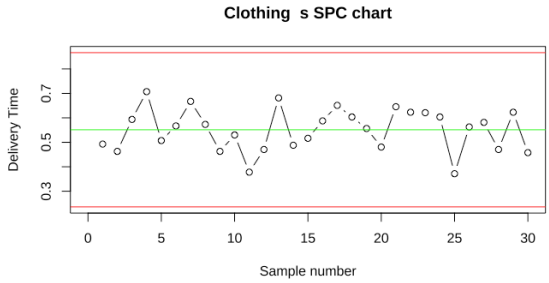


Figure 14: Clothing s SPC chart

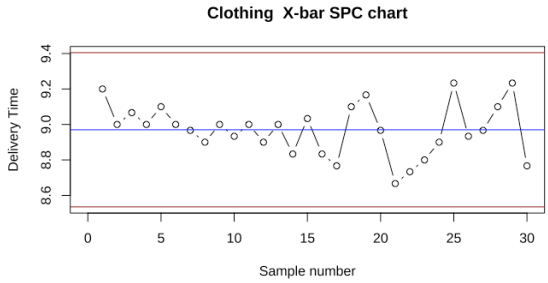


Figure 15: Clothing X-bar SPC chart

3.1.2.3 Household:

The 6th sample of the s-bar chart for the household class almost exceeds the upper limit, this chart is however still stable and produces an accurate x-bar chart.

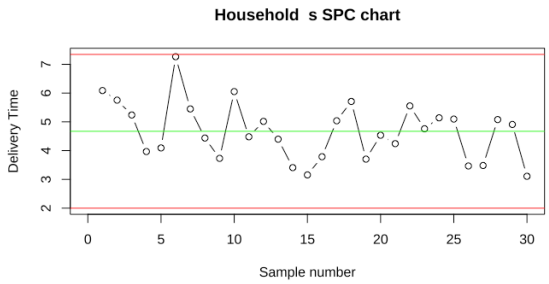


Figure 16: Household s SPC chart

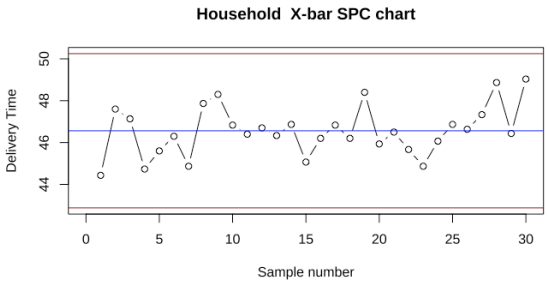


Figure 17: Household X-bar SPC chart

3.1.2.4 Luxury:

Like the s-bar charts of the previous three classes, the s-bar chart for luxury is stable and produces an accurate x-bar chart.

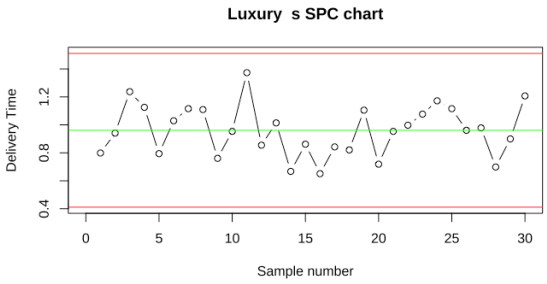


Figure 18: Luxury s SPC chart

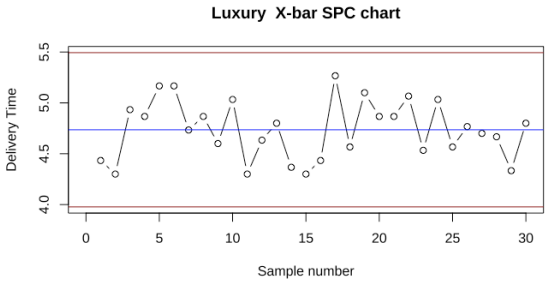


Figure 19: Luxury X-bar SPC chart

3.1.2.5 Gifts:

The same can be said for the s-bar and x-bar charts for the class gift.

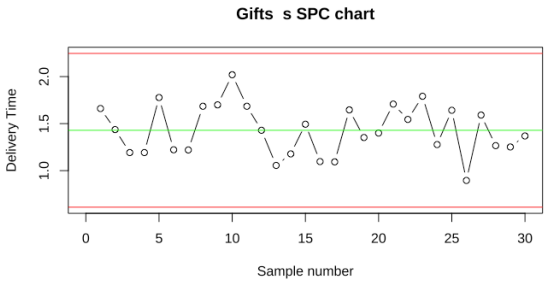


Figure 20: Gifts SPC chart

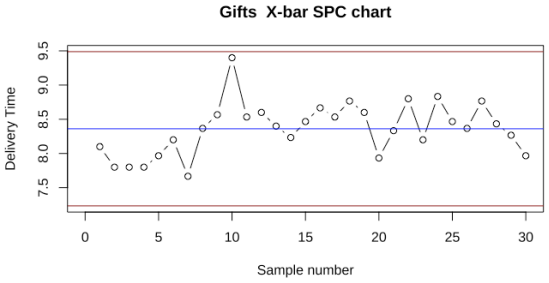


Figure 21: Luxury X-bar SPC chart

3.1.2.6 Food:

As seen on the s-bar chart, sample 19 for the food class is out of bounds resulting in an unstable s-bar chart. The cause of this instability must be tracked and rectified immediately. Due to this instability an inaccurate x-bar chart is created.

