ECSA Graduate Attribute 4 Project

BY: CORNELIA MARGARETHA KLOPPERS - 23958502

The Department of Industrial Engineering, Stellenbosch University.

19 October 2022

Cornelia Margaretha Kloppers (Getha)

Third year Industrial Engineering undergraduate student

23958502

Abstract

This report is a discussion on the statistical analysis of the sales data of an online business. The project was programmed in R and all the visualisation in this project was therefore also generated in R-studio.

The aim of the analysis was to identify certain buying patterns and trends within the data. These patterns and trends will be discussed in this report. Statistical process control was performed on the delivery times of this company to determine the stability of their processes. Recommendations were made on how certain processes could be improved, for example, reducing the average delivery time of technology items by 3 days will give optimal results. Furthermore, the reliability of current services were calculated and discussed.

After the explorative data analysis was done, it was important to conclude that the company's processes is not flawless and that corrective measures should be taken to improve their current processes.

Table of contents

List of figures

List of tables

Glossary

CL – centre line

H_0 – null hypothesis

IQR – interquartile range

LCL – lower control limit.

LSL – lower specification limit

NA – no value is available

UCL – upper control limit

USL – upper specification limit

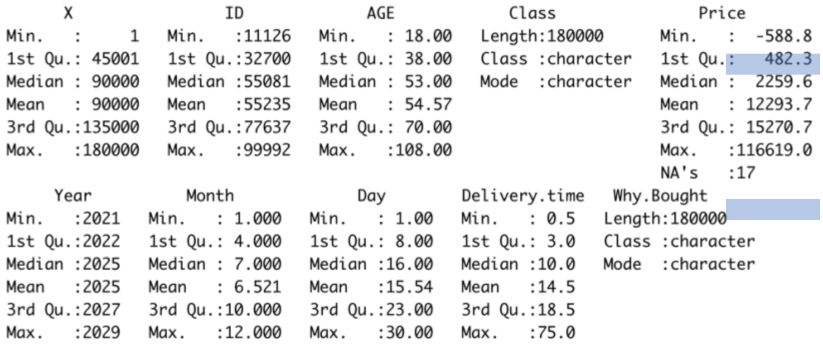
Introduction

Client sales data of an online business was given to be analysed by descriptive statistic measures. Different buying patterns have to be identified and recommendations have to be made on where corrective measures should be taken. The aim of this report is to measure the quality of the services of this online business and to discuss the knowledge gained from the data.

Firstly, the original dataset will be cleaned from any invalid instances. Descriptive statistics will be performed and the feature distributions and relationships will be discussed. Statistical process control will then be conducted and the results will be presented. The delivery process will be optimised, MANOVA tests will be performed on features and the reliability of current services will be measured.

1. Data wrangling

After getting insights of the data, it was clear that there were several invalid instances that had to be removed. The data was ordered in terms of year, month, day and index values. After that, the ordered dataset was split into two different datasets, one containing all the valid instances and the other all the invalid instances. A new primary ID column was added for each of these instances, while the original ID was kept as the secondary ID. This was done to keep track of which instances was removed. Whenever new instances are added to the data, the code can be executed and the two datasets will update automatically.



```
       X                ID              AGE             Class               Price
 Min.   :     1   Min.   :11126   Min.   : 18.00   Length:180000      Min.   :  -588.8
 1st Qu.: 45001   1st Qu.:32700   1st Qu.: 38.00   Class :character   1st Qu.:   482.3
 Median : 90000   Median :55081   Median : 53.00   Mode  :character   Median :  2259.6
 Mean   : 90000   Mean   :55235   Mean   : 54.57                      Mean   : 12293.7
 3rd Qu.:135000   3rd Qu.:77637   3rd Qu.: 70.00                      3rd Qu.: 15270.7
 Max.   :180000   Max.   :99992   Max.   :108.00                      Max.   :116619.0
                                                                      NA's   :17
      Year           Month             Day          Delivery.time   Why.Bought
 Min.   :2021   Min.   : 1.000   Min.   : 1.00   Min.   : 0.5    Length:180000
 1st Qu.:2022   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 3.0    Class :character
 Median :2025   Median : 7.000   Median :16.00   Median :10.0    Mode  :character
 Mean   :2025   Mean   : 6.521   Mean   :15.54   Mean   :14.5
 3rd Qu.:2027   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:18.5
 Max.   :2029   Max.   :12.000   Max.   :30.00   Max.   :75.0
```

Figure 1:Summary of original dataset

Seventeen missing values were removed. All of these missing values occurred in the “price” column. These values had to be removed because calculations with the price column would result in errors in the presence of “NA” values. From the summary of the data, five negative price values was noticed and removed. Price can only have a positive value and therefore these instances was considered invalid.

2. Descriptive Statistics

2.1 Data quality: continuous features

The valid dataset consists of 179 978 instances. Eight of ten features in the dataset was considered continuous features. All of these features is of the type "integer", except for price which is of the type "numerical". After calculating various statistical measures and feature characteristics, a data quality table was created and the data was investigated. It is clear that there is no missing values, as the incomplete instances had already been removed.

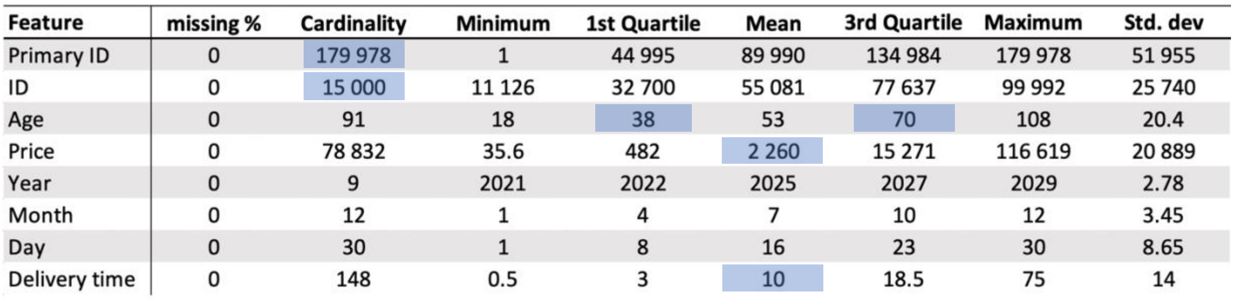
Primary ID and ID: Primary ID has a cardinality equal to the total amount of instances. As expected, this means that each instance has a unique Primary ID. Primary ID will have an evenly distributed graph. ID, on the other hand, has a cardinality of 15 000. This ID possibly represents the customers of the business. The ID feature has a range of 11 126 to 99 992.

Age: Fifty percent of the customers are between the age of 38 and 70 with a mean age of 53 years.

Price: The minimum price value is R35.60, while the highest price value is R116 619. Due to the large difference between the third quartile and the maximum, possible outliers was investigated. An instance was considered an outlier if it was more than 1.5 times the interquartile range from the 1st or 3rd quartile. There was 22 171 outliers detected for this feature. This can be due to more expensive products or business-to-business transactions.

Year, Month and Day: From figure 2 it is clear that the dataset contains data from 2021 to 2029. When looking at the statistical values month and day, it is clear that there are data from all 12 months of the year. There is also 30 days in a month as expected. Therefore, nothing unusual was found within these features.

Delivery Time: The mean delivery time is 10 hours with an interquartile range (IQR) of 15.5 hours. Due to large maximum of 75 hours in comparison to the IQR, further investigation led to 17 516 outliers. This can be due to certain products that have delivery complexities, or customers require deliveries over a longer distance as they are far away from where the warehouse or shop is situated.

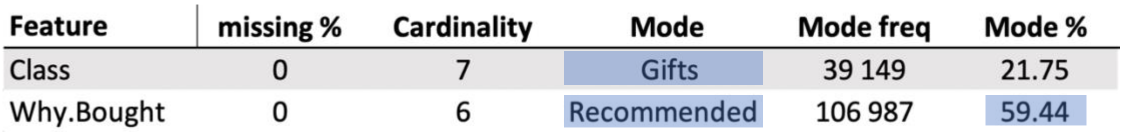


Feature	missing %	Cardinality	Minimum	1st Quartile	Mean	3rd Quartile	Maximum	Std. dev
Primary ID	0	179 978	1	44 995	89 990	134 984	179 978	51 955
ID	0	15 000	11 126	32 700	55 081	77 637	99 992	25 740
Age	0	91	18	38	53	70	108	20.4
Price	0	78 832	35.6	482	2 260	15 271	116 619	20 889
Year	0	9	2021	2022	2025	2027	2029	2.78
Month	0	12	1	4	7	10	12	3.45
Day	0	30	1	8	16	23	30	8.65
Delivery time	0	148	0.5	3	10	18.5	75	14

Table 1: Data quality report: continuous features

2.2 Data quality: categorical features

There are two categorical features in the dataset, “class” and “why bought”. There are seven different classes of products of which gifts are sold the most. From the sales, 59.44% is gained due to a recommendation by someone. Therefore, word by mouth is very important for this business.



