Quality Assurance ECSA Project

Name: Lewis Seymour

Student number: 23704659

For the department of Industrial Engineering

Abstract

This report is about analysing sales data provided by the company and determining different relationships, correlations and distributions within the data. The report goes deeper into the delivery time's associated with each class and statistical process control is performed on this data. It is determined that 'Household' and 'Gift' delivery times need to be looked at closer as it contains many samples outside of the control limits. Type I and II analysis is also being done on various aspects of the data. There is also a section dedicated to optimising the delivery process in which it is determined that the mean delivery time should be decreased by 3 hours to have the lowest cost loss. MANOVA analysis is also done on the data, determining whether there are relationships between 'Price' and 'Delivery time' features relating to the 'Class' and 'Why.Bought' features, which is determined to be dependent on each other. Lastly there is a focus on service reliability and determining whether backup machines are required for 'Technology' product manufacturing as well as how many days the business can expect reliably deliveries in a year, which is determined to be 293 days in a year.

List of figures:

List of Tables:

Table of Contents

Introduction:

The focus of this report is to analyse the data provided by the sales department by looking at different relationships between features and comparing them to each other. It is also crucial to make important deductions from the data which can aid management in making decisions. The importance lies in a general understanding of what the data is saying and not focused on specific values, but rather on the bigger picture.

Thus this report contains an explanation and preparation of the data that was provided . This data will then be used to build descriptive statistics models which will allow for the understanding of trends, distributions and variations in the data provided. There will then be an emphasis on statistical process control of the delivery times for the technology products sold. From this statistical process control the deductions made will be used to try and optimise the delivery process as well as determine the Type I error relating these deductions. In this section there will also be a solution for optimizing the delivery cost given certain penalty costs. Furthermore a MANOVA will be done on the data, determining whether there are relationships between 'Price' and 'Delivery time' features relating to the 'Class' and 'Why.Bought' features. Lastly there will be a focus on service reliability and determining whether backup machines will be required for 'Technology' product manufacturing as well as how many days the business can expect reliably deliveries in a year.

1. Part 1: Data Wrangling

The data that was received from the sales department contains a lot of information that will be able to be used for analysis and decisions. The data first needs to be reordered and cleaned to enable for accurate analysis to be performed on the data. It is important to understand what measures are taken to create the data to be used for analysis as it impacts the credibility of the analysis.

1.1 Initial data:

The data received from the sales department came in the form of a column separated vector file and contains valuable information about the business and its sales. The initial data is randomly ordered and contains 180000 instances with 10 features.

The 10 features are:

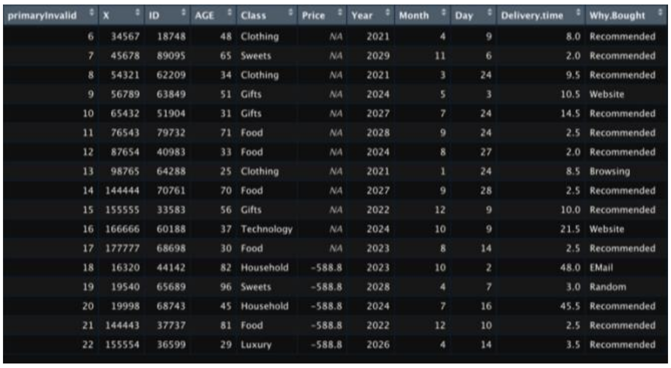
- X – a class index
- ID – Identification number of the client
- Age – what age the clients are
- Class – the type of product bought by client
- Price – price of the product
- Year – the year that the product was bought
- Month – the month that the product was bought
- Day - the day that the product was bought
- Delivery time – the time it takes to deliver the product to the client
- Why bought – the reason that client bought the product

The initial data however contains some incomplete instances which need to be taken care of and thus the data needs to be cleaned.

1.2 Validating data:

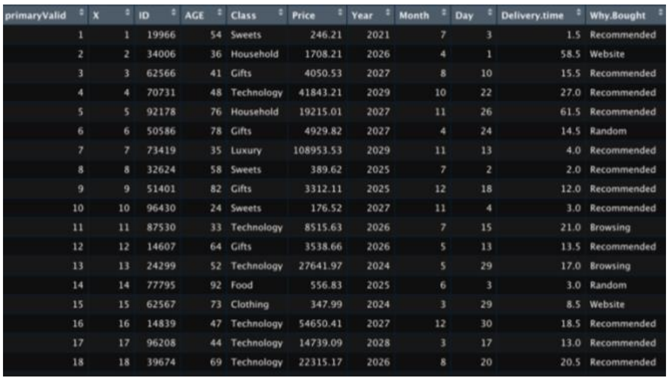
The initial data is tested to determine which instances are invalid and a separate dataset is made with these invalid instances. The instances observed to be invalid are instances which contain missing values as well as negative price values.

The initial data contains 17 instances with invalid events and 5 instances with negative price values. Therefore there are a total of 179978 valid data instances from which valuable information can be determined from. These 22 invalid instances are removed and added to a new dataset for invalid data and a new dataset is also created for the valid data. These datasets are given new index numbers to allow for proper analysis of data. Below an extract of the valid and invalid data can be seen as well as their new index numbers (primaryInvalid/ primaryValid) compared to their old index numbers (X).



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

Table 1: Extract of invalid dataset with 22 instances



primaryValid	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
12	12	14607	64	Gifts	3538.66	2026	5	13	13.5	Recommended
13	13	24299	52	Technology	27641.97	2024	5	29	17.0	Browsing
14	14	77795	92	Food	556.83	2025	6	3	3.0	Random
15	15	62567	73	Clothing	347.99	2024	3	29	8.5	Website
16	16	14839	47	Technology	54650.41	2027	12	30	18.5	Recommended
17	17	96208	44	Technology	14739.09	2028	3	17	13.0	Recommended
18	18	39674	69	Technology	22315.17	2026	8	20	20.5	Recommended

Table 2: Extract of valid dataset with 179978 instances

2. Part 2: Descriptive statistics

In this section descriptive statistics is done on the valid dataset that was created. This information allows the business to make very important deductions from the data that is analysed. The data contains categorical and continuous features which both allow for important classification and understanding of the data. The purpose of the descriptive statistics is to enable valuable analysis on the measures of central tendency, measures of variability, and frequency distribution of the data. There will also be a focus on finding interesting trends in the data that allow for valuable deductions in the data.

2.1. Features statistics:

The valid data being analysed only contains 2 categorical features (Class and Why Bought) and 9 continuous features (Primary Key, X, ID, Age, Price, Year, Month, Day, and Delivery Time.) The Class, Why Bought, Delivery Times, Year and Age features will now be looked at in more depth to understand their impact on the business. A special focus is put on the Price feature as this feature informs the business of sales revenue per product. These chosen features are valuable features to understand the trends of where, when and how the business makes the most sales revenue.

2.1.2 Class:

The 'Class' features plays an important role in the analysis of data because understanding which product types the clients are buying and what income is associated with each item allows the business to understand crucial information regarding which products work and which need to be looked at.

As can be seen in Table 3, the 'Gifts' class is the most frequent, meaning that 'Gifts' is the type of product that is most popular and bought most frequently (39148 times) by the client. It can also be deduced that between the top 3 most frequently bought product types (Gifts, Technology and Clothing) that they add up to 56.7% of all total frequency of sales.

Class	Mode Rank	Mode Frequency	Percentage	Mean Price	Std Price
Gifts	**1**	**39148**	**21.8%**	**2961.84**	**1611.12**
Technology	**2**	**36347**	**20.2%**	**29508.06**	**16368.40**
Clothing	**3**	**26403**	**14.7%**	**640.52**	**296.41**
Food	**4**	**24580**	**13.7%**	**407.77**	**163.06**
Sweets	**5**	**21565**	**12.0%**	**304.02**	**156.65**
Household	**6**	**20066**	**11.1%**	**11008.11**	**6263.84**
Luxury	**7**	**11869**	**6.6%**	**64857.12**	**30082.28**

Table 3: Class categories vs mode, mode frequency, mode percentage, mean price and standard deviation of price.

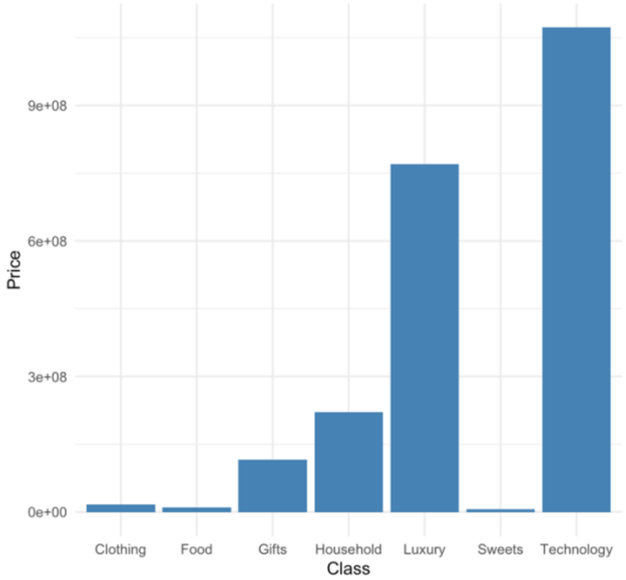


Figure 1: Bar plot of price per class

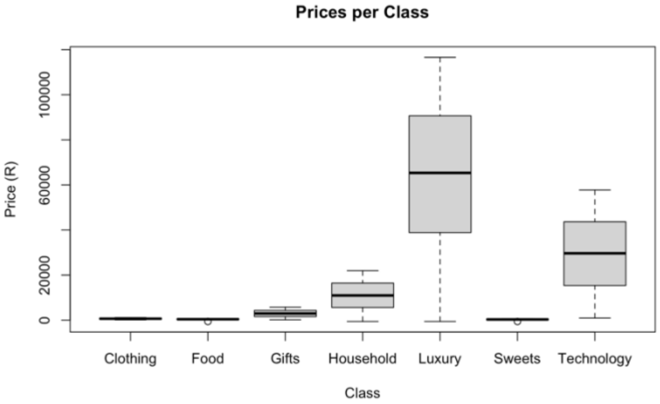


Figure 2: Box and whisker plot of price per class

The figures and tables above allow for crucial information to be deduced about the price data associated with each class. It can be seen in Figure 1 that the 'Technology' product brings in the most amount of revenue, this is not surprising as it seems to have a reasonably high frequency (2nd most frequent sale with 36347) along with the second highest mean price of R29508.06, with only 'Luxury' items having a higher mean price (R64857.12). It is also important to note that by looking at Figure 2, the spread of the prices can be seen per class and that there seems to be a reasonably large spread and variation of the prices for 'Luxury', 'Technology' and 'Household' products, relative to the large scale of the figure. This large spread means that there is a high degree of variability in the prices, but at the same time it can be deduced that even with that variability the different classes still hold their same rank in terms of price per product. It can also be deduced from Figure 2 that all the classes seem to have a normally distributed price per product.

Another important deduction form the above tables and figures (Table 3 and Figure 1 and 2) is that 'Gifts', 'Technology' and 'Clothing' seem to be the most frequently bought products but that 'Technology', 'Luxury' and 'Household' products seem to bring in the most sales revenue. 'Luxury' items bring in the second most revenue but is the least frequently bought and a similar deduction can be made for Household items.