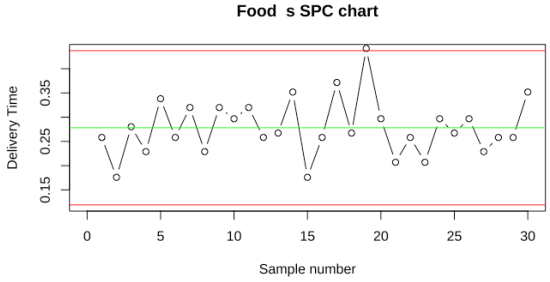


Figure 22: Food SPC chart

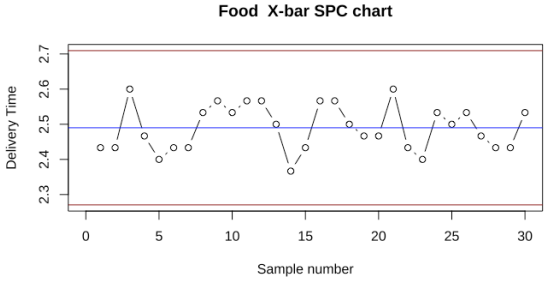


Figure 23: Food X-bar SPC chart

3.1.2.7 Sweets:

Like the food class, the sweets class is also unstable, except that in this case the 18th sample is out of bounds causing the instability.

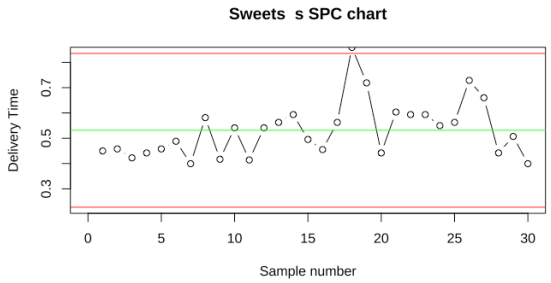


Figure 24: Sweets SPC chart

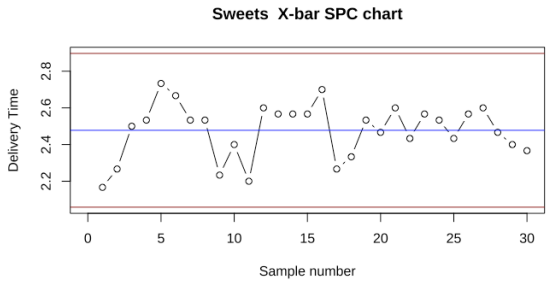


Figure 25: Sweets X-bar SPC chart

3.2 X&s-charts for every class of sale (All samples)

After plotting X&s-charts for the first thirty samples, further X&s-charts of all the data instances for every class were created.

3.2.1 Technology

The X&s-charts for the technology process follow the necessary rules to be considered in control, despite a few samples that are considered to be unstable, since they exceed the limits.

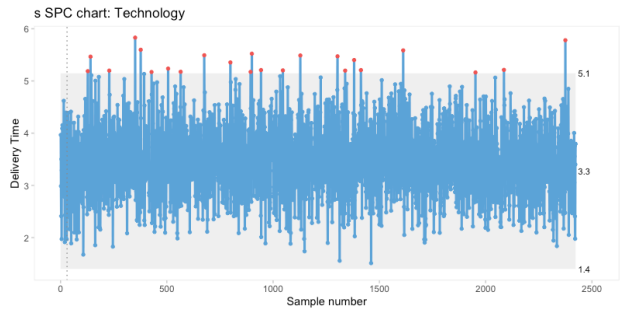


Figure 26: s SPC chart: Technology

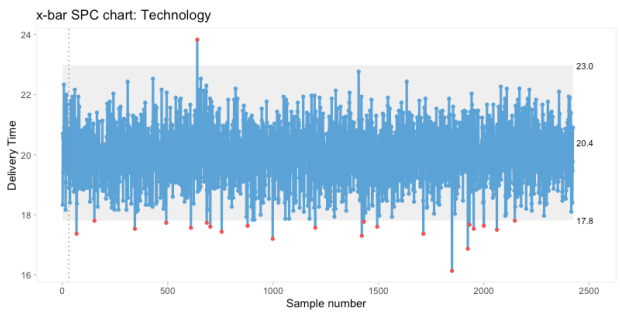


Figure 27: x-bar SPC chart: Technology

3.2.2 Clothing

Although the clothing process is less stable than the technology process due to more frequent unstable instances, the X&s-charts for the clothing process largely follow the necessary rules to be considered in control.