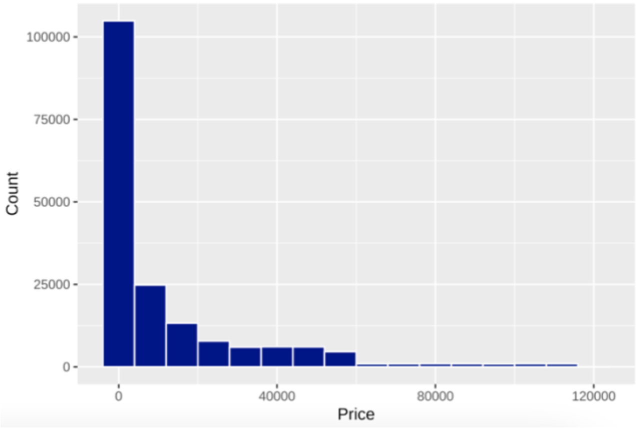
Feature	missing %	Cardinality	Mode	Mode freq	Mode %
Class	0	7	Gifts	39 149	21.75
Why.Bought	0	6	Recommended	106 987	59.44

Table 2: Data quality report: categorical features

2.3 Visualizations: Distribution of values

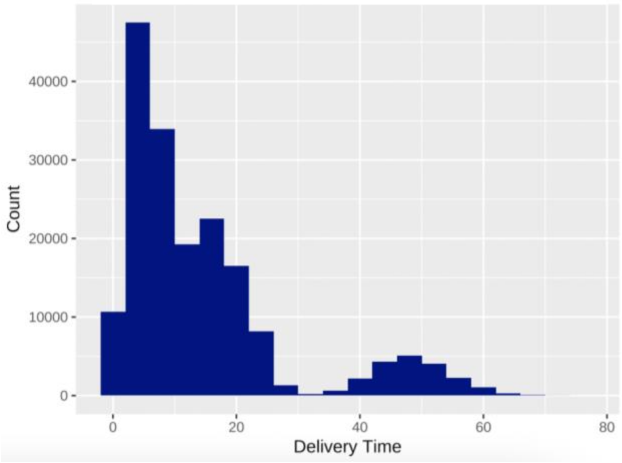
Histograms will be displayed for the relevant continuous features and bar plots for the categorical features.

Figure 2: Histogram of Price feature distribution



The price feature distribution was right-skewed. There were many sales of low prices and very few sales of a large prices. The outliers that was previously mentioned can now be visually seen by looking at the tail of the graph.

Figure 3: Histogram of delivery time distribution



As there are two peaks in the graph, delivery time is considered a bimodal distribution. After removing the outliers of the data, the data was a bit skewed to the right as seen in figure 4. Figure 5 was then drawn of the delivery times above 30 hours and it was found to be relatively normally distributed. The reason for this two peaks should be investigated.

Figure 4: First peak delivery time distribution

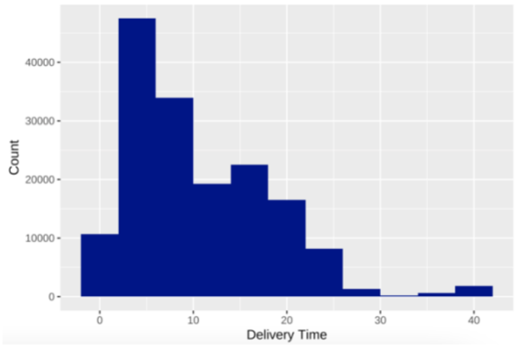
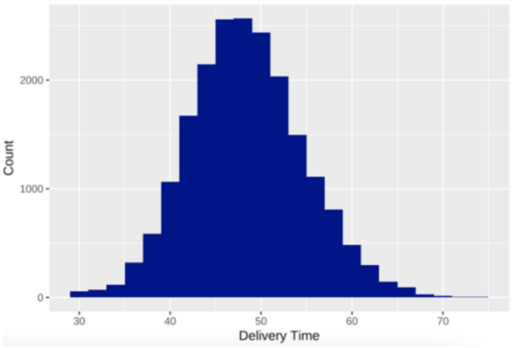


Figure 5: Second peak delivery time distribution



The bar plots of the distribution of the two categorical features can be seen in figure 6 and 7. From the plots it is clear that “gifts” and “recommended” is the two modes. Figure 7 also supports the statistic that “recommended” covers almost 60% of all the instances.

Figure 7: Distribution of class feature

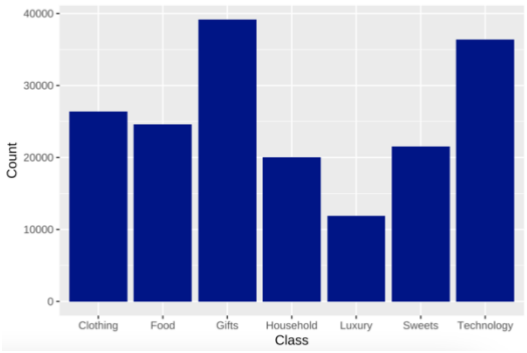
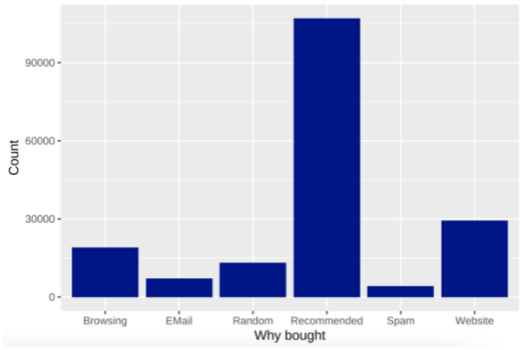


Figure 6: Distribution of why Bought



2.4 Visualizations: Relationships between features

A few graphs was drawn to investigate the relationship between the features. Only the graphs that gave clear insights on the relationships and trends in the data, will be discussed in this section.

As mentioned in section 2.3, the cause of the two peaks in the distribution of delivery times had to be investigated. Considering figure 8, it is clear that the second peak is completely caused by the high delivery time of household products. It is also clear that the delivery times for “Gifts”, “Household” and “Technology” follows a relatively normal distribution. Figure 9 gives further confirmation that “Food” and “Sweets” has a short average delivery time. This makes sense as it can be perishable and therefor customers expect fast delivery of these goods. Household products have the largest average delivery time. This can be due to delivery complexities like the size of the products that have to delivered.