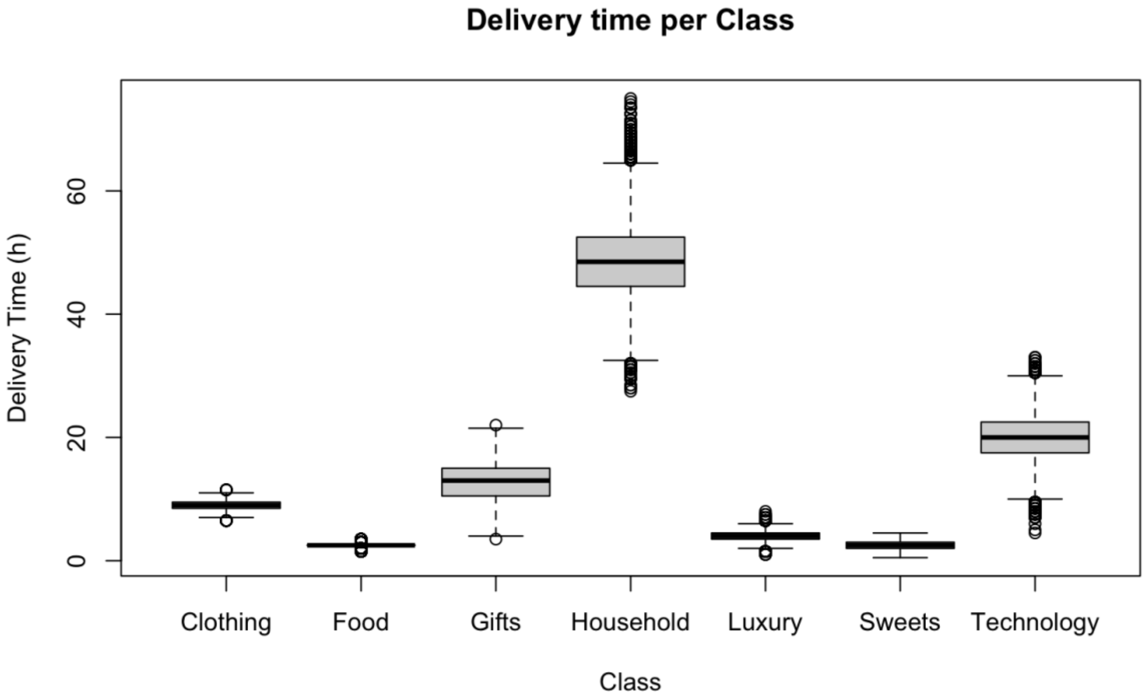


Figure 3: Box and whisker plot of delivery time per product class

In Figure 3 it can be noted that the delivery times for 'Food', 'Sweets' and 'Luxury' items all seem to be very low (below 5 hours.) This is likely due to the fact that 'Food' and 'Sweets' have a low shelf life and need to have a very quick turnaround time from order to delivery in order to gain maximum utilization from those products. 'Luxury' items likely have a quick delivery time because the price of the products are very expensive, thus it is likely a high priority for the business to deliver the products quickly to the high paying clients. From figure 3 it can also be deduced that 'Household' products have a slow delivery time (more than 40 hours.) This is likely due to the fact that these products include large furniture and products that require complex transport from warehouse to customer. These products are bound to have a long lifespan and thus the delivery time is short relative to its lifespan, but it is still something that could be looked at and possibly improved. Another thing to mention is that the delivery times in most class cases seem to be normally distributed and that 'Household', 'Gifts' and 'Technology' products have a relatively large spread of delivery times (for 'Houshold' it can be seen that delivery time ranges from around 30 hours to 70 hours.)

2.1.2 Why.Bought:

Another important feature to have a look at is the one relating to why a product is bought. This gives valuable information regarding how the business got to making the sale.

As can be seen in Table 4, the ‘Recommended’ category is the most frequent occurring reason for why a product is bought (1069867 times). This means that the business heavily relies on word of mouth to spread the news about the products (around 60% of clients buy the products due to it being recommended.) This shows that it is crucial for the business to maintain good quality in the products as well as keep improving customer relationships. The fact that the clients are being kept happy is evident and this needs to be maintained. It is also important to ensure that the business’ website is updated regularly as this category also contributes a large amount towards why clients bought a product (16%) as can be seen in Table 4 below.

Why.Bought	Mode Rank	Mode Frequency	Percentage	Mean Price	Std Price
Recommended	1	**106988**	**59.4%**	**13441**	**22762**
Website	2	**29447**	**16.4%**	**11021**	**16933**
Browsing	3	**18994**	**10.6%**	**16131**	**22486**
Random	4	**13121**	**7.3%**	**4288**	**9571**
EMail	5	**7225**	**4.0%**	**6661**	**13351**
Spam	6	**4208**	**2.3%**	**9361**	**15365**

Table 4: Why.Bought categories vs mode, mode frequency and mode percentage

Furthermore, when looking at Figure 4 it can be understood that majority of the sales revenue comes in from the ‘Recommended’ category, this makes sense seeing as it accounts for about 60% of the sales. In Figure 4 it can also be seen that ‘Browsing’ and ‘Website’ account for the second most sales revenue, this also relates to it being second and third most frequent reasons for sales. This shows that there is a relatively strong correlation between Price vs Why.Bought and Frequency vs Why.Bought for this feature, whilst for the Class feature this was not the case as a high frequency for a class does not mean a high sales revenue for that class (eg. Gifts is most frequently bought but 4th highest sales revenue.)

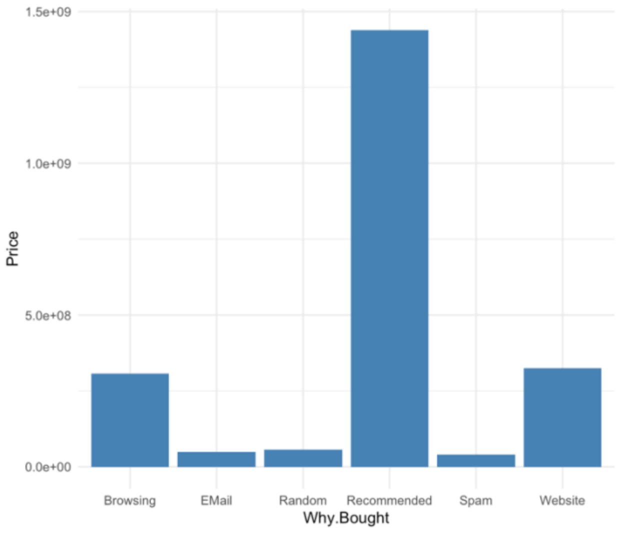


Figure 4: Bar graph of price per category of the product was bought

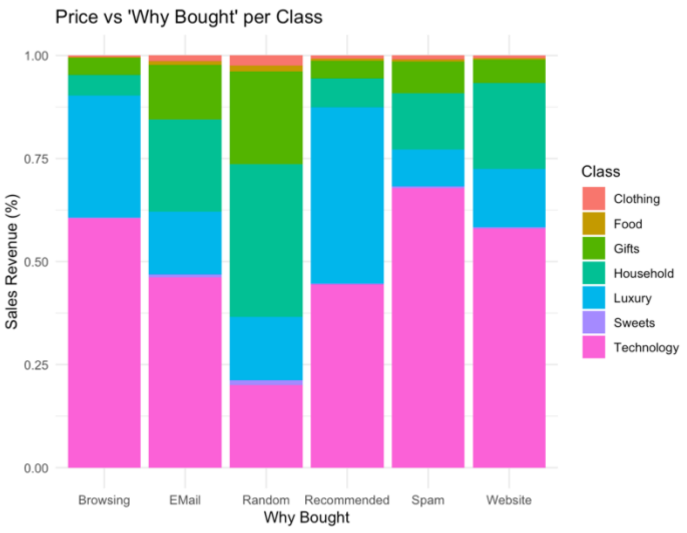


Figure 5: Stacked bar graph showing price distribution per class for each ‘Why.Bought’ reason

Another interesting deduction can be made when looking at Figure 5. When looking at specifically the ‘Recommended’ category it can be seen that ‘Technology’, ‘Luxury’ and ‘Household’ products make up about 90% of the revenue. When taking this into account with the understanding that the ‘Recommended’ category makes up 60% of the reason for the sales, it can be deduced that 54% (0.9x0.6) of the total revenue of the business is accounted for by ‘Recommended’ clients for ‘Technology’, ‘Luxury’ and ‘Household’ products. This is important to note as this combination of features plays a vital role in securing a large amount of revenue.

It is also important to note the distribution of the ‘Why.Bought’ categories. In Figure 6 it can be seen that for all categories the data is positively skewed. This means that majority of the sales prices are in the lower region of the box plot (less than R5000.) This can be confirmed by looking at Figure 7 where it can be seen that majority of the sales are made under R7500. It can also be seen in Figure 7 that there are still accounts of sales being sold up to over R100000, and this is the reason for the skewed data and many outliers in the data. This also gives an understanding as to why the standard deviation of the different ‘Why.bought’ categories are so high in Table 2.

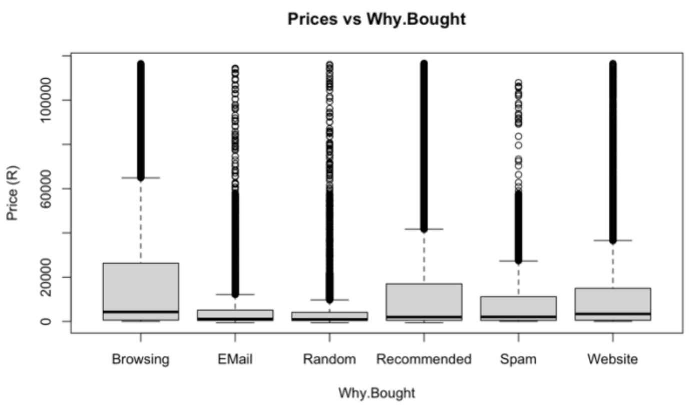


Figure 6: Box and whisker plot for price distributions per reason why the product is bought.

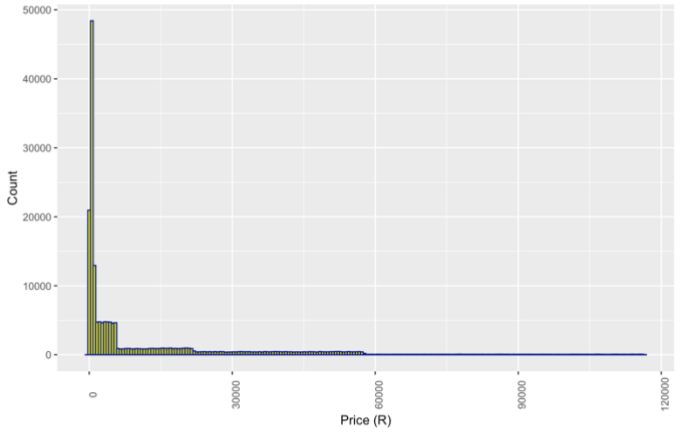


Figure 7: Histogram showing the distribution of prices of items purchased.

2.1.3 Price vs Year and Age:

From Figure 8 it can be seen that there is a spike in sales frequency and revenue in the first year (2021) and thereafter it subsides only to increase with an upward gradient periodically until the last year (2029.) Looking into this even further it can be noted in Figure 10 that in 2021 there is a large influx of Household items that were sold (40% of total sales compared to a normal aggregate of 10% of total sales in other years,) this is likely due to some sort of economic need for Household items in that year.