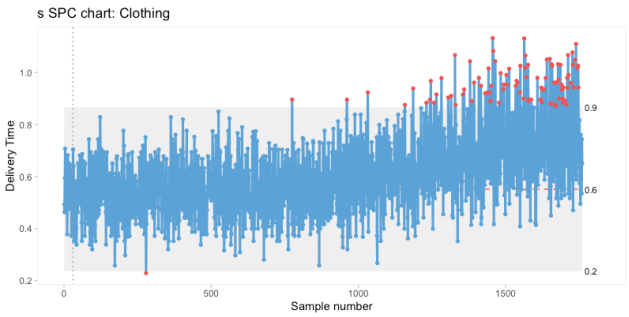


Figure 28: s SPC chart: Clothing

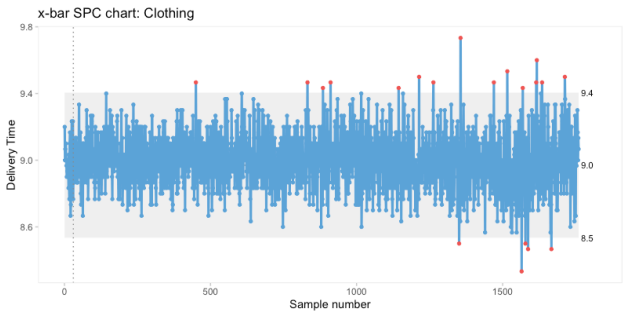


Figure 29: x-bar SPC chart: Clothing

3.2.3 Household

The X&s-charts for the household process breaks some of the necessary rules to be considered in control. Consecutive number of data points fall outside of the control limits and are on one side (above) the average as seen in the X-bar chart. The household process is therefore unstable.

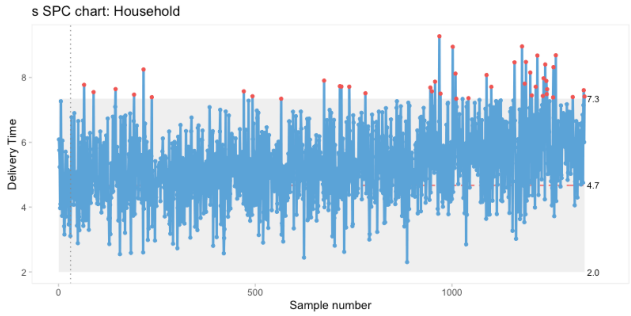


Figure 30: s SPC chart: Household

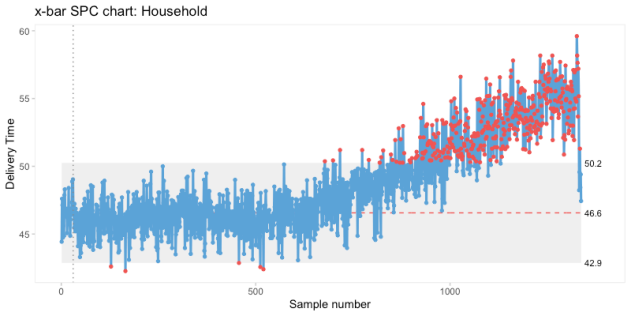


Figure 31: x-bar SPC chart: Household

3.2.4 Luxury

Like, the X&s-charts for the household process, the charts for the luxury process also break the rules that define a process as being in control. In this case, a consecutive number of data points fall outside of the control limits and are on one side (below) the average as seen in the X-bar chart. The luxury process is therefore unstable.

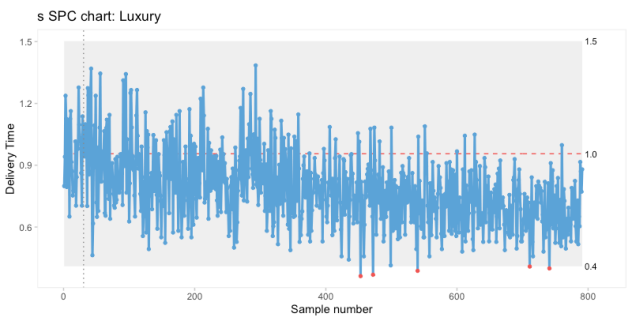


Figure 32: s SPC chart: Luxury

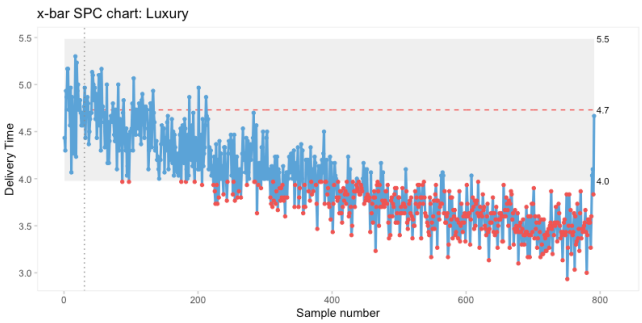


Figure 33: x-bar SPC chart: Luxury

3.2.5 Gifts

The process for class gifts follows a similar trend to the household class and is therefore also out of control.