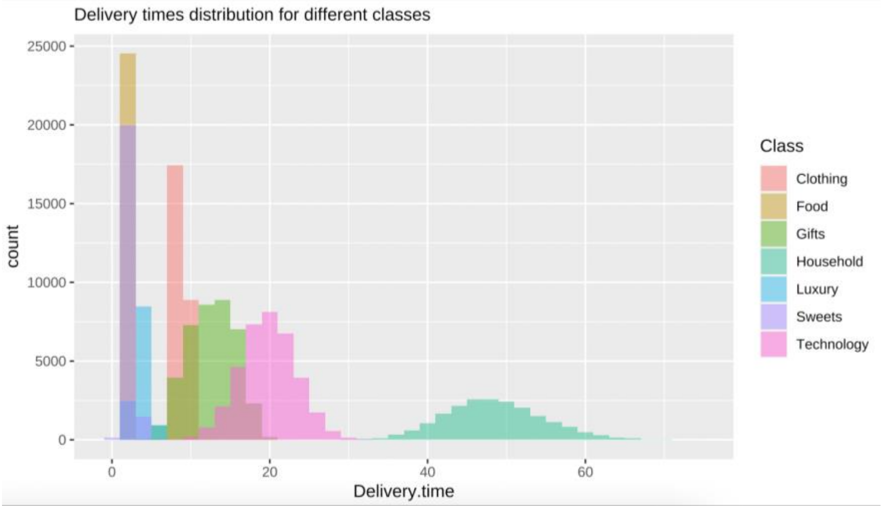


Figure 8: Histogram of delivery times per product class

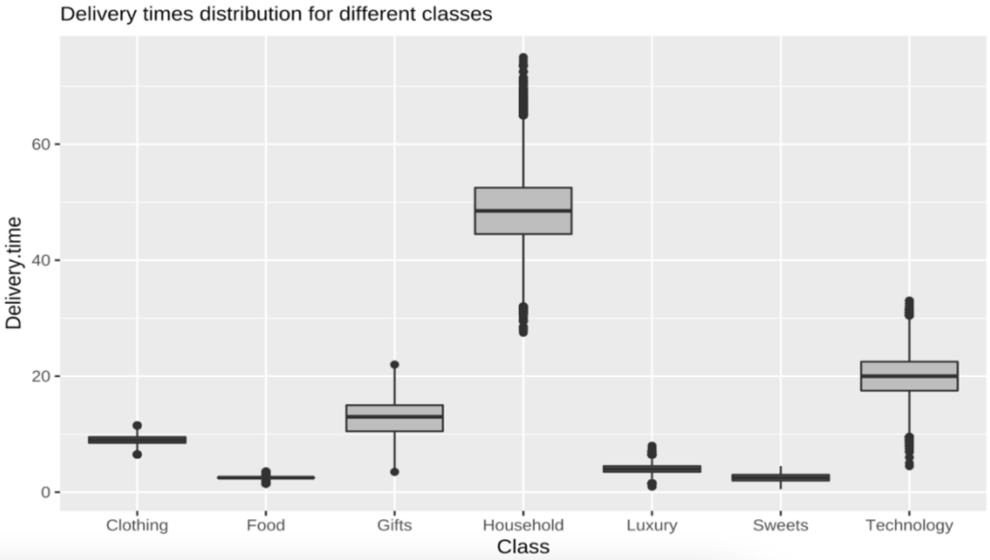


Figure 9: Boxplot of delivery times per product class

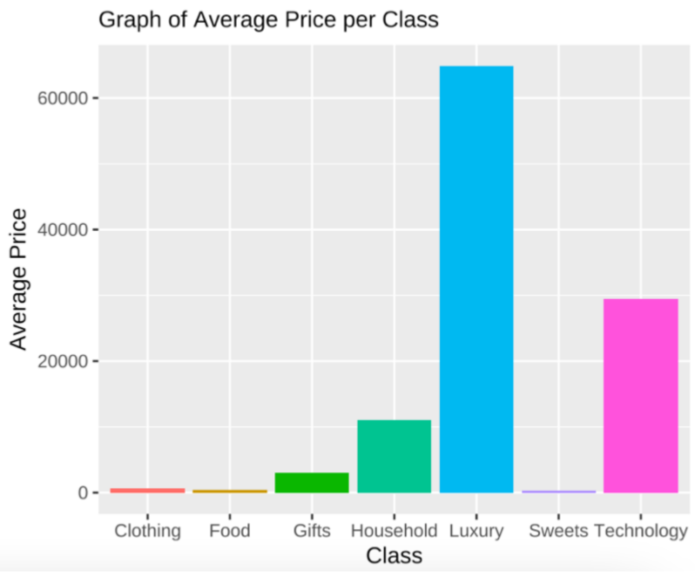


Figure 10: Average Price of sales per Class

As expected, the average price of luxury goods is the highest when investigating figure 10. Technology and household products have an average price which is considerably less than luxury goods, but they can still be valued as expensive products. Extra effort should go into the control and management of these goods.

In figure 11 it is clear that there is no seasonality in monthly income when looking at the total sales over years 2021 to 2029. In figure 12 it can be seen that there is variability in the total income of each month. It is clear that, for example, income will always increase in December months. It is also important to notice in figure 13 that there were a lot more transactions in 2021 than any other year. The amount of transactions per class is relatively constant from 2022 onwards.

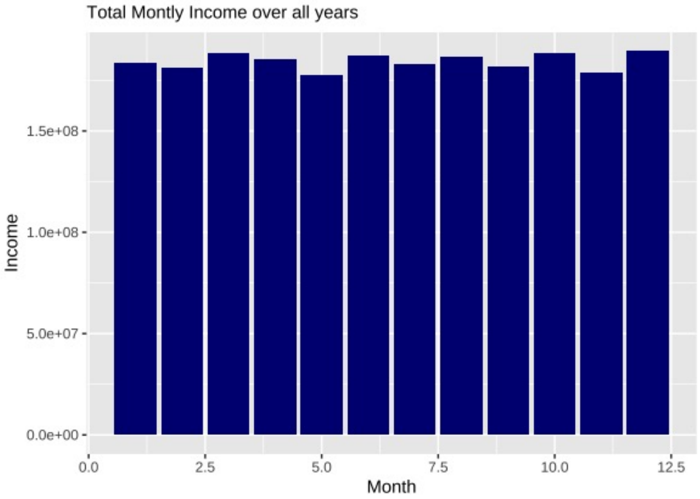


Figure 11: Total monthly income over all the years

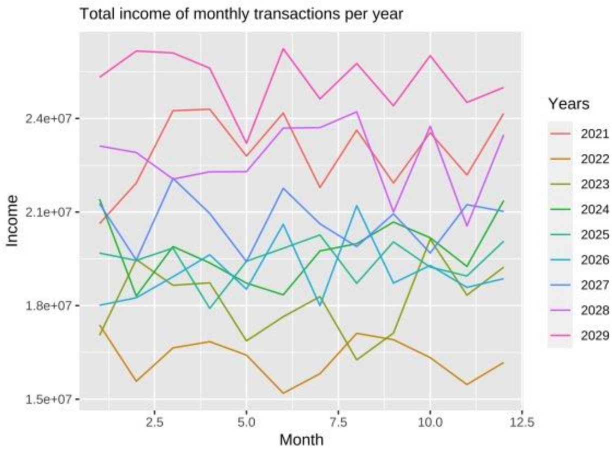


Figure 12: Line plot of total monthly income per year

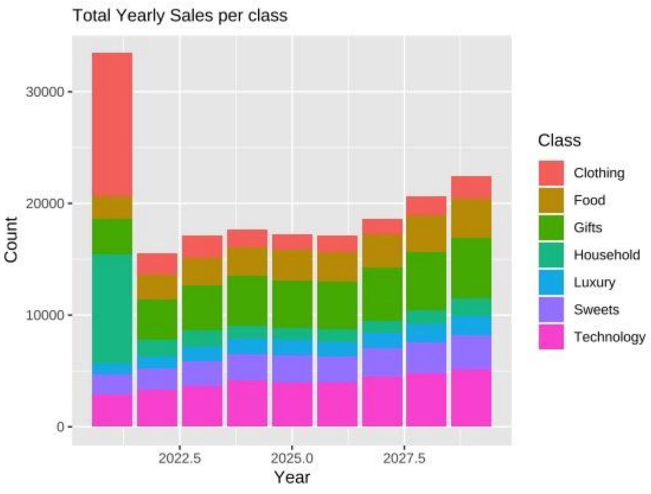


Figure 13: Barplot of yearly transactions by class

2.5 Process Capability Indices

Table 3: Process Capability Indices

Table 3: Process Capability Indices	
C_P	1.142
C_{PU}	0.380
C_{PL}	1.905
C_{PK}	0.380

Given an upper specification limit (USL) of 24 hours and a lower specification limit (LSL) of zero hours for delivery times, the process capability indices were calculated for the class "technology". The mean delivery time was calculated to be 20.01 hours. An LSL of zero is sensible as delivery time cannot be a negative value and thus the smallest possible value is 0.

C_P shows whether a distribution will fit within the specified USL and LSL limits. A value of 1 indicates that the distribution will fit perfectly within the limits while values larger than 1 indicates that the distribution is wider than the given limits. C_{PK} is an indication of where the average lies with reference to the centre of the specification. The larger the difference between these two values, the more offset is the overall average (PQSystems, 2022). From the capability indices in table 3, it is clear that the distribution has shifted to the right which means that the average delivery time is larger than the centre (12) of the two limits.