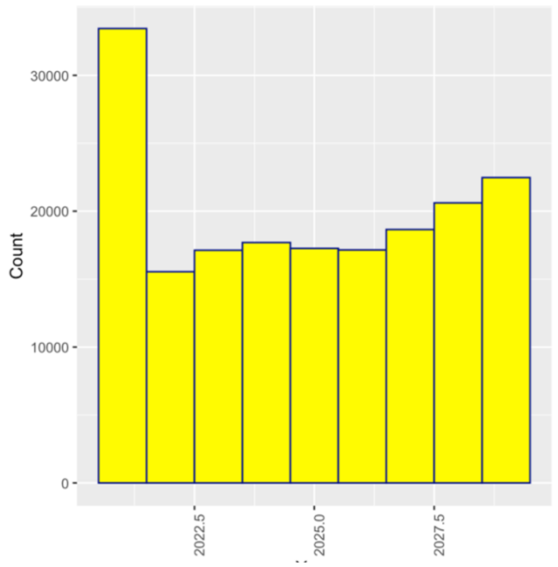


Figure 8: Histogram showing frequency count of items bought per year

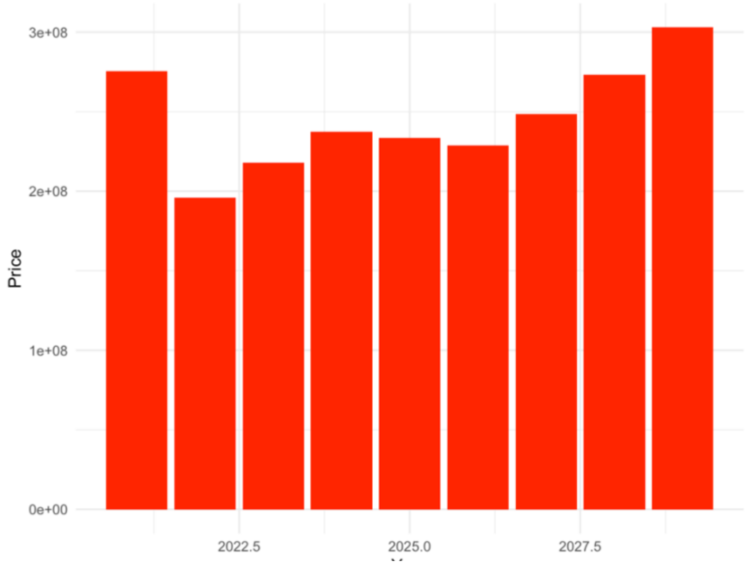


Figure 9: Bar Graph showing total revenue per year

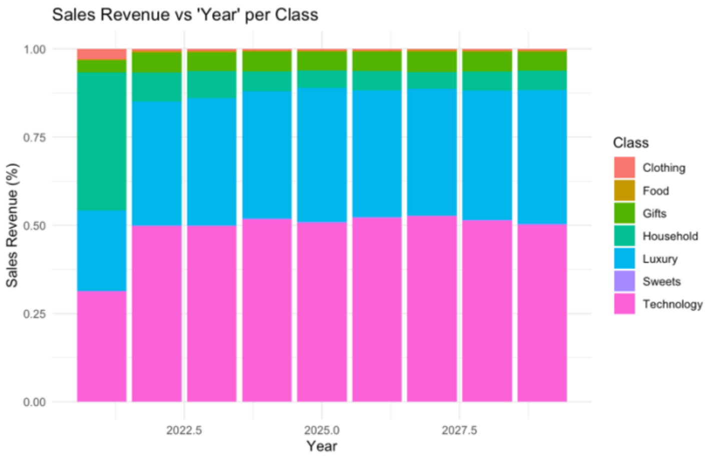


Figure 10: Stacked bar graph showing the percentage of products sold per year

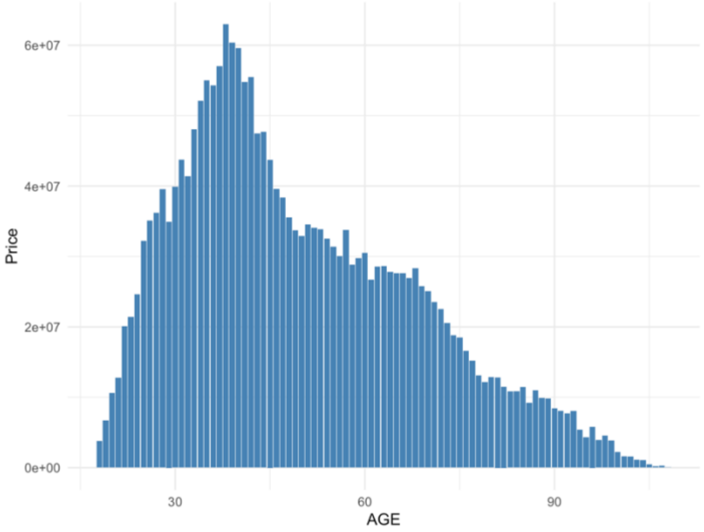


Figure 11: Distribution of the age vs total sales.

The final deduction that will be made from the data is that the business seems to have clients from age 18 to 108 with majority of the sales revenue coming from clients age 30 t0 45. The data follows a distribution that is skewed to the right, this also just reaffirms the deduction that majority of the sales revenue lie in the lower age region of the data.

2.2. Process capability

An USL of 24 and a LSL of 0 is used to determine the process capacity indices of the process delivery times for the 'Technology' class products.

When analysing the Technology class delivery times it is determined that the class has a mean delivery time of 20 hours, a standard deviation of 3.5 hours and a 5 point number summary which can be seen in Table 5 below.

Minimum	1st Quartile	Median	3rd Quartile	Maximum
4.5	17.5	20	22.5	33

Table 5: 5-point number summary of technology delivery times in hours.

CP	CPU	CPL	CPK
1.142207	0.3796933	1.90472	0.3796933

Table 6: Process capacity indices for technology delivery times

The process is not sufficiently capable seeing as the CP is greater than 1. Thus this process does not seem to be capable of delivering the technology products within the specified time when the instance falls in the region of deviation. This can be looked at and improved by making changes to the delivery process which make the standard deviation smaller for the technology class's delivery times.

The process is far from centred as it can be seen that the CPK (0.3796) is a reasonably far margin away from the CP value which is 1.142, because CPK is far from the CP value of 1.142. This can be improved by adjusting the process to fit the target better. Lastly a LSL of 0 does make sense as the delivery time cannot be less than 0 and due to that has a minimum value of 0.

3. Part 3: Statistical Process Control for X&s-charts

X and S charts play an important role in understanding the mean (X) and standard deviation (s) for a particular process over time. The standard deviation distribution of the data allows the business to see what the spread is of the data, thus identifying how large the variability of the data is from the mean that is determined from the samples selected.

3.1. Understanding control charts

When Figure 12 is considered, it can be noted that in a stable process 68.3% of the data instances should fall between ± 1 sigma, 95.5% of the data instances should fall between ± 2 sigma and 99.7% of the data instances should fall between the UCL (upper control limit) and LCL (lower control limit.) (Macros, Q.I. (2020))

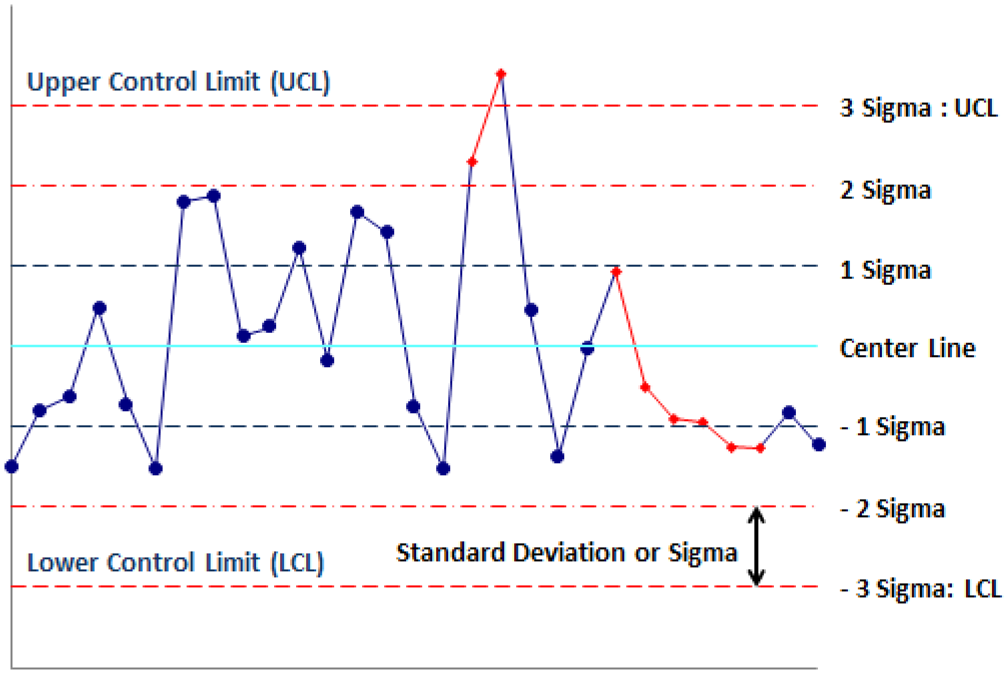


Figure 12: Explanation diagram the UCL, U2Sigma, U1Sigma, CL, L1Sigma, L2Sigma and LCL. (Macros, Q.I. (2020))

3.2. Initialising and Analysing control charts

In order to analyse and understand the following control charts it is important to note that the initial charts were created with 30 samples containing 15 instances each. The purpose of this is to initialise the data for an error-free period and then develop a control chart which can be used to analyse the rest of the samples in the dataset in order to remove all the samples which fall outside of the control limits determined during the initialisation of the charts. The grey region on the control charts figures indicate the accepted region between the UCL and LCL and the centre line is indicated by a bold grey line.

Below in Table 7 and 8 are the control limits and sigma values for the X-bar charts and S-Charts for the first 30 samples. In Figures 13 to 26 are the control charts for the first 30 samples with explanations.

X-Bar-Chart Control Limits:

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22.9731	22.1069	21.2407	20.3744	19.5082	18.6420	17.7758
Clothing	9.4047	9.2598	9.1149	8.9700	8.8251	8.6802	8.5353
Household	50.2462	49.0182	47.7902	46.5622	45.3342	44.1062	42.8783
Luxury	5.4935	5.2409	4.9882	4.7356	4.4829	4.2302	3.9776
Food	2.7093	2.6362	2.5631	2.4900	2.4169	2.3438	2.2707
Gifts	9.4879	9.1123	8.7367	8.3611	7.9855	7.6099	7.2343
Sweets	2.8968	2.7571	2.6175	2.4778	2.3381	2.1984	2.0588

Table 7: X-bar chart table showing the mean UCL, U2Sigma, U1Sigma, CL, L1Sigma, L2Sigma and LCL of delivery times per class.

S-Chart Control Limits:

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.1799	4.5518	3.9237	3.2955	2.6674	2.0393	1.4111
Clothing	0.8664	0.7614	0.6563	0.5512	0.4462	0.3411	0.2360
Household	7.3432	6.4528	5.5623	4.6719	3.7814	2.8910	2.0005
Luxury	1.5109	1.3276	1.1444	0.9612	0.7780	0.5948	0.4116
Food	0.4372	0.3842	0.3312	0.2781	0.2251	0.1721	0.1191
Gifts	2.2460	1.9737	1.7013	1.4290	1.1566	0.8842	0.6119
Sweets	0.8352	0.7340	0.6327	0.5314	0.4301	0.3288	0.2275

Table 8: S-chart table showing the standard deviation UCL, U2Sigma, U1Sigma, CL, L1Sigma, L2Sigma and LCL of delivery times per class.