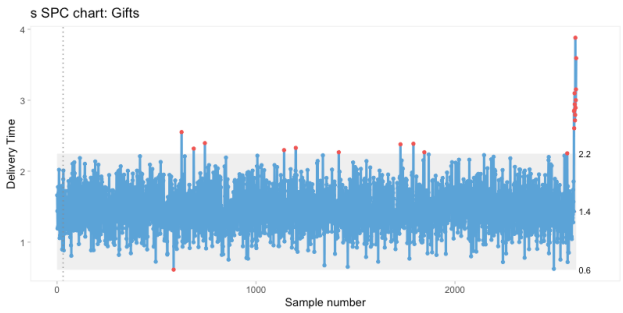


Figure 34: s SPC chart: Gifts

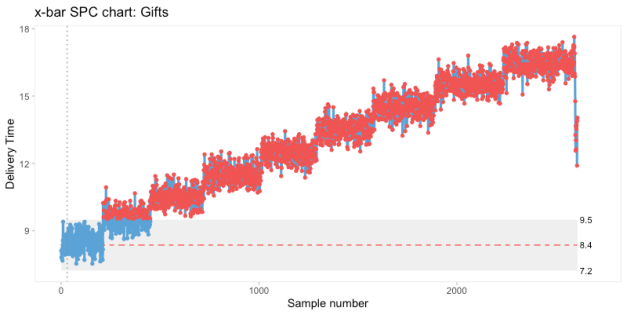


Figure 35: x-bar SPC chart: Gifts

3.2.6 Food

The X&s-charts for the food process follow the necessary rules to be considered in control, despite a small number of instances falling outside of the required specifications.

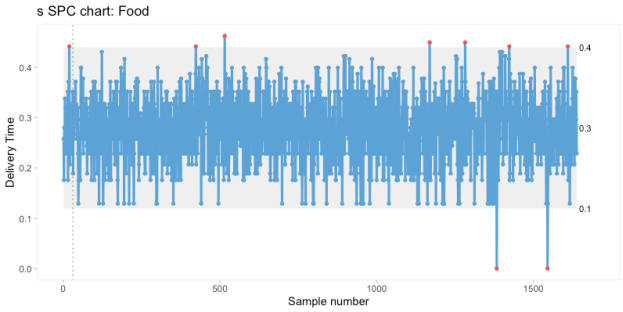


Figure 36: s SPC chart: Food

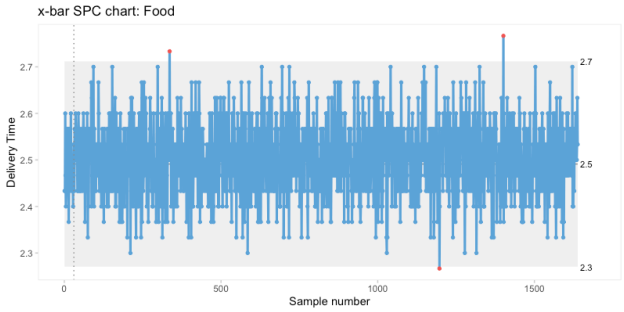


Figure 37: x-bar SPC chart: Food

3.2.7 Sweets

The X&s-charts for the sweets process follow the necessary rules to be considered in control and is the process with the least amount of unstable instances.

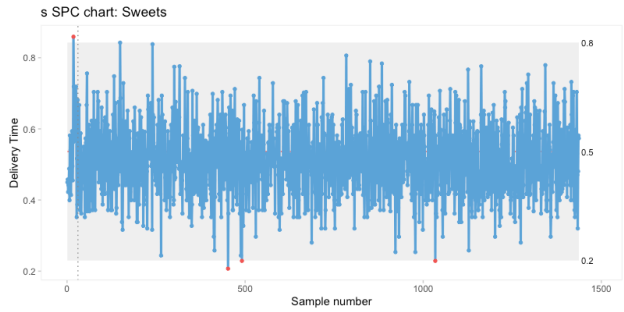


Figure 38: s SPC chart: Sweets

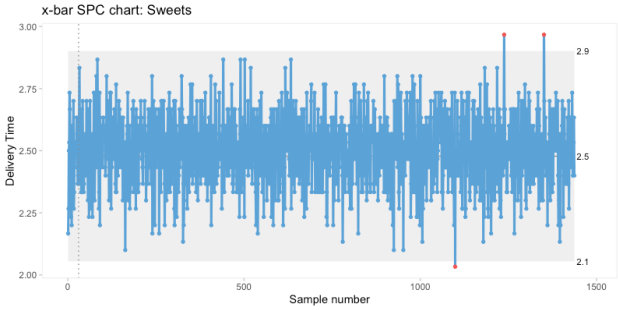


Figure 39: x-bar SPC chart: Sweets

The food and sweets processes were both considered unstable when looking at the first 30 samples. However, after reviewing the X&s-charts over all the data instances, the processes were mostly stable.

On the other hand, household, luxury and gifts which were initially stable for the first 30 samples are considered out-of-control when reviewing all the data instances. Improvements should be made to these three processes to stabilize them.

4. Optimising the delivery processes

4.1 Sample numbers that gave indications of out of control

4.1.a Sample means outside of the outer control limits

The following table indicates the number of sample means outside the control limits for each class. From the table it is clear that gifts has the most sample means outside of the control limits and is therefore the least stable process. On the other hand food and sweets has the least amount of sample means outside the control limits and are therefore the most stable processes. From the table it is also clear that the household, luxury and gifts classes are the three processes that need to be improved. This was also confirmed by the X&s-charts of all the samples.