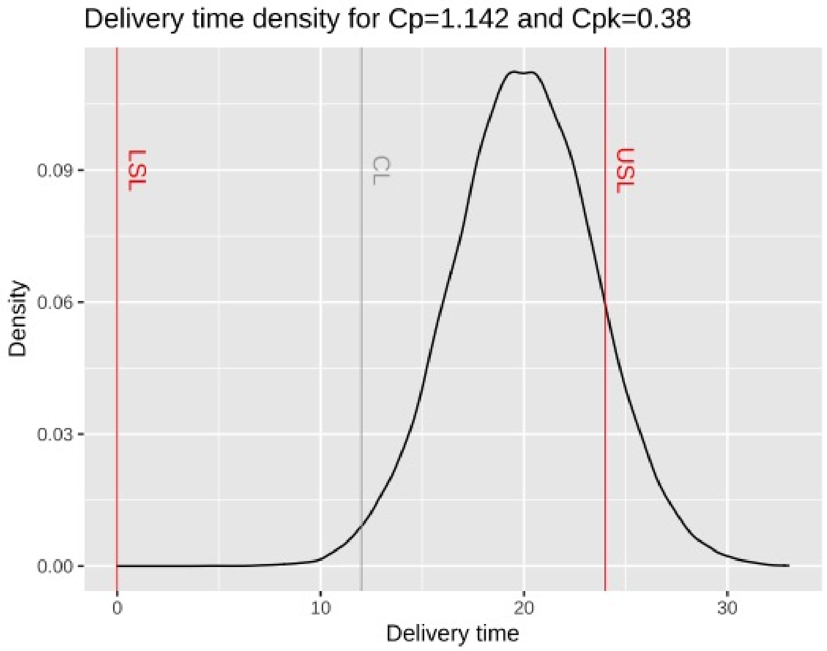


Figure 14: Delivery time distribution for technology

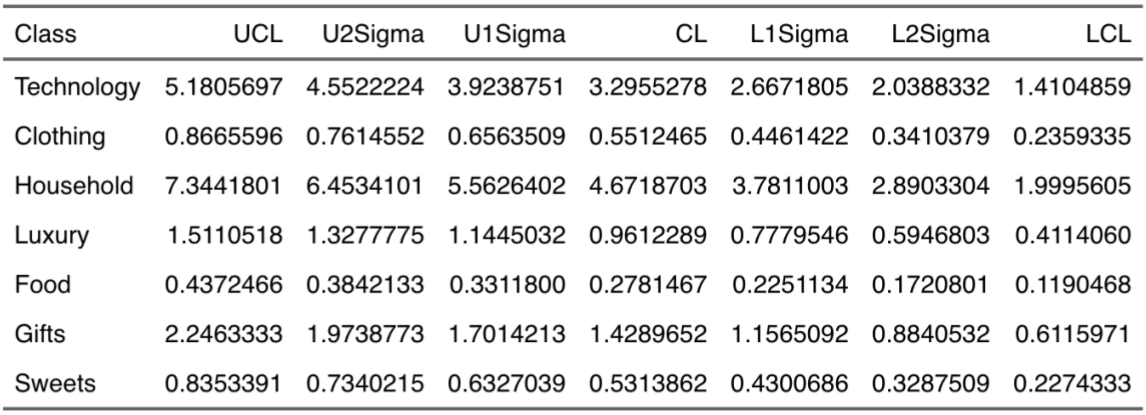
3. Statistical Process Control

Statistical Process Control (SPC) is the use of statistical techniques to measure, monitor and control the performance of a process. The process involves analysing data to determine process behaviour and variation (Hessing, n.d.). Before performing the analysis, the data was ordered according to ascending year, month, day and index. The delivery times was sampled in sets of 15 for each class respectively.

3.1. X-bar and S Charts: 30 samples

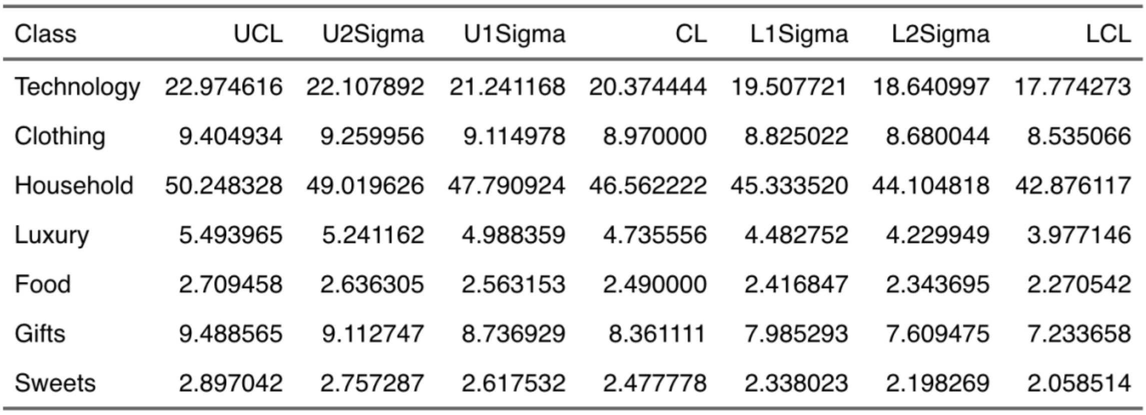
X-bar and S tables:

The 30 oldest data samples, according to date, were used to calculate the control limits for the X and S-chart respectively. The X-chart uses the mean of each sample to determine the control limits. For the S chart limits, the standard deviation of each sample were calculated as it is a sensitive indicator of the variability in each class. The different limits that were calculated for each class can be seen in table 4 and 5.



Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.1805697	4.5522224	3.9238751	3.2955278	2.6671805	2.0388332	1.4104859
Clothing	0.8665596	0.7614552	0.6563509	0.5512465	0.4461422	0.3410379	0.2359335
Household	7.3441801	6.4534101	5.5626402	4.6718703	3.7811003	2.8903304	1.9995605
Luxury	1.5110518	1.3277775	1.1445032	0.9612289	0.7779546	0.5946803	0.4114060
Food	0.4372466	0.3842133	0.3311800	0.2781467	0.2251134	0.1720801	0.1190468
Gifts	2.2463333	1.9738773	1.7014213	1.4289652	1.1565092	0.8840532	0.6115971
Sweets	0.8353391	0.7340215	0.6327039	0.5313862	0.4300686	0.3287509	0.2274333

Table 4: Control limits for S-charts



Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22.974616	22.107892	21.241168	20.374444	19.507721	18.640997	17.774273
Clothing	9.404934	9.259956	9.114978	8.970000	8.825022	8.680044	8.535066
Household	50.248328	49.019626	47.790924	46.562222	45.333520	44.104818	42.876117
Luxury	5.493965	5.241162	4.988359	4.735556	4.482752	4.229949	3.977146
Food	2.709458	2.636305	2.563153	2.490000	2.416847	2.343695	2.270542
Gifts	9.488565	9.112747	8.736929	8.361111	7.985293	7.609475	7.233658
Sweets	2.897042	2.757287	2.617532	2.477778	2.338023	2.198269	2.058514

Table 5: Control limits for X-charts

X-bar and S Charts:

- In the following graphs, the horizontal orange lines represents the outer control limits (UCL and LCL) and the horizontal blue line represents the centre line. It is important that the S-Chart is analysed first as it is an indication of the validity of the samples. Values outside of the outer control limits are considered outliers and should be removed. Accurate analysis of the X-bar charts can only be done if there are no “out of control” samples in the S-chart.

The S-Charts of the 30 samples for each class can be seen in figure 15. “Sweets” and “Food” each has one sample with values higher than their upper control limit. These samples should be removed before an accurate analysis can be done on their X-charts.