Technology initialisation:

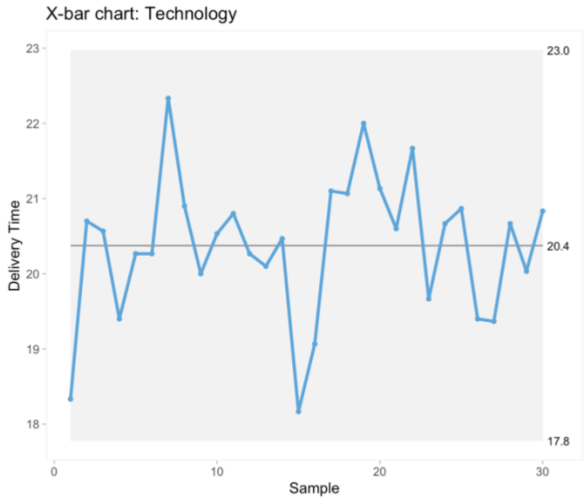


Figure 13: X-Bar control chart showing mean delivery times for technology for 30 samples

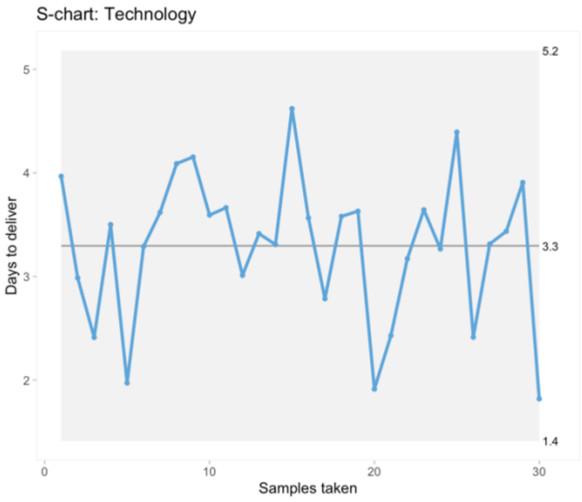


Figure 14: S control chart showing standard deviation of delivery times for technology.

In Figure 13 and Figure 14 it can be noted that all the samples fall within the control limits and thus the mean and standard deviation samples are in control for technology. The mean remains within the limits of 23 and 17.8 hours and the standard deviation remains within 5.2 and 1.4 hours.

Clothing initialisation:

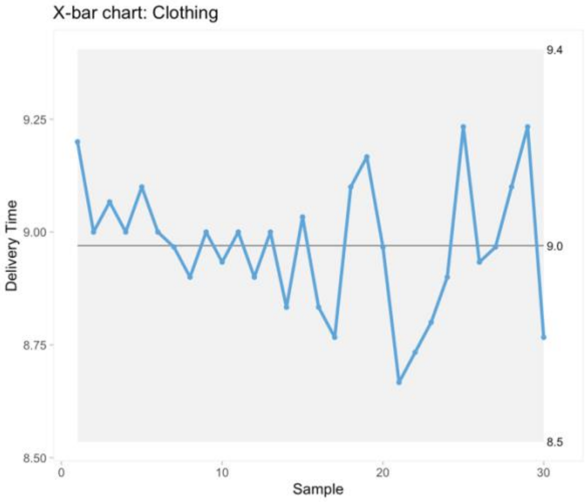


Figure 15: X-Bar control chart showing mean delivery times for clothing for 30 samples

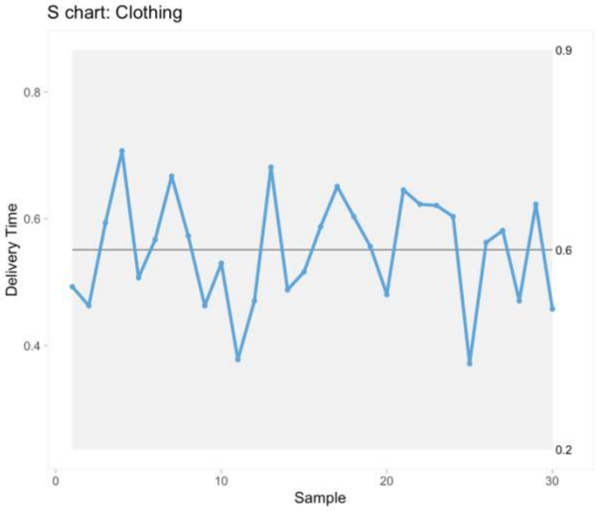


Figure 16: S control chart showing standard deviation of delivery times for clothing.

In Figure 15 and Figure 16 it can be noted that all the samples fall within the control limits and thus the mean and standard deviation samples are in control for clothing. The mean remains within the limits of 9.4 and 8.5 hours and the standard deviation remains within 0.9 and 0.2 hours.

Housing initialisation:

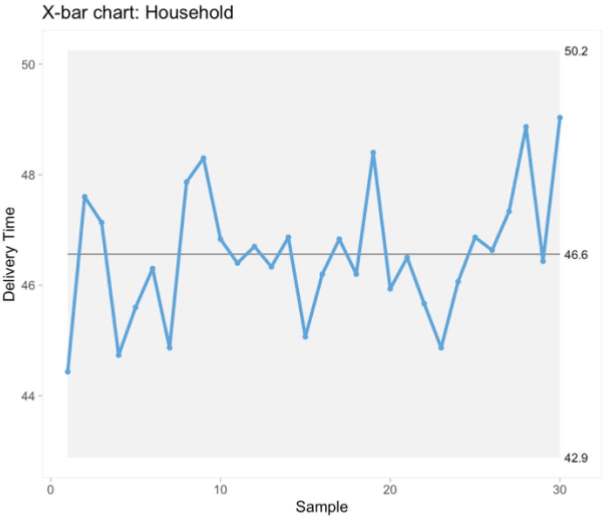


Figure 17: X-Bar control chart showing mean delivery times for housing items for 30 samples

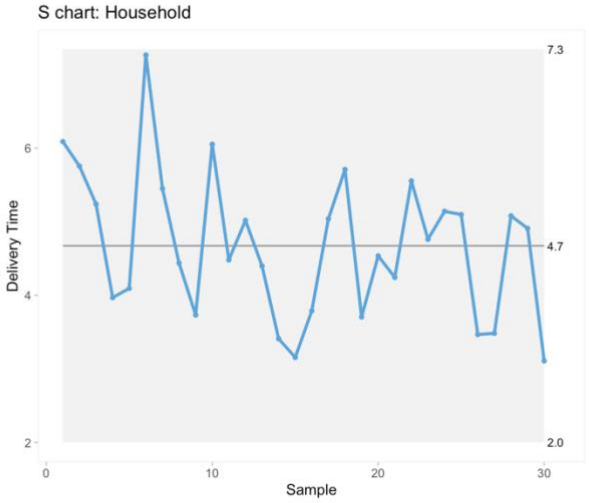


Figure 18: S control chart showing standard deviation of delivery times for housing items.

In Figure 17 and Figure 18 it can be noted that all the samples fall within the control limits and thus the mean and standard deviation samples are in control for housing items. The mean remains within the limits of 50.2 and 42.9 hours and the standard deviation remains within 7.3 and 2 hours.

Luxury initialisation:

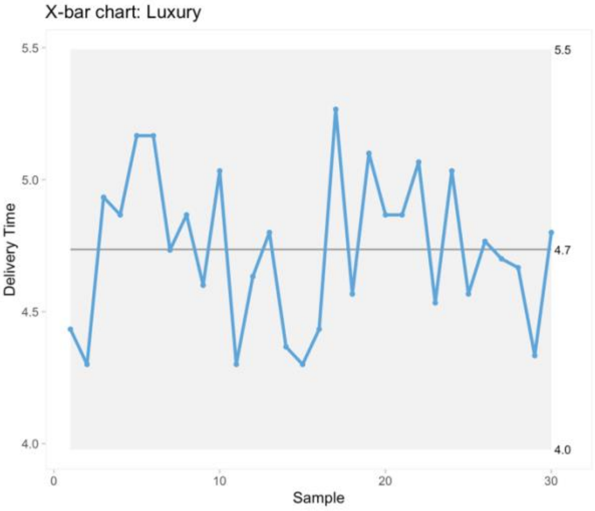


Figure 19: X-Bar control chart showing mean delivery times for luxury items for 30 samples

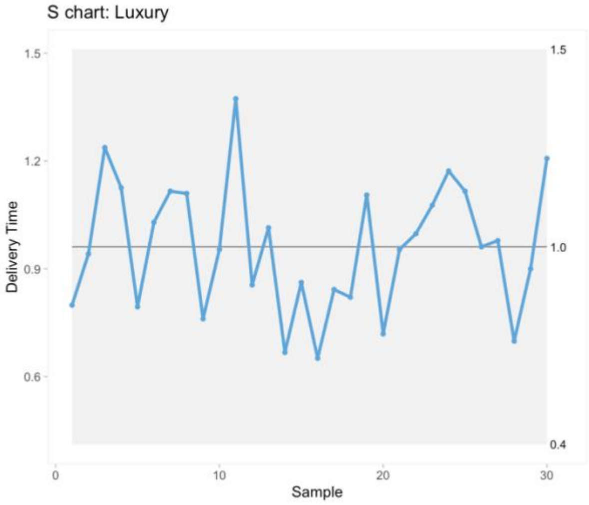


Figure 20: S control chart showing standard deviation of delivery times for luxury items.

In Figure 19 and Figure 20 it can be noted that all the samples fall within the control limits and is thus the mean and standard deviation samples are in control for luxury items. The mean remains within the limits of 5.5 and 4 hours and the standard deviation remains within 1.5 and 0.4 hours.

Food initialisation:

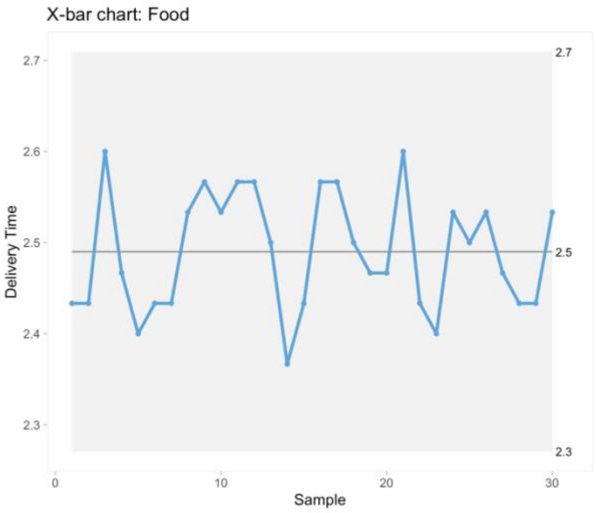


Figure 21: X-Bar control chart showing mean delivery times for food items for 30 samples

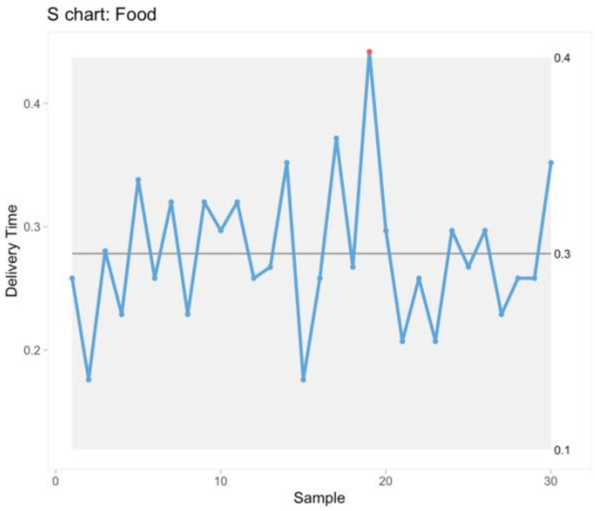


Figure 22: S control chart showing standard deviation of delivery times for food items.