Class	Total found	1st	2nd	3rd	3rd Last	2nd Last	Last
Clothing	20	450	832	885	1635	1667	1713
Household	392	128	165	457	1330	1331	1334
Food	3	336	1197	1401	NA	NA	NA
Technology	23	67	152	344	2000	2062	2147
Sweets	3	1099	1238	1351	NA	NA	NA
Gifts	2289	212	215	217	2607	2608	2609
Luxury	441	87	97	175	786	787	790

Table 5: Sample means outside of the outer control limits

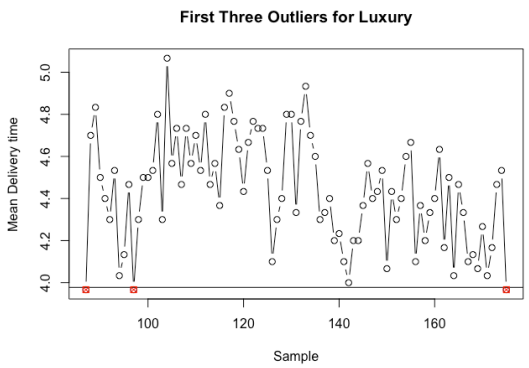


Figure 40: First three outliers for luxury

Figure 41: Last three outliers for luxury

The following two graphs for the delivery times of luxury confirm the data in the table. As seen in the graphs the red dots are the first three and last three outliers for luxury, since they are below the lower-class limit. These red dots correspond to the values in the table which is sample: 87, 97 and 175 as the first three outliers and sample: 786, 787 and 790 as the last three outliers.

4.1.b Most consecutive samples of "s-bar or sample standard deviations" between -0.3 and +0.4 sigma-control limits

As seen in the table, technology and gifts had the most number of samples in the sigma range between -0.3 and +0.4. Both technology and gifts had six samples in this sigma range respectively, of which sample 1191 and sample 1334 were the last samples to be given in the range respectively. Luxury only had three samples in this range, which is the lowest between the seven classes.

Class	maximum between -0.3 & 0.4 sigma	Position of first	Position of last
Clothing	5	665	665
Household	4	253	761
Food	5	752	905
Technology	6	1191	1191
Sweets	5	692	692
Gifts	6	1334	1334
Luxury	3	230	230

Table 6: Most consecutive samples of "s-bar or sample standard deviations" between -0.3 and +0.4 sigma-control limits

To confirm that the values indicated in the table is correct, graphs for both technology and gifts were created in the correct ranges. As seen on the graphs, both technology and gifts have 6 red samples between the two black lines that indicate the +0.4 and -0.3 control limits respectively. This shows that both of them have six consecutive instances between the specified sigma limits.

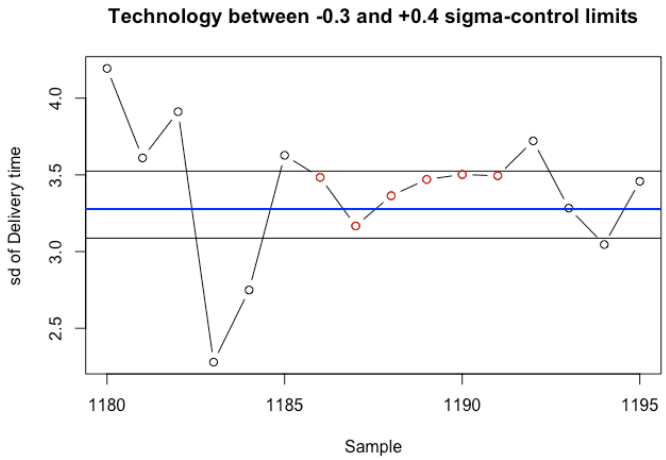


Figure 42: Technology between -0.3 and +0.4 sigma-control limits

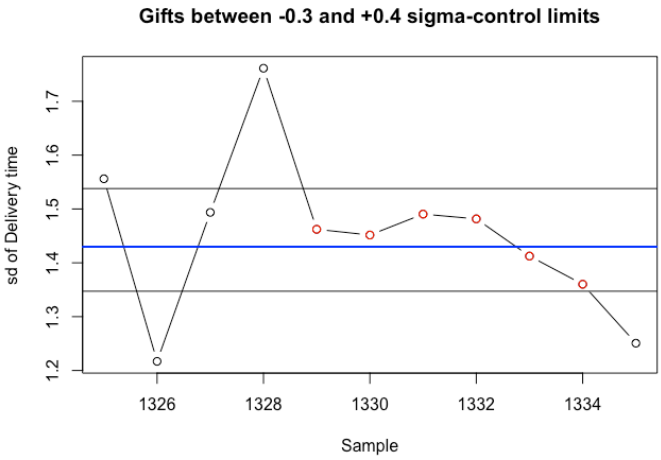


Figure 43: Gifts between -0.3 and +0.4 sigma-control limits

4.2 Type I (Manufacturer's) Error

A type I error or false-positive occurs when the null hypothesis is rejected, when in fact it should have been accepted (Bhandari, 2022). The null-hypothesis (Ho) for the process is: "the process is in control and centred on the centreline calculated using the first 30 samples" and the alternative

hypothesis (Ha) is: “is that the process in not in control and has moves from the centreline or has increased or decreased in variation.”.