Figure 15: S-charts of 30 samples

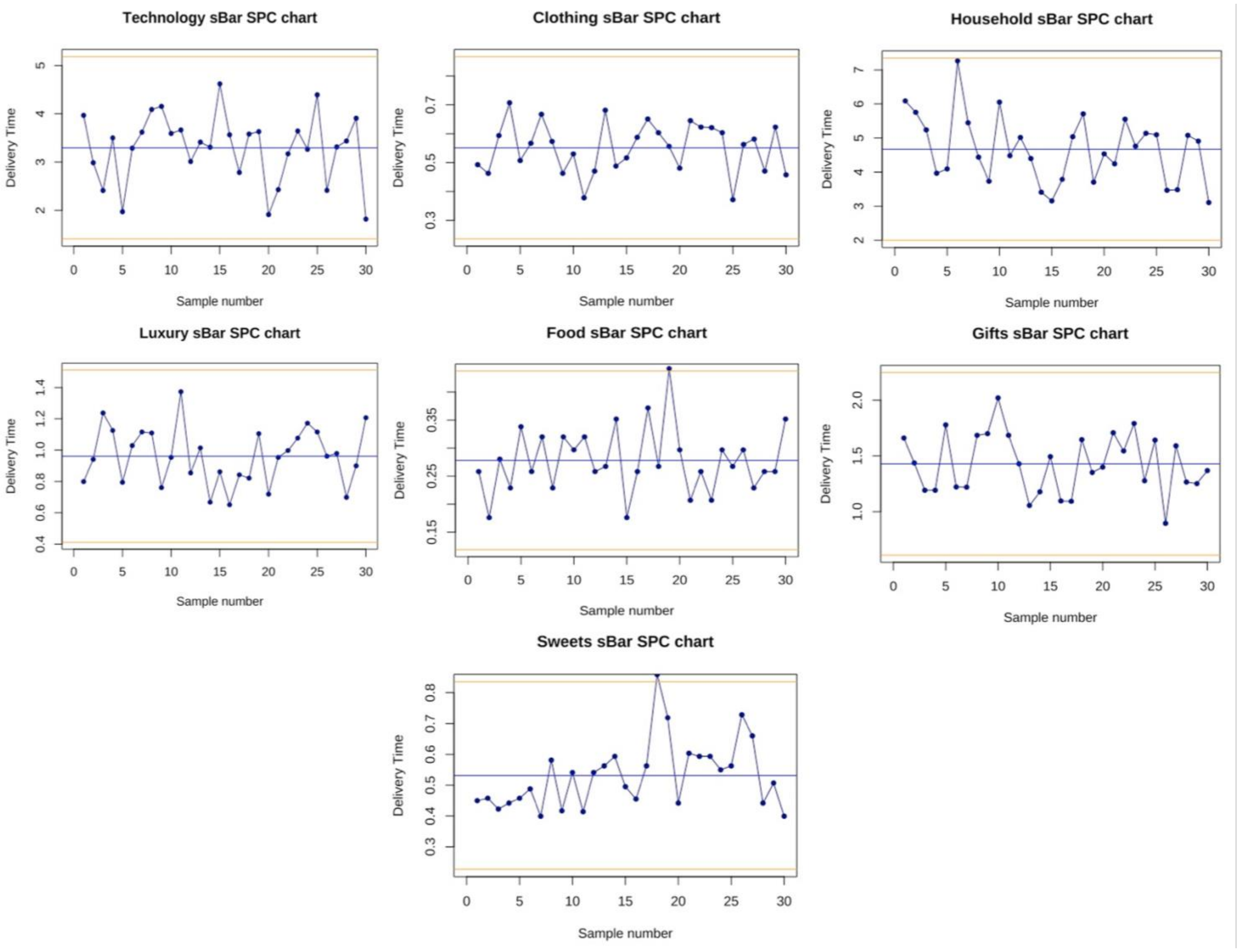
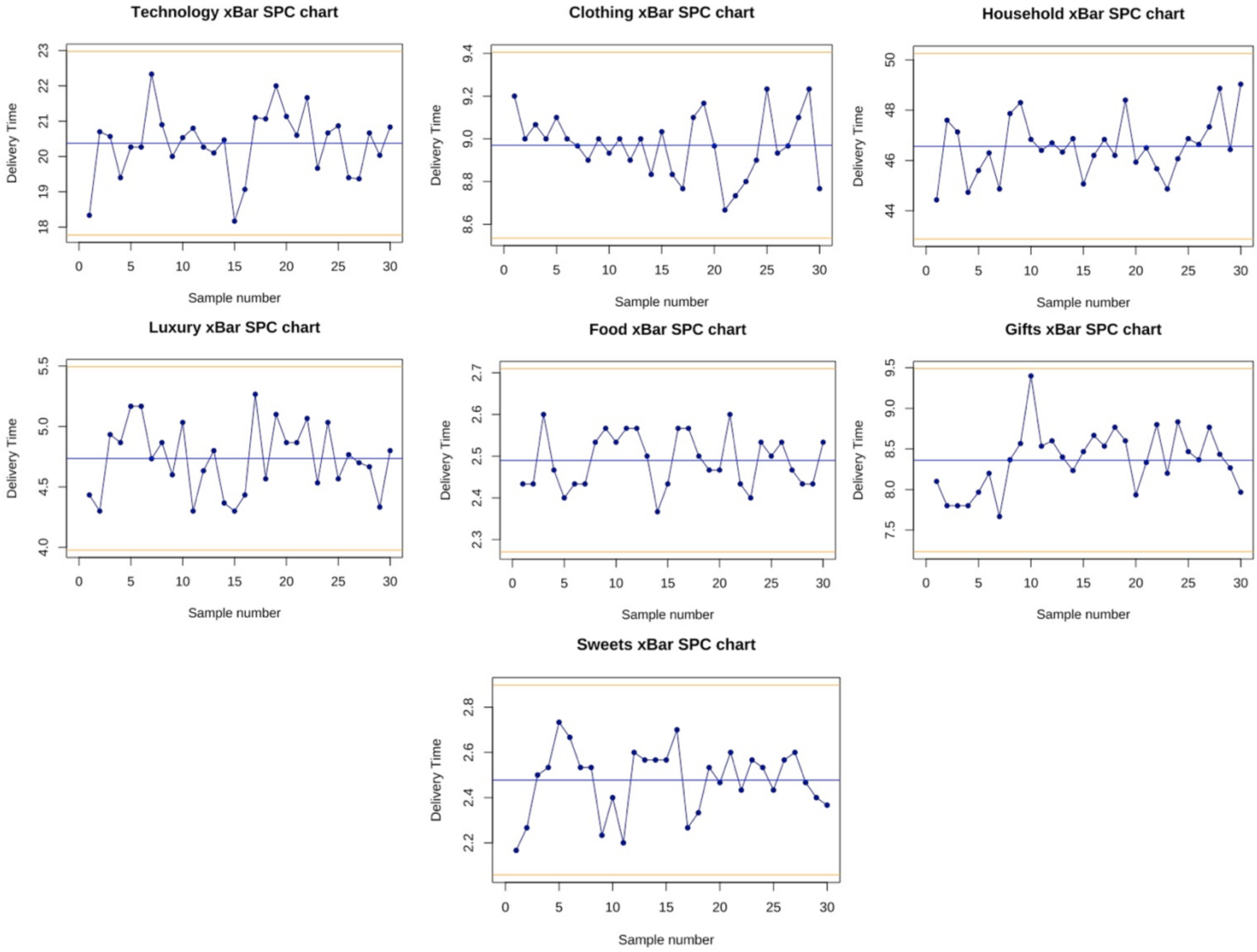


Figure 16 is a representation of all the X-charts of the first 30 samples of each class. The X-chart for “Food” and “Sweets” cannot be used for analysis of the samples against the control limits. All the other classes are in control and there are also no samples outside of the control boundaries in the X-charts.

Figure 16: X-charts of 30 samples



3.2. Process Control of all samples

As the process continued, more samples were generated and the X-bar and S charts were drawn for all the samples in each class. Information regarding the process stability and control was extracted from the graphs and will be discussed in this section.

The assumption was made that the samples of each class is normally distributed. Recall that the S-charts is evaluated first and then the X-charts. When the S-chart has samples that is out of control, the cause of these irregularities has to be identified. Only after the S-chart is corrected, the X-chart can be analysed. In the following graphs, none of the out of control samples were removed and therefore not all of the X-charts represents accurate results.

Technology:

In the S-chart for technology, only a few random instances are exceeding the upper limit and is "out of control". As all the outliers exceeds the UCL, it is possible that there was a slight increase in the average standard deviation of the samples. From the S-chart, the process is considered stable and in control. The X-chart further motivates this statement with few instances out of control.