In Figure 21 it can be noted that all the mean samples fall within the control limits and is thus the mean is in control for food items. In Figure 22 there is a sample which falls outside of the control limit for the standard deviation control chart (sample 19) and should be removed.

Gifts initialisation:

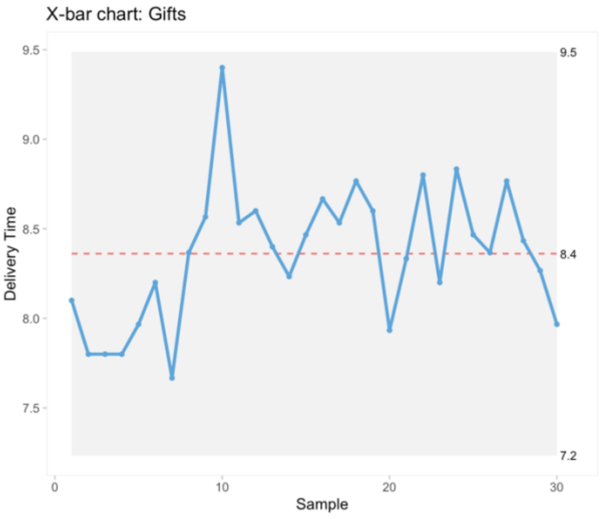


Figure 23: X-Bar control chart showing mean delivery times for gift items for 30 samples

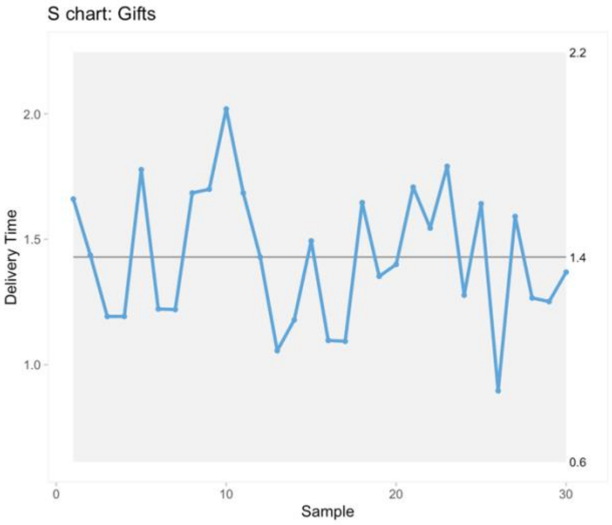


Figure 24: S control chart showing standard deviation of delivery times for gift items.

In Figure 23 and Figure 24 it can be noted that all the samples fall within the control limits and is thus the mean and standard deviation samples are in control for gift items. The mean remains within the limits of 9.5 and 7.2 hours and the standard deviation remains within 2.2 and 0.6 hours.

Sweets initialisation:

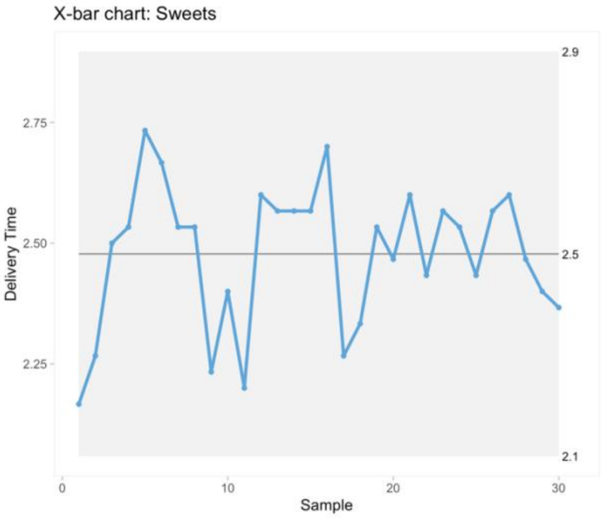


Figure 25: X-Bar control chart showing mean delivery times for sweet items for 30 samples

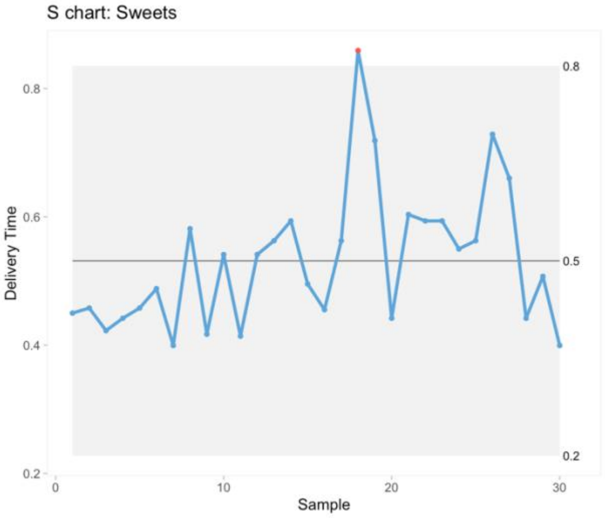


Figure 26: S control chart showing standard deviation of delivery times for sweet items.

In Figure 25 it can be noted that all the mean samples fall within the control limits and is thus the mean is in control for sweet items. In Figure 26 it can be noted that sample 18 falls outside of the upper control limit of the standard deviation graph and should be removed.

3.3. Statistical Process control for all samples

In this section it is important to use the control limits determined from the previous section for the first 30 samples to control the xBar and S charts. If samples do fall outside of these initial control limits they should be removed. The control charts indicating the mean and standard deviation of the delivery time for all the samples following on from and including the first 30 samples is now plotted for each class:

3.3.1 Technology:

From Figure 27 and 28 below it can be noted that there are around 2423 samples which need to be controlled. Take special note that the upper and lower control limits are generated from the first 30 samples only. These figures also show that for the X-bar chart 17 samples must be removed and for the S chart 16 samples must be removed, as they lie outside of the control limits. These figures also give us an indication that comparing to the mean values, the variation per sample does not deviate too much and is there for a reliable figure to understand the aggregate sales.

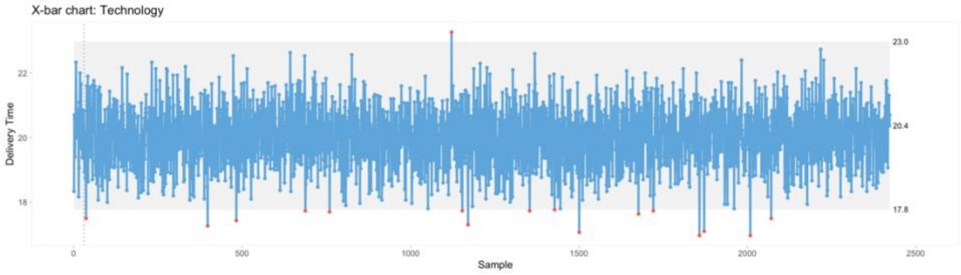


Figure 27: X-Bar control chart showing delivery time mean for all technology samples

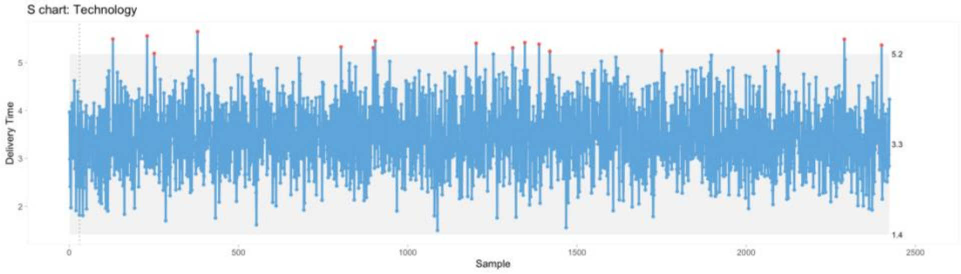


Figure 28: S control chart showing delivery time standard deviation for all technology samples

3.3.2 Clothing:

In Figure 29 and 30 below it can be noted that there are around 1760 samples which need to be controlled. Take special note that the upper and lower control limits are generated from the first 30 samples only. These figures also show that for the X-bar chart 17 samples must be removed and for the S chart 98 samples must be removed, as they lie outside of the control limits. In Figure 30 it can be noted that there is an increase in samples that fall outside of the limit, meaning that the reliability of deliveries became lower in the later samples. These samples which fall outside of the limits should also be investigated by going to the data and determining what caused their deviation.

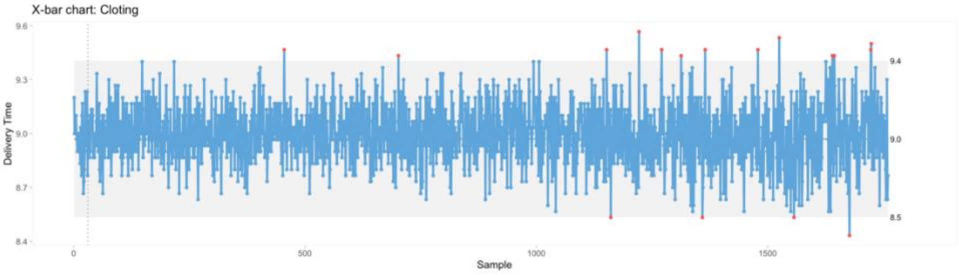


Figure 29: X-Bar control chart showing delivery time mean for all clothing samples

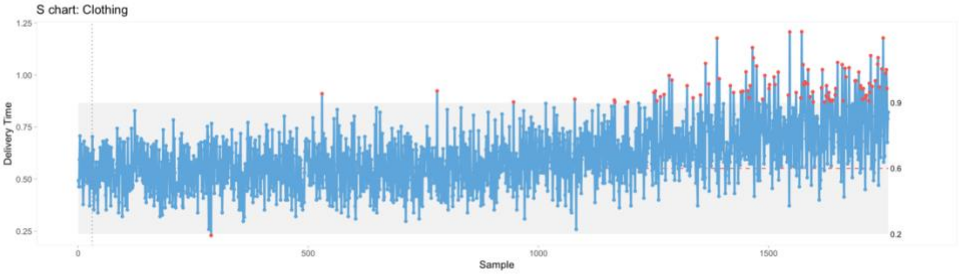


Figure 30: S control chart showing delivery time standard deviation for all clothing samples

3.3.3 Household:

In Figures 31 and 31 it can be well noted that the samples on the second half of the control chart vary heavily from those in the first half. The later samples for the X-bar chart have much larger average delivery time. This indicated that towards the later stages of the data the delivery times take much longer than the start. There are 400 samples which fall outside of the control limits for the X-bar chart. These samples which fall outside of the limits should also be investigated by going to the data and determining what caused their deviation.