4.2.a,b
In the case of (a), a type I error will occur when a process is in control, but it is identified as out-of-control. In other words, although the sample lies between the lower and upper control limits, we declare that it lies outside of these limits. As seen in the table, there is a 0.27% chance that we are making this type I error and accepting the alternative hypothesis instead. This is very low and the chances are therefore small. This value must be compared with the alpha value of this process and based on this relationship, the null-hypothesis is accepted or rejected.

In the case of (b), a type I error will occur when a value lies between -0.3 and +0.4 sigma-control limits, but we declare that it lies outside of these limits. There is a 72.67% chance of this error happening. This is a very high probability and this error is therefore likely in the case of (b). This value must also be compared with the alpha value of this process and based on this relationship, the null-hypothesis is accepted or rejected.

Rule	Probabilities	Probability %
A	0.00269979606326019	0.269979606326019
B	0.726666836200723	72.6666836200723

Table 7: Type I errors for A and B

4.3 Best profit for Technology’s delivery time

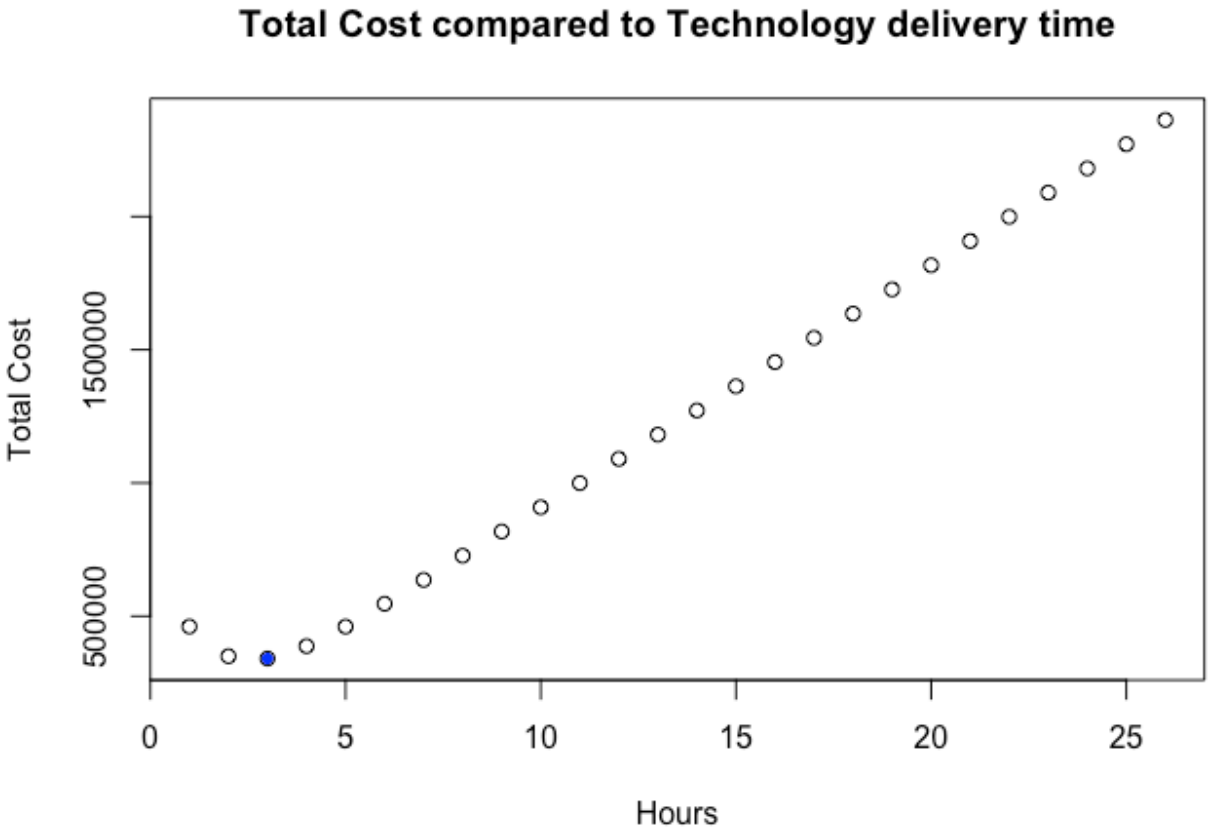


Figure 44: Total Cost compared to Technology delivery time

The following graph compares the average delivery time in hours for the technology class and the cost it is accompanied with. As seen on the graph the lowest cost, and therefore the highest profit will occur when the delivery times for technology is centred around three hours. This will result in a minimum cost of R340870. The current average delivery time for technology is 20.01095 hours and results in a cost of R1817350. This value must be reduced to 3 hours to improve the profit of the business. This is similar to Taguchi loss, since any increase or decrease in this value (three hours) will create dissatisfaction (higher costs).

4.4 Type II (Consumer's) Error

A type II error or false-negative occurs when the null hypothesis is accepted, when in fact it should have been rejected (Bhandari, 2022). This error can be reduced by reducing the statistical power of the process.

4.4.a
In the case of (a), a type II error will occur when a process is out-of-control, but it is identified as in control. In other words, although the sample lies outside the lower and upper control limits, we declare that it lies between these limits. As seen on the graph, there is a 48.82% chance that we are making this error, which is fairly high. The business should be vigilant, since delivery times can be longer without the business noticing. This can lead to unhappy clients.