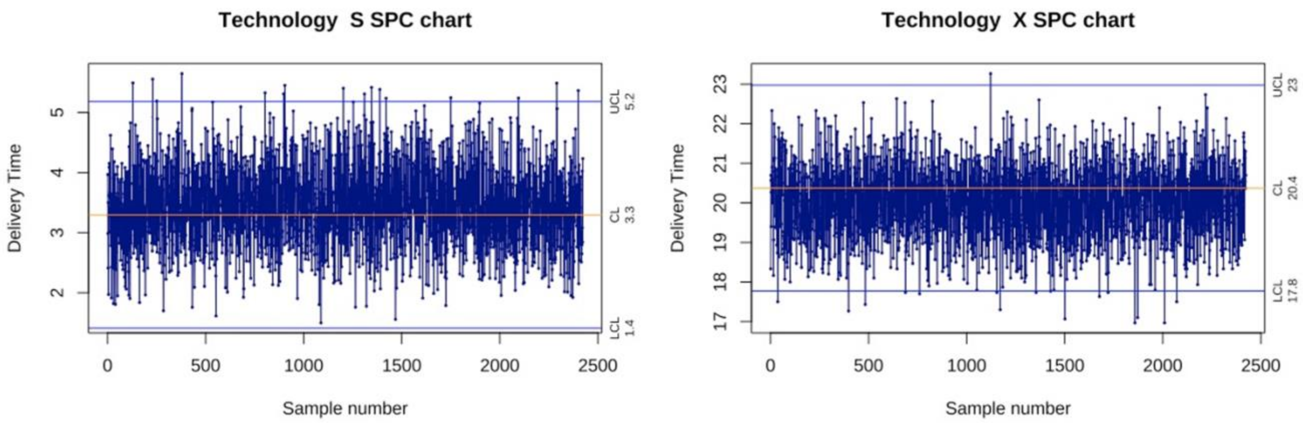


Figure 17: Xbar-S Charts for Technology

Clothing:

From the S-chart, it is clear that there is an upward trend in the standard deviation for the delivery of clothing. Due to this gradual increase, the delivery process for clothing is considered unstable. The cause of this trend needs to be investigated. Even though the X-chart looks promising with only a few outliers, this chart cannot be used for an accurate analysis.

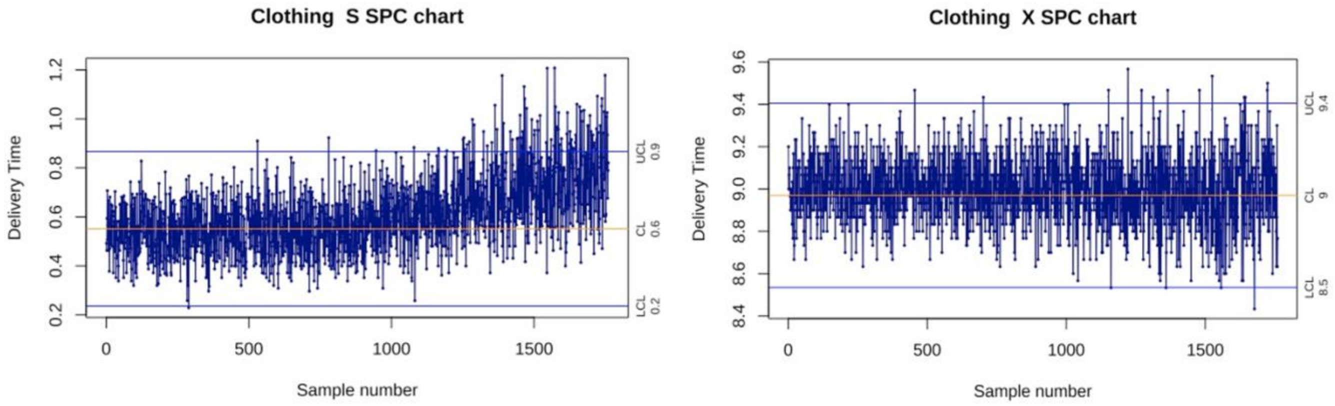


Figure 18: Xbar-S Charts for Clothing

Household:

When analysing the S-chart, an upward trend was noticed in the standard deviation of the delivery times for household products. In the X-chart this upward trend is further highlighted, with almost all samples out of control from the 900th sample and onward. The process is thus unstable and out of control.

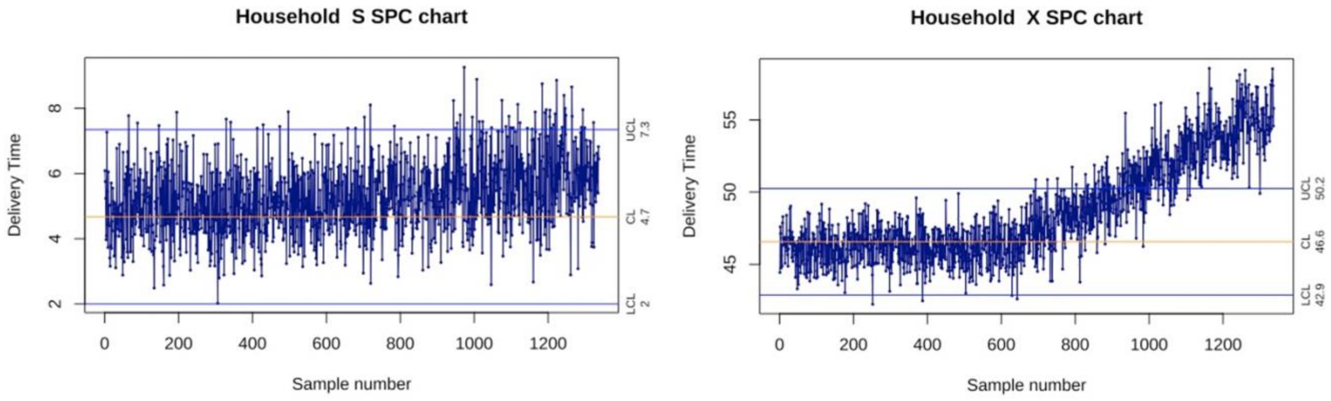


Figure 19: Xbar-S charts for Household

Luxury:

In the S-chart there is a downward trend in the sample standard deviation for luxury goods. Even though there are only three samples outside of the control limits, the values is not evenly distributed around the centre line. The S-chart is within limits and therefore the X-chart can be used to draw accurate conclusions of the process. The variation in the process can clearly be seen in the X-chart and it provides more insight about the clear downward trend in the average delivery time of luxury products. The process is unstable and the cause of this shift needs to be investigated as all the samples are currently below the LCL.

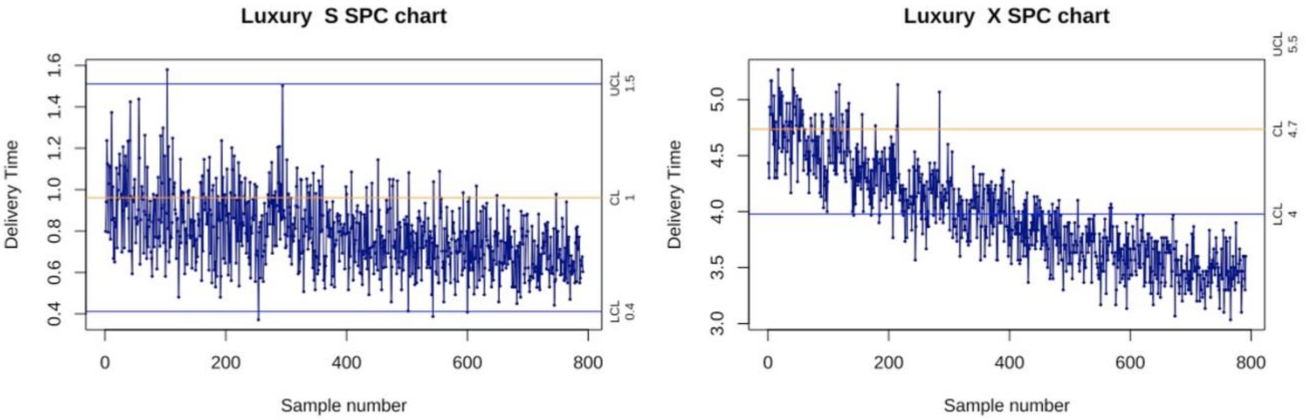


Figure 20: Xbar-S Charts for Luxury

Food:

The S-chart for food has no trend, but there are 5 outliers below the LCL with a standard deviation of zero. The reason for this occurrence should be investigated and the samples should be removed before analysing the X-chart. The samples that are in control in the S-chart can accurately be analysed in the X-chart. As the majority of samples are in control in the S-chart and only a few samples are “out of control” in the X-chart, the process is considered stable.