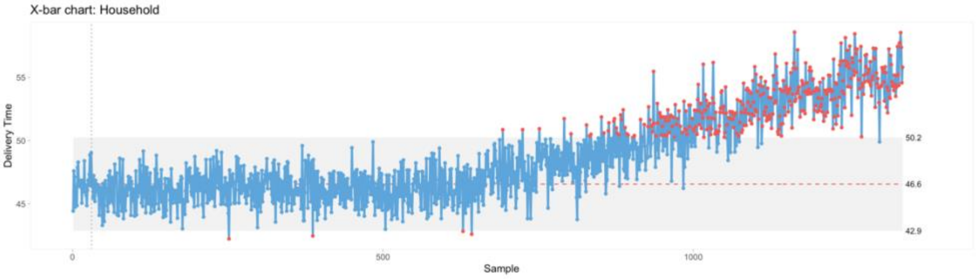


Figure 31: X-Bar control chart showing delivery time mean for all household samples

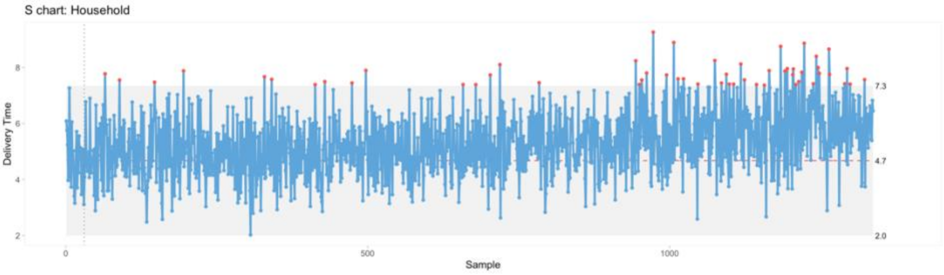


Figure 32: S control chart showing delivery time standard deviation for all household samples

3.3.4 Luxury:

In Figure 33 it can be noted that many sample means vary from that of the first 30 samples control limits. There are 434 samples that should be removed, which is a lot considering there are only 791 samples in the Luxury sample set. In Figure 34 it can be noted that there are very few samples out of control (4 samples.) There is clearly a large decrease in mean delivery times for luxury items and this is not a bad thing, but new control limits should probably be investigated.

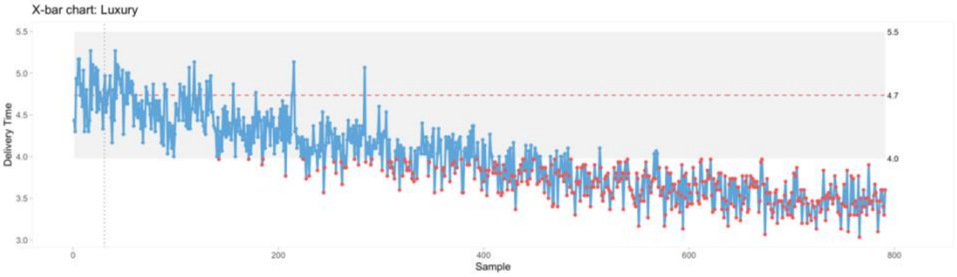


Figure 33: X-Bar control chart showing delivery time mean for all luxury samples

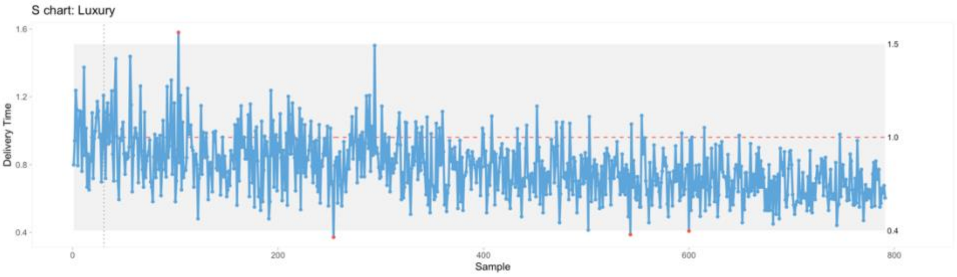


Figure 34: S control chart showing delivery time standard deviation for all luxury samples

3.3.5 Food:

The food samples consist of 1638 samples and in both Figures 35 and 36 it can be noted that the samples who’s mean and standard deviation values fall outside the control limits is less 4 and 5 respectively. This indicates a very reliable food delivery time, which is important as good service reliability is crucial for food deliveries due to its low shelf life.

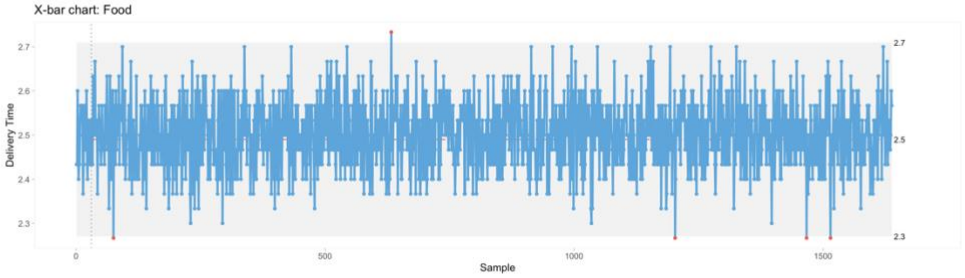


Figure 35: X-Bar control chart showing delivery time mean for all food samples

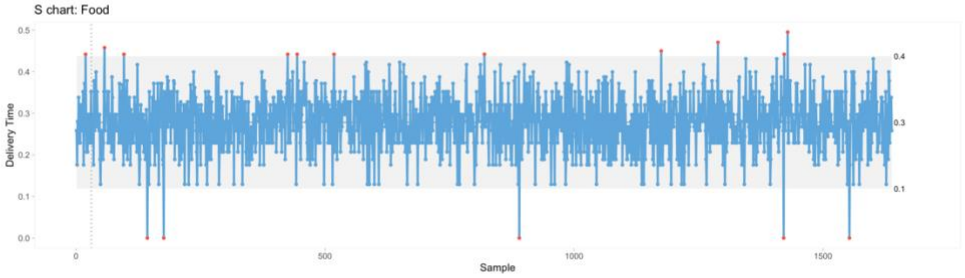


Figure 36: S control chart showing delivery time standard deviation for all food samples

3.3.6 Gifts:

In figure 37 it can be noted that the 2290 samples out of the 2609 lie outside of the control limits. It can also be noted that there are 9 different incremental groupings of samples, this likely indicates the change in delivery time per year from 2021-2029. This indicates a low reliability of delivery time as it is constantly changing. Looking at the samples, it might be better to choose a different of samples to calculate the control limits, but even doing so the same issue would occur. Thus a large issue is occurring and causing the mean delivery time to increase for the gift samples and this should be investigated.

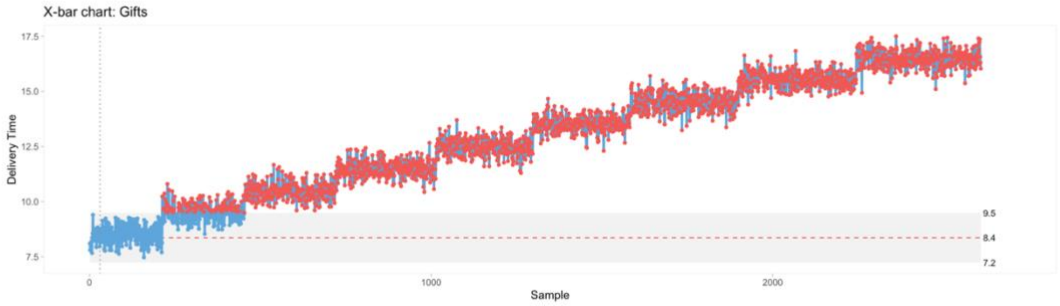


Figure 37: X-Bar control chart showing delivery time mean for all gift samples

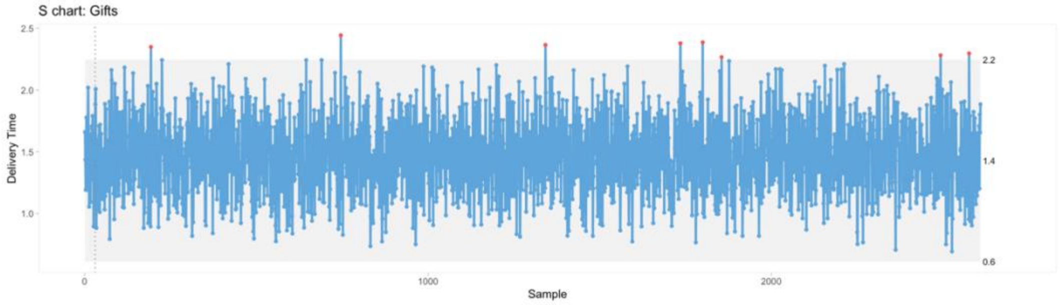


Figure 38: S control chart showing delivery time standard deviation for all food samples

3.3.7 Sweets:

The sweets samples consist of 1437 samples and in both Figures 39 and 40 it can be noted that the samples who's mean and standard deviation values fall outside the control limits is 5 and 1 respectively. This indicates a very reliable sweet delivery time with few variation in delivery time over the entire sample set.

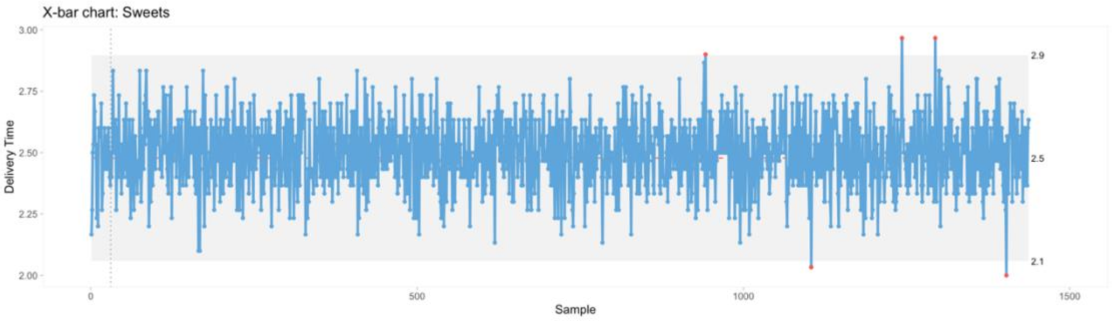


Figure 39: X-Bar control chart showing delivery time mean for all sweets samples

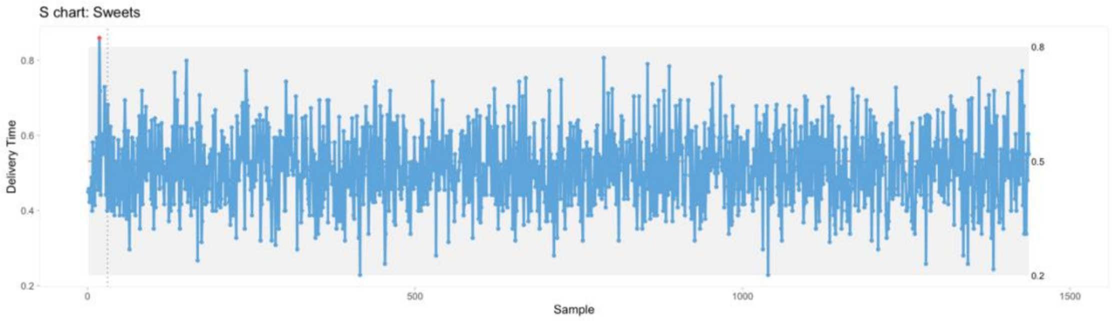


Figure 40: S control chart showing delivery time standard deviation for all sweets samples

4. Part 4: Optimising the delivery process

In this section there is an intention to analyse information in order to optimise the delivery process. All data samples are used in the analysis of information.