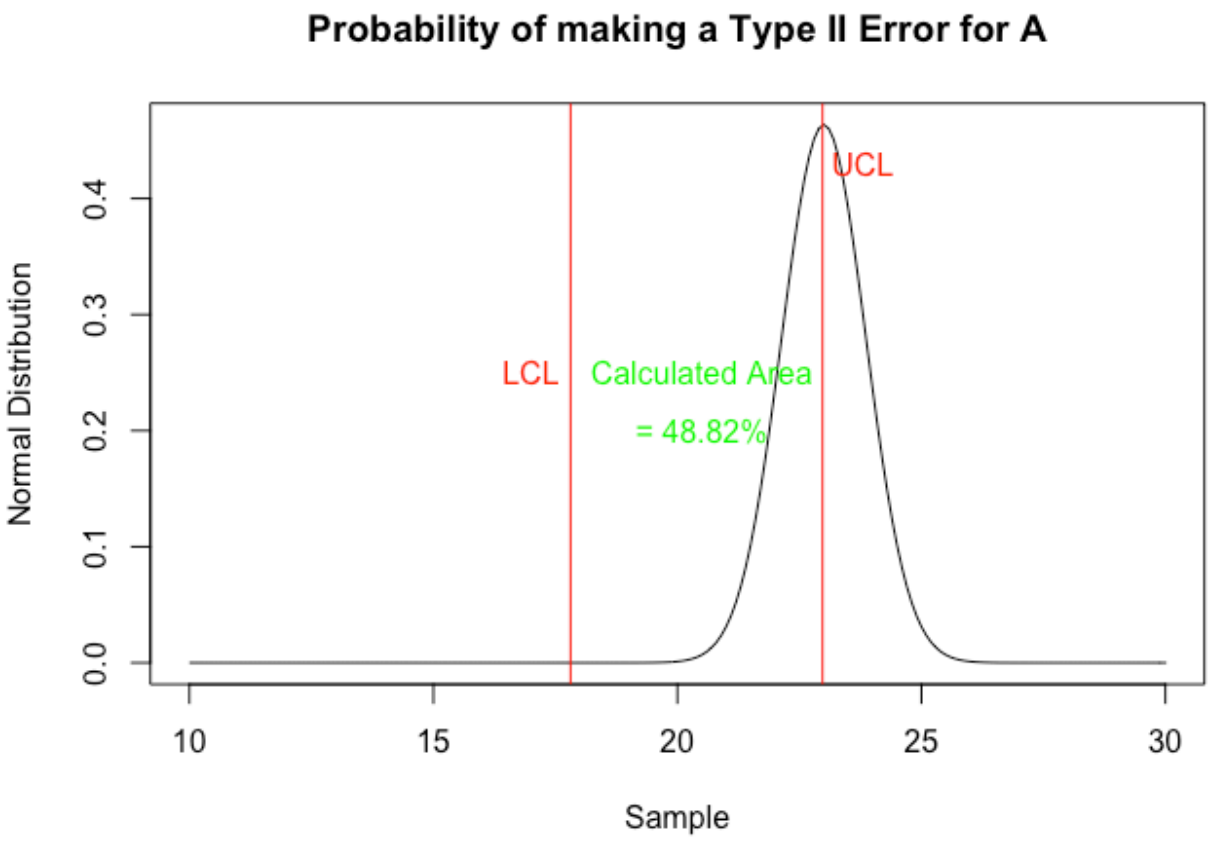


Figure 45: Probability of making a Type II error for A

5. DOE and MANOVA

A test will be conducted to determine whether the class of a product has an influence on its price and delivery time.

For the test the null-hypothesis (Ho) is: The type of class does have an influence on the price and delivery times of a product

The alternative hypothesis (Ha) is: Class has no influence on the price and delivery times of a product

In order to conduct the test a MANOVA table was created with Price and Delivery.times as the two dependent variables and Class as the independent variable. After crating the MANOVA table, the p value (practically zero) was smaller than the selected alpha value of 0.05 as seen in the table. This means that the null-hypothesis can be accepted and class does indeed influence the price and delivery time.

	Df	Pillai	approx F	num Df	den Df	Pr(>F)	
independent_var	6	1.6797	157291	12	359942	< 2.2e-16	***
Residuals	179971						

Table 8: MANOVA table

To prove this result a post-hoc test is conducted.

Figure 46: The delivery time and price for each class

As seen in the graph, products from the Household class tend to have longer delivery times and products from the Luxury class tend to have higher prices. This corresponds to the null-hypothesis

that the class does indeed have an influence on delivery times and the prices of products. Also see figure 8, which confirms that certain classes tend to have longer delivery times than others.

6. Reliability of the service and products

6.1 Do Problem 6 and 7 of chapter 7

6.1.1 Problem 6

A blueprint specification for the thickness of a regrigerator part at Cool Food, Inc. is 0.06 ± 0.04 centimeters (cm). It costs $45 to scrap a part that is outside the specifications. Determine the Taguchi loss function for this situation.