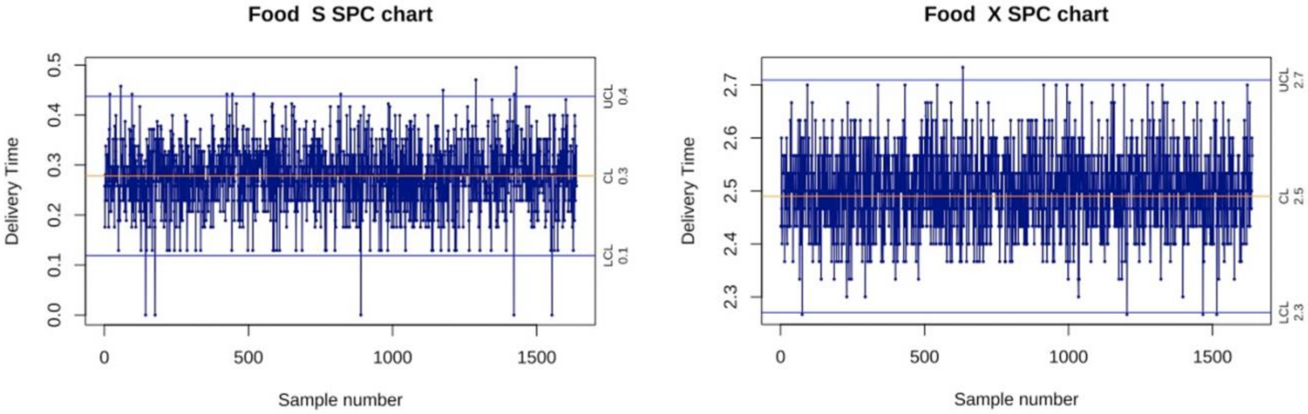


Figure 21: Xbar-S Charts for Food

Gifts:

The S-chart for gifts has no clear trend and there are only a few values above the UCL. The X-chart can thus be used to draw accurate conclusions. When looking at the X-chart for gifts, there is a definite upward trend in the average delivery time of gifts. This process is unstable and the cause of the increasing delivery times should be investigated.

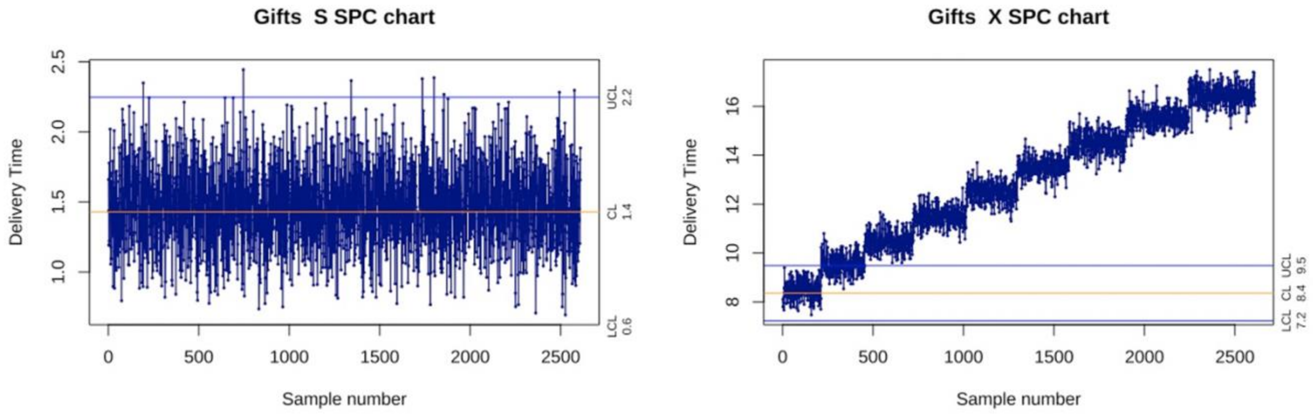


Figure 22: Xbar-S Charts for Gifts

Sweets:

There is no noticeable trends in the standard deviation for delivery time of sweets. There is only one outlier in the S-chart and this sample was already noticed in the sweets graph in figure 15. The X-chart also contains only five samples that is outside of the control limits. The process is considered stable and in control.

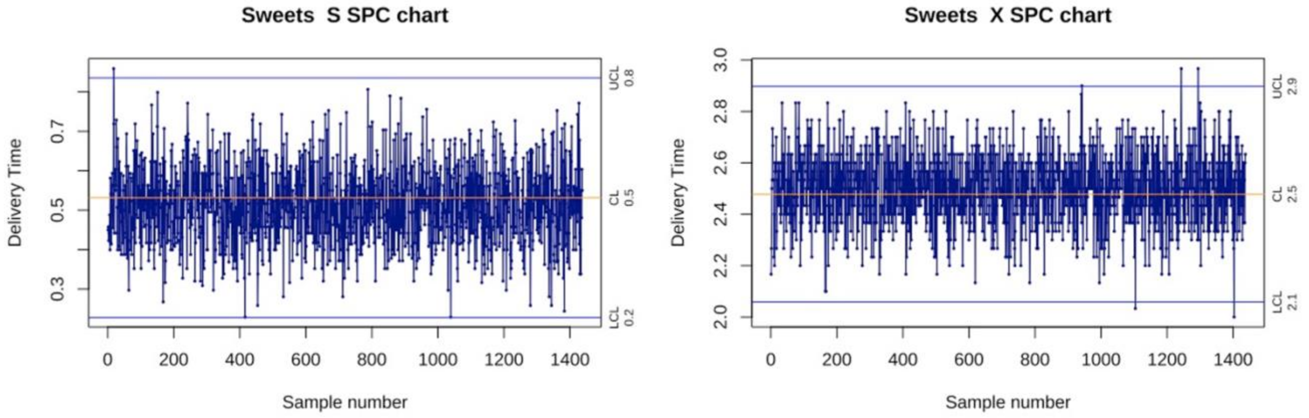


Figure 23: Xbar-S Charts for Sweets

4. Optimisation of the delivery process

Further analysis had to be done on the control charts of each class. This will help to investigate and accurately identify problems that needs to be corrected in each process.

4.1 Xbar-S Charts with Rules

Rule A:

A test was conducted to identify which samples were not within the specified control limits of each class. The control limits are +3 and -3 sigmas away from the average or centre line. The total amount of "out of control" samples is indicated in the table below, as well as the sample numbers of the first 3 and last 3 outliers of each class.

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

Table 6: Samples identified with Rule A

The "Food" and "Sweets" class have five outliers each. The X-charts in figure 21 and 23 serves as confirmation that this statistic is correct. In figure 24 and 25 where the red dots indicate outliers, it can be seen that these outliers are random and therefor, these samples can be removed. The process does not necessarily need any improvements.

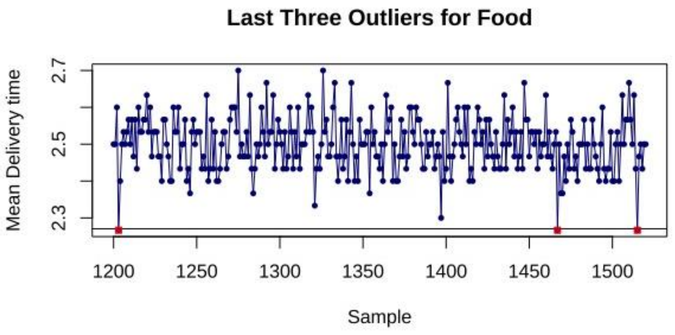


Figure 24: Last three outliers for Food

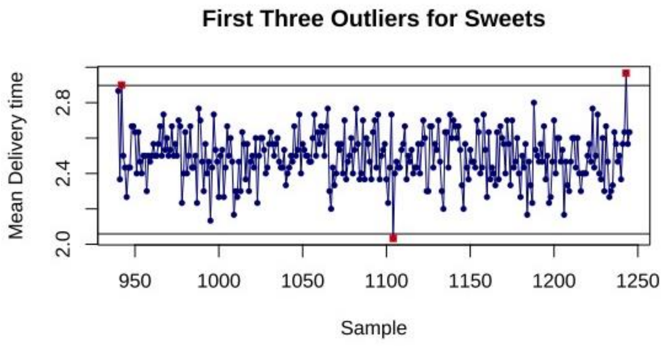


Figure 25: First three outliers for Sweets

When investigating the outliers for household products, it is clear that the first three outliers is randomly below the LCL. The process reaches a stage where almost all the consecutive samples is outliers above the UCL, as seen in figure 27. This is an indication of the upward trend in the average delivery time for household products. The reason for this increase needs to be investigated and improvements should be made to the process to keep it stable. This increase could be due to the company expanding their variety of household products. An example of this could be that used to only sell household appliances but they expanded to furniture, which is larger and has more delivery complexities.