4.1. Samples out of control

For all classes there are some samples that fall outside of the control limits and thus need to be looked at.

A) Sample means which lie outside of the control limits:

In Table 9 below are the samples from each of the classes that fall outside of the control limits. For 'Technology' the first 3 samples that fall outside of the control limits is sample 37, 398 and 483. For 'Technology' the last 3 samples that fall outside of the control limits is sample 1872, 2009 and 2071 with a total samples of 17 falling outside of the control limits for this class.

Furthermore, an interesting value to mention is the total samples of 'Gifts' that fall outside of the control limits (2290 samples.) This is obviously a major issue and needs to be assessed to determine why this is the case, or another initial sample should be used that more accurately represents all the samples. The relatively high values for Household and Luxury samples also indicate that these class delivery times are not under control.

Class	1st	2nd	3rd	3rd last	2nd last	last	Total samples
Technology	37	398	483	1872	2009	2071	17
Clothing	455	702	1152	1677	1723	1724	17
Household	252	387	629	1335	1336	1337	400
Luxury	142	171	184	789	790	791	434
Food	75	633	1203	NA	1467	1515	5
Gifts	213	216	218	2607	2608	2609	2290
Sweets	942	1104	1243	NA	1294	1403	5

Table 9: Sample means per class that are outside of the control limits

B) Most consecutive sample standard deviations per class between -0.3 and 4 sigma:

For this section it can be noted in Table 10 that most the class of ‘Food’ contains the most consecutive samples which fall within -0.3 and 0.4 sigma. The range of samples in which this falls is between sample 92 and 97 for the technology samples.

Class	Most Consecutive	Last Sample
Technology	6	2410
Clothing	4	1750
Household	3	1320
Luxury	4	764
Food	7	1636
Gifts	5	2604
Sweets	4	1437

Table 10: Most consecutive samples per class and last sample

4.2. Type I (Manufacturer's) Error for A and B

Manufacturer’s errors occur if there is nothing wrong with a process, but it is thought that there is something wrong with the process and thus the process stops to be inspected. This happens when instances which are incorrectly defined as being outside the control limits when they are actually inside the control limits, or when the control limits are faulty.

When doing calculations for 4.1.A to determine the p-value -3 was used. The upper and lower control limits represents the maximum and minimum z-values (-3 and 3). When calculating 4.1.B. the -0.3 and 0.4 sigma values were used as this is the region in which the values had to fall for B.

A:

P(Type I Error for A) = pnorm(-3)*2

= 0.002699796

= 0.270 %

Thus indicating that there is a 0.27% probability that a sample mean falls within the control limits and it is thought to be incorrect.

B:

P(Type I Error for B) = pnorm(0.4) – pnorm(-0.3)

= 0.2733332

= 27.33332%

Thus indicating that there is a 27.33% probability that a sample standard deviation that falls within the -0.3 and 0.4 sigma limits is correct and it is thought to be incorrect.

4.3. Optimizing the technology delivery cost

The business is interested in optimising the technology delivery cost. There is a penalty cost of R329 per item-late-hour which is incurred due to a loss in sales if the item delivery takes longer than 26 hours. There is also a cost of R2.5 per item per hour to reduce the average cost by 1 hour. The current mean delivery time for technology is 20.01095 hours.

After iterating through the number of cost loss function which can be calculated as follow:

Cost Loss(x) = Number of late days(x)*329 + 2.5*x*(Number of instances)

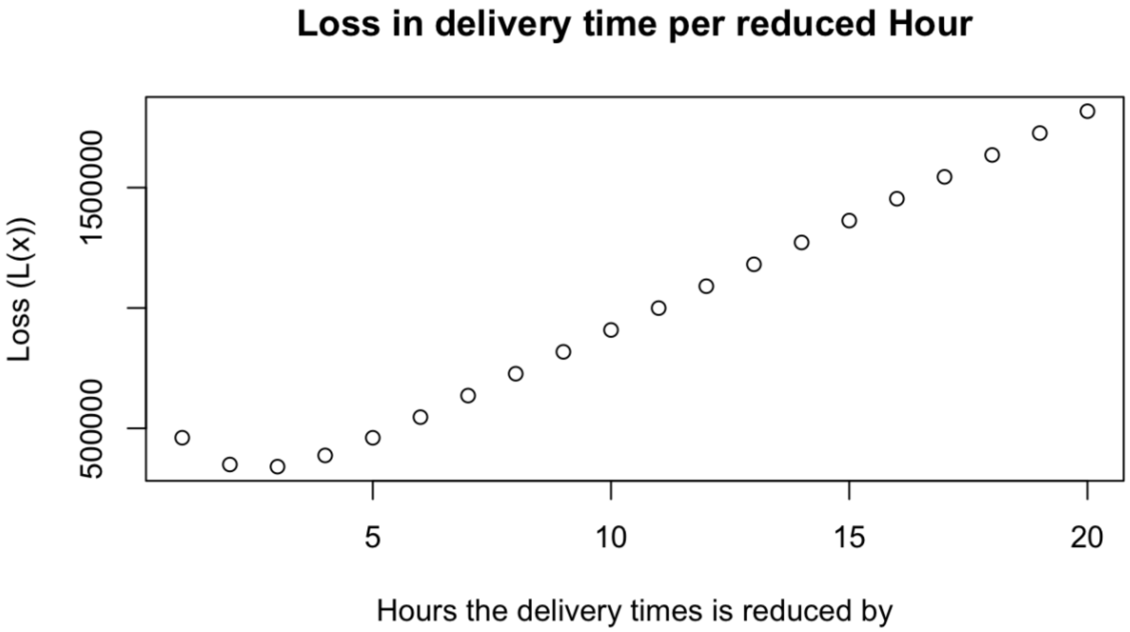


Figure 41: Cost loss function per hours that the delivery time is reduced

After evaluating the iterations it can be determined that the minimum loss cost will be R 340870 and this occurs when the delivery time is reduced by 3 hours (as can be seen in Figure 41.) Thus indicating that the technology delivery process will be at its most optimal when the average delivery time is reduced by 3 hours and equal to 17.01095.

A Taguchi loss function is used to determine the loss that occurs from a product being outside of its specification (Dirkse van Schalkwyk, T. (2022).) In the case of the cost loss function the intention was to find the optimisation point at which the costs for the delivery time is the lowest. This is similar to the Taguchi loss function because the mean values are used to determine the cost loss. Both determine the cost loss but do so in different ways.

4.4. Type II error

A consumers error (Type II error) refers to a false negative. For a type II error, the Ha is true, but it is not identified and this is because the sample Xbar value being between the LCL and UCL. In this case the control limits are known but the mean is to be shifted to 23

The probability that the mean value will lie within the UCL and LCL is calculated:

P(Type II) = pnorm(UCL, mean,standard deviation) - pnorm(LCL, mean,standard deviation)

=pnorm(22.2973,23, 0.8662197) - pnorm(17.77579,23, 0.8662197)

= 0.4876147

Thus indicating that there is a **0.4876** chance that if a delivery process mean shifts to 23 and a sample falls within the control limits, it will not be realised that the mean has shifted.

5. Part 5: DOE and MANOVA

Throughout the report a big emphasis has been put on the ‘Class’ and ‘Why.Bought’ features, with major ways of comparing these two features being the sales revenue (‘Price’ feature) and the ‘Delivery Time’ feature. These features allow the business to determine which class of items bring in the most revenue as well as what the main reason was for bringing in customers and generating the most revenue. Thus a MANOVA will be set up for ‘Why.Bought’ and ‘Class’ features using ‘Price’ and ‘Delivery Time’ features and the ‘Delivery Time’ feature will especially give the business valuable information about service and reliability relating to the products and reasons for why it is bought.

5.1. Class MANOVA

When performing the Hypothesis test for the ‘Class’ feature it is important to determine the null hypothesis (H_0) and alternative hypothesis (H_1) in order to determine what the business is trying to predict. When looking at the ‘Class’ feature, the business would like to determine whether ‘Price’ and ‘Delivery Time’ influence the the ‘Class’ feature. Thus:

- H_0: Price and Delivery time **does not** influence the ‘Class’
- H_1: Price and Delivery time **does** influence the ‘Class’

After performing the MANOVA for the ‘Class’ feature it can be noted in Table 11 and Table 12 that from both ‘Price’ and ‘Delivery time’ that the P value in both cases is less than $2.2\text{x}10^{-16}$. This is extremely small and indicates that the null hypothesis (H_0) is rejected and the alternative hypothesis (H_1) is accepted. Thus indicating that the ‘Class’ feature is influenced by the Price and Delivery time features. This is further confirmed when looking at Figure 1 and seeing that if a class of ‘Technology’, ‘Luxury’ or ‘Household’ items are purchased they have a much larger sales revenue than the rest of the classes (confirming that price plays a role in categorising the different classes.)

Furthermore looking at Figure 3 it can be noted that Household items have a far longer distribution of delivery time than other items and thus it is understood why the ‘Class’ feature does matter when relating to Price and Delivery time. This figures also brings to light that the delivery service reliability is less for ‘Technology’, ‘Gifts’ and ‘Household’ items as they have a larger spread of data and thus it can be looked at to reduce the variation of these delivery times.

Another analysis that can be made from Figure 3 is that for ‘Clothing’, ‘Food’, ‘Luxury’ and Sweets items the distribution of delivery times are very small, thus indicating a high service reliability for these items.

Response Price:

	DF	Sum sq	Mean sq	F value	Pr(>F)
Class	6	5.7168e+13	9.5281e+12	80258	< 2.2e-16
Residual	179971	2.1366e+13	1.1872e+08		

Table 11: MANOVA for sales revenue relating to 'Class' feature

Response Delivery time:

	DF	Sum sq	Mean sq	F value	Pr(>F)
Class	6	33458565	5576427	629429	< 2.2e-16
Residual	179971	1594452	9		

Table 12 : MANOVA for delivery time relating to 'Class' feature

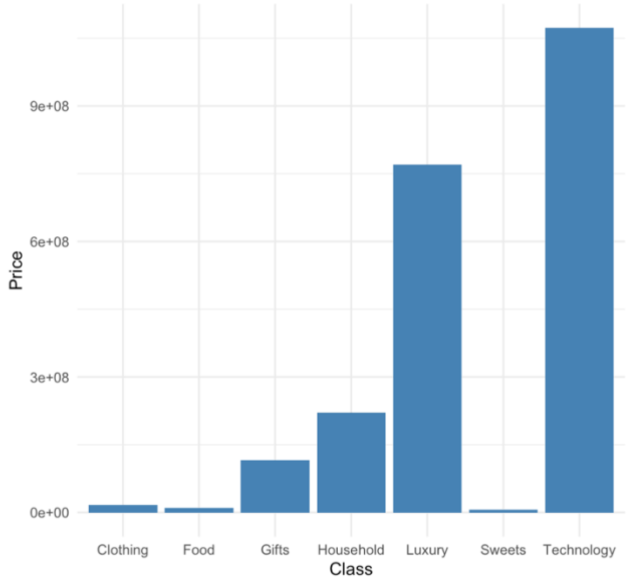


Figure 1: Bar plot of price per class

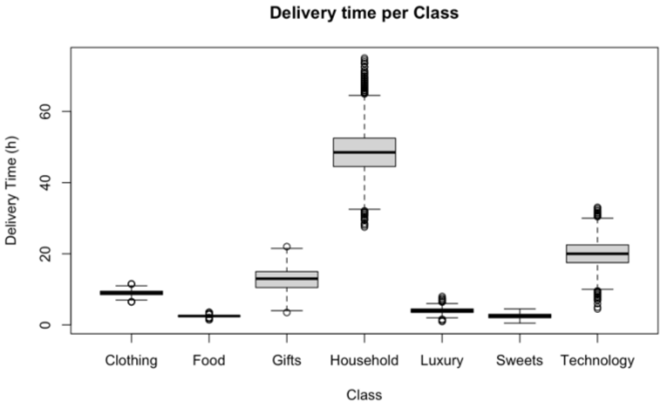


Figure 3: Box and whisker plot

5.2. Why.Bought MANOVA

When looking at the ‘Why.Boughts’ feature, the business would like to determine whether the ‘Why.Bought’ feature is influenced by the ‘Price’ and ‘Delivery time’ features, thus:

- H_0: Price and Delivery time **does not** influence the ‘Why.Bought’ feature.
- H_1: rice and Delivery time **does** influence the ‘Why.Bought’ feature.