T = 0.06 (Quality dimension target)
LSL = 0.06-0.04 (Lower Specification Limit)
USL = 0.06 + 0.04 (Upper Specification Limit)

Taguchi Loss Funtion:

$L(x) = k(y - m)^2$
$45 = k(0.04)^2$
$k = 45/(0.04^2)$
$k = 28125$

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

6.1.2 Problem 7

A team was formed to study the refrigerator part at Cool Food, Inc. described in Problem 6. While continuing to work to find the root cause of scrap, they found a way to reduce the scrap cost to $35 per part.

(a) Determine the Taguchi loss function for this situation.
(b) If the process deviation from target can be reduced to 0.027 cm, what is the Taguchi loss?

(a) $L(x) = k(y-m)^2$
$35 = k(0.04)^2$
$k = 35/(0.04^2)$
$k = 21875$

$L(x) = 21875(y-0.06)^2$

(b) $L(x) = k(y-m)^2$
$L(0.027) = 21875(0.027)^2$
$L(0.027) = 15.95$

6.2 Problem 27 of chapter 7

Magnaplex, Inc. has a complex manufacturing process, with three operations that are performed in series. Because of the nature of the process, machines frequently fall out of adjustment and must be repaired. To keep the system going, two identical machines are used at each stage; thus, if one fails,

the other can be used while the first is repaired. The reliabilities of the machines are as follows:
Machine Reliability A = 0.85 B = 0.92 C = 0.90

(a) Analyze the system reliability, assuming only one machine at each stage (all the backup machines are out of operation).
(b) How much is the reliability improved by having two machines at each stage?

(a) System reliability = 0.85*0.92*0.9
System reliability = 0.7038

(b) Machine A reliability with two machines: 1 - (1 – 0.85)^2 = 0.9775
Machine B reliability with two machines: 1 - (1 – 0.92)^2 = 0.9936
Machine C reliability with two machines: 1 - (1 – 0.90)^2 = 0.99

New system reliability = 0.9775*0.9936*0.99
New system reliability = 0.96153

As determined above, without backup machines, the business will have a system reliability of 70.38% and with backup machines, the business will have a system reliability of 96.15%. The backup machines therefore make a significant difference in the reliability but add several extra costs. The business should therefore investigate only adding one or two backup machines.

If only machine A had a backup: System reliability: 0.9775*0.92*0.9 = 0.80937

Which is a 10% improvement compared to having no backup machines. This will be the best solution for the business.

6.3 Reliable delivery times

Currently the business has twenty vehicles available. With these twenty vehicles, the business is guaranteed to have 46.97018 ≈ 50 days in a year where all twenty vehicles are available. If the business increases the number of vehicles to twenty-one, there will be 314.5344 ≈ 315 days where all twenty-one vehicles are available. This will ensure a vast improvement in reliable delivery and the business should therefore opt for an extra vehicle.

When looking at the drivers, the total number of days per year when we have all 21 drivers available is x days per year.

By combining these two solutions, the number of days when reliable delivery times can be expected for twenty vehicles and twenty-one vehicles can be calculated.

For twenty vehicles: (46.97018*341.0693)/(365*365)*365
= 43.89 ≈ 44 days

For twenty-one vehicles: (314.5344*341.0693)/(365*365)*365
= 293.91 ≈ 294 days

The business should buy an extra vehicle to improve the reliability of their deliveries.

Conclusion

To help the online business through means of data analysis, the data was first cleaned by removing all missing values and negative data instances. Thereafter, the distribution of certain features and their correlations were discussed to gain a deeper understanding of the data. The stability of each class was then determined and out-of-control processes were highlighted. For the fourth part of the report, the cost of delivery for technology was optimised and the different types of statistical errors were discussed. Thereafter, in part five the proportionality or lack thereof between features were discussed. Finally, various statistical problems were solved to make crucial business decisions.

Bibliography

mountz, 2018. *Understanding and Using Cpk.* [Online]
Available at: https://www.mountztorque.com/Understanding-and-Using-Cpk#:~:text=What%20is%20Cpk%3F,to%20around%20your%20average%20performance.
[Accessed 10 October 2022].

Institute, S., 2022. *Six Sigma DMAIC Process - Control Phase - SPC - Out of Control.* [Online]
Available at: https://www.sixsigma-institute.org/Six_Sigma_DMAIC_Process_Control_Phase_SPC_Out_Of_Control.php
[Accessed 12 October 2022].

Bhandari, P., 2022. *Type I & Type II Errors | Differences, Examples, Visualizations.* [Online]
Available at: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/#:~:text=In%20statistics%2C%20a%20Type%20I%20error%20means%20rejecting%20the%20null,hypothesis%20when%20it%27s%20actually%20false.
[Accessed 15 October 2022].

Radečić, D., 2022. *MANOVA in R – How To Implement and Interpret One-Way MANOVA.* [Online]
Available at: https://www.r-bloggers.com/2022/01/manova-in-r-how-to-implement-and-interpret-one-way-manova/#:~:text=Interpret%20MANOVA%20in%20R%20With%20a%20Post%2DHoc%20Test,-The%20P%2DValue&text=By%20doing%20so%2C%20we%27ll,independent%20variable%20%E2%80%93%20the%20
[Accessed 16 October 2022].