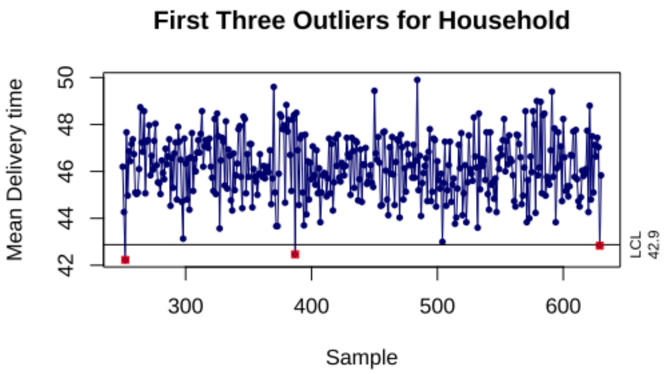


Figure 26: First three outliers for Household

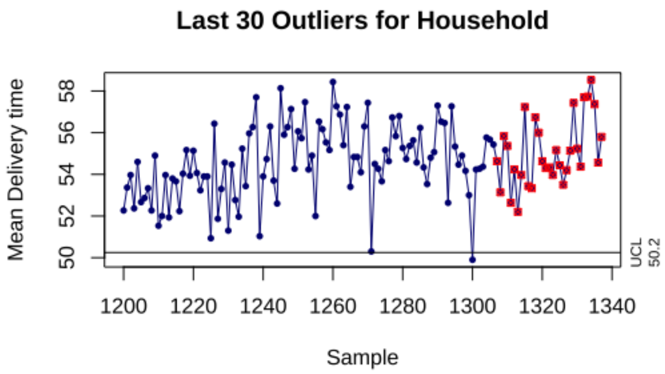


Figure 27: Last 30 outliers for Household

Rule B:

For rule B, the most consecutive S-bar samples between -0.3 and +0.4 sigma control limits were found. The most consecutive samples between those limits were seven for the class “Food”. This indicates that for the range of samples in figure ::, the standard deviation of the samples is very close to the mean standard deviation. The second longest consecutive samples within these limits is six, and it occurred in the “Technology” class. In the figures below, “LL” and “UL” indicates the lower and upper limits calculated with -0.3 and +0.4 sigma, respectively.

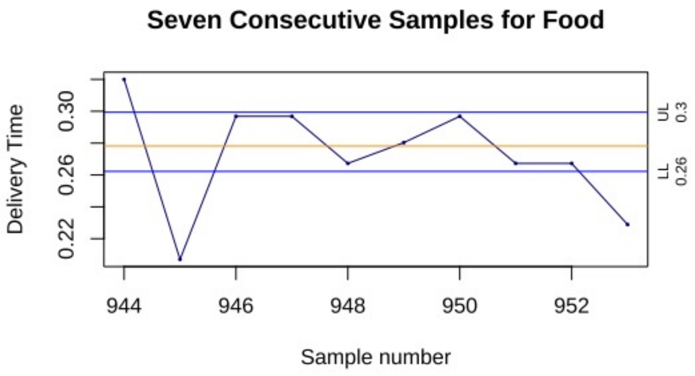


Figure 29: Consecutive samples for Food

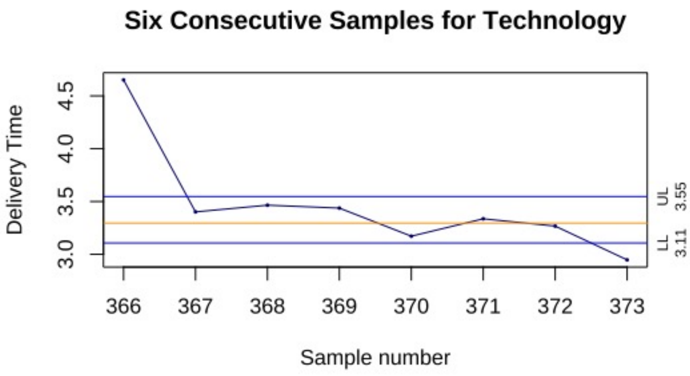


Figure 28: Consecutive samples for Technology

4.2 Probability of a Type 1 error

For a type 1 error, the null hypothesis (H_0) states that the process is in control and close to the centre line, where the closeness is measured relative to certain limits in rule A and B. A type 1 error (alpha) occurs when you reject the null hypothesis, when the null hypothesis is actually true (Bhandari, 2021). In other words, you treat the process as if it is out of control when it is stable. This is considered as a waste as you unnecessarily stop or investigate the process. The probability of making a type 1 error for each of the rules were calculated.

Rule A:

Z-values = +3 and -3

$$Type\ 1\ error = pnorm(-3) + \left(1 - pnorm(3)\right)$$
$$= 0.001349898 + (1 - 0.9986501)$$
$$= 0.002699796 = \mathbf{0.27}\%$$

There is a 0.27% probability that you will classify a process as “out of control” when it is not.

Rule B:

Z values = +0.4 and -0.3

$$Type\ 1\ error = (pnorm(0.4) - pnorm(-0.3))$$
$$= 0.6554217 - 0.3820886$$
$$= 0.2733332 = \mathbf{27.33}\%$$

There is a 27% probability that you will classify a process as “out of control” when it is not.

The probability of incorrectly rejecting the null hypothesis, and classifying 7 consecutive samples as “out of control” when they are actually in control, is 0.011% (see calculations below).

$$Type\ 1\ error = (pnorm(0.4) - pnorm(-0.3))^7$$
$$= (0.6554217 - 0.3820886)^7$$
$$= 0.0001139846 = \mathbf{0.011}\%$$

4.3. Optimising the reduction in delivery times for Technology products.

The company experiences a loss of R329/item-late-hour when the delivery time of technology products is more than 26 hours. The company can reduce the average delivery time at a cost of R2.5/item/hour. The optimal reduction in delivery time was optimised by minimizing the total cost.

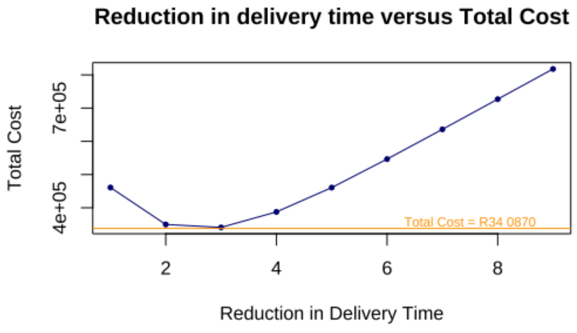


Figure 30: Reduction in delivery time vs cost

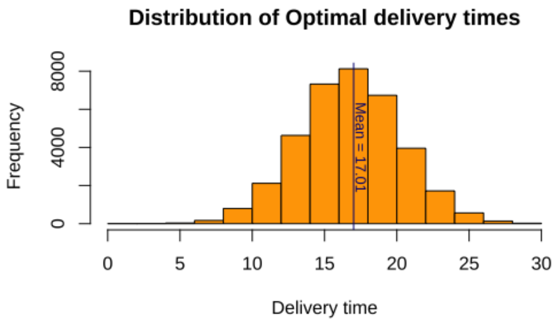


Figure 31: Histogram of optimal delivery times

From figure 30, it is clear that the optimal reduction in delivery time should be three hours. This will result in a cost of R340 870. The new distribution for delivery time can be seen in figure 31, with an optimal mean delivery time of 17.01 hours.

4.4 Probability of a Type 2 error

The null hypothesis states that the process is in control. A type 2 error occurs when the null hypothesis is not rejected when it is actually wrong (Bhandari, 2021). The probability that we fail to reject the null hypothesis when the process is out of control, was calculated. The UCL and LCL for the class “Technology” was used and the delivery process average was shifted to 23 days.

$\boldsymbol{UCL} = 22.97462$

$\boldsymbol{LCL} = \mathbf{17.77427}$

$\sigma = 0.8667238$

$\mu = 23$

$\boldsymbol{Type\ 2\ error} = \boldsymbol{pnorm}(\boldsymbol{UCL}, \mu, \sigma) - \boldsymbol{pnorm}(\boldsymbol{LCL}, \mu, \sigma)$

$= \mathbf{0.4883177} - \mathbf{8.234208}^{-\mathbf{10}}$

$= \mathbf{0.4883177} = \mathbf{48.83}\%$

There is a probability of 48.83% that we won’t reject the null hypothesis when the process, with a mean delivery time of 23, is actually out of control.

5. MANOVA

MANOVA (multivariate analysis of variance) tests are used to determine whether there are a relationships between various dependent variables (Chetty & Jain, 2021). A significance level of 0.05 was used in the analysis