When performing the MANOVA for the ‘Why.Bought’ feature it can be noted in Table 13 and Table 14 that from both ‘Price’ and ‘Delivery time’ that the P value in both cases is $2.2\text{x}10^{-16}$. This is extremely small and due to this small P value, the null hypothesis (H_0) is rejected and the alternative hypothesis (H_1) is accepted. Thus indicating that the ‘Why.Bought’ is influenced by the ‘Price’ and ‘Delivery time’ of an instance. This is further confirmed when looking at Figure 4 seeing that if a reason for a product being bought is ‘Recommended’, it makes a much larger sales revenue than the other reasons for why a product was bought.

Price:

	DF	Sum sq	Mean sq	value	Pr(>F)
Class	5	1.5742e+12	3.1484e+11	736.26	< 2.2e-16
Residual	179972	7.6960e+13	4.2762e+08		

Table 13: MANOVA for sales revenue relating to ‘Why.Bought’ feature

Delivery time:

	DF	Sum sq	Mean sq	value	Pr(>F)
Class	5	783320	156664	822.74	< 2.2e-16
Residual	179972	34269697	190		

Table 14: MANOVA for delivery time relating to ‘Why.Bought’ feature

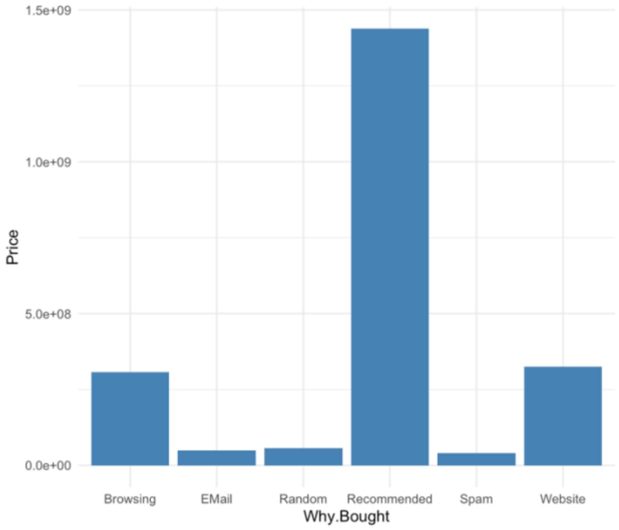


Figure 4: Bar graph of price per category of the product was bought

6. Part 6: Reliability of the service and the products

It is crucial for food deliveries to be kept cool during transit. The business has a subsidiary, Lafrideradora, who makes components for their units. The business would like to compare different loss functions to determine the best one that will misimise the loss of the business.

6.1. Tuguchi Loss Function

Form the given information the Taguchi loss function can be determined:

Problem 6

Given: Loss= R45, part thickness = 0.06+-0.04cm.

$$\mathbf{L(x) = K(y\text{-}m)^2} \quad (1)$$

$$\mathbf{45 = K^*(0.04)^2}$$

$$\mathbf{K = 28125}$$

$$\mathbf{L(x) = 28125^*(y\text{-}0{,}04)^2}$$

Problem 7

Given: a) Loss= R35, part thickness = 0.06+-0.04cm. b) (y-m)=0.027

a)

$$\mathbf{L(x) = K(y\text{-}m)^2}$$

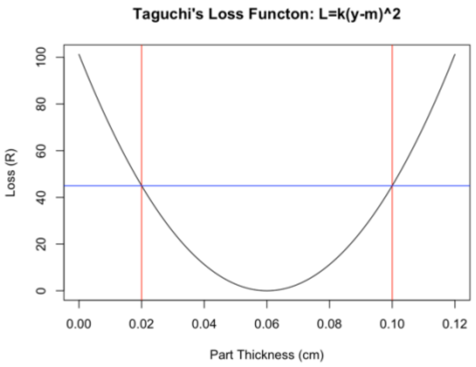
$$\mathbf{35 = K^*(0.04)^2}$$

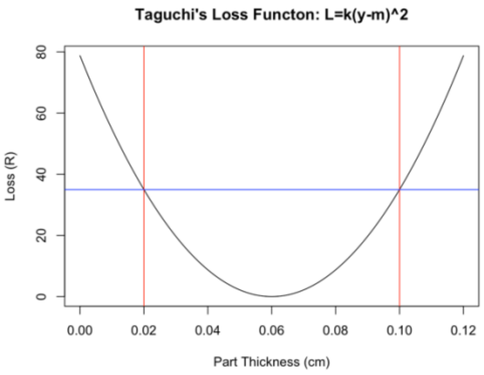
$$\mathbf{K = 21875}$$

$$\mathbf{L(x) = 21875^*(y\text{-}0{,}04)^2}$$

b)

$$\mathbf{L(x) = 21875^*(0.027)^2 = R15.95\ per\ part}$$





When looking at Figure 42 and 43 it can be noted that there is a certain region in which scrap loss does not occur, but as soon as the tolerance deviation of a part is passed then a loss starts to originate. It can also be noted from (1) that if the deviation of a part is decreased, it will have a significant effect on the increasing of the loss cost and therefore another way should be found to reduce the loss cost per part.

Figure 42: Taguchi Loss function with scrap loss cost being R45

Figure 43: Taguchi Loss function with scrap loss cost being R35

that in both c : be scraped is when the part has a thickness of 0.02-0.1cm and therefor staying within this rage significantly increases the reliability of the part not being scrapped, which will keep the cost lower. This therefore has a large impact on the service reliability of the food deliveries as a working, cost effective refrigerator system is vital for ensuring fresh deliveries of food items.

6.2. System reliability

Problem 27:

Given information: P(A) = 0.85, P(B)= 0.92, P(C)= 0.9.

a) If no backup machines are working

$$\textbf{Reliability Probability} = P(A)*P(B)*P(C)$$

$$= 0.85*0.92*0.9$$

$$=0.7038$$

Thus the reliability probability of the manufacturing of the technology items **without** a backup machine is 0.7038.

b)

$$\textbf{Reliability} = [1-(1-P(A)) \text{ x } (1-P(A))] \text{x } [1-(1-P(B))*(1-P(B))]*[1-(1-P(A))*(1-P(A))]$$

$$=0.9775 \text{x} 0.9936 \text{x} 0.99$$

$$= 0.9615$$

Thus, the reliability probability of the manufacturing of the technology items **with** a backup machine is 0.9615.

Therefore using the system with a backup machine increases the reliability of the manufacturing process by 25.8%. Due to the high frequency of sales per period of the technology products, and the need to maintain a high service level for the technology class with our customers as it is one of the business's main sources of income, it will be wise to go for the option with backup machines. This will aid the business in terms of not falling behind on production. An alternative option for this would be to only have a backup machine for machine A (which has the lowest reliability) and this increases the reliability to 0.9775x0.92x0.9 = 80.94%. The final recommendation is however for Magnaplex to keep the backup machines running for the technology products to maintain a high reliability in the production line.

6.3. Days to expect reliable delivery times

The business has 21 vehicles and 21 drivers available per day to do deliveries. The requirement is for at least 20 vehicles to be driving to reach service reliability. With the given information the days per year which the business can expect to do reliable deliveries is calculated (using the binomial function in r.):

The initial probability values are not known and therefore an iterative process was used to determine the roots for initial probability values to give an accurate answer on the binomial probability of the events occurring.

P(reliable delivery vehicles per day) = 0.8615412

P(reliable delivery drivers per day)= 0.9344269

Thus the number of days in a year in which to expect a reliable delivery is:

P(reliable delivery vehicles per day)xP(reliable delivery drivers per day)x365

0.8615412x0.9344269x365

= 293.8422 = 293 days

If the number of delivery vehicles is increased to 22 then the P(reliable delivery vehicles per day) = 0.8615661 and thus the number of days in a year in which to expect a reliable delivery is:

P(reliable delivery vehicles per day)xP(reliable delivery drivers per day)x365

0.8615661x0.9344269x365

= 293.8508 = 293 days

Therefore adding one extra vehicle has very little to no change on the overall number of days in a year that a delivery will be on reliable.

7. Conclusion

In conclusion the data contains crucial information which allows for analysis of different products, why it was bought and when it was bought as well as the delivery times that it incurred before getting to customers.

The ‘Technology’ and ‘Luxury’ items make up majority of the sales revenue for the company, far exceeding all the other classes combined. ‘Food’ and ‘Sweets’ had the fastest delivery times and this is likely due to the low shelf life that these products had which required prompt delivery times. Furthermore the items that were ‘Recommended’ and which fell in the ‘Technology’, ‘Luxury’ and ‘Household’ class made up 54% of the total revenue of the company.

Adding on to this; ‘Household’, ‘Luxury’ and ‘Gift’ items contained many samples that had delivery time means that fell outside of the initialised control limits for the mean. ‘Household’ and ‘Gift’ items need to be investigated further to determine the cause of this. Furthermore, in order to optimise the ‘Technology’ delivery time the business calculated that decreasing the mean delivery time by 3 hours (from 20.01095h to 17.01095h) will give the lowest cost loss and thus optimise the delivery process. After doing a MANOVA test it was determined that both ‘Class’ and ‘Why.bought’ products are dependent on ‘Price’ and ‘Delivery time’ per item. It was also determined that the company had 293 expected reliable delivery time days in a year.

Lastly the company was heading in a good direction and had a stable and reliable revenue. There were some delivery times which should be looked at and imporved, but over all the information given gained the company valuable input into understanding who the company sold to, what they sold, how they sold it as well as how long it took to deliver the products to clients.

8. References

Hessing, T. (2021) *X bar S control chart*, *Six Sigma Study Guide*. Available at: https://sixsigmastudyguide.com/x-bar-s-chart/ (Accessed: October 13, 2022).

Macros, Q.I. (2020) *What are control chart limits - how do you calculate control limits?*, *Control Chart Limits | UCL LCL | How to Calculate Control Limits*. KnowWare International Inc. Available at: https://www.qimacros.com/free-excel-tips/control-chart-limits/index2.php (Accessed: October 13, 2022).

Dirkse van Schalkwyk, T. (2022) Statistical Methods in Quality Assurance Part 1 summary. Quality Assurance. QA344. Stellenbosch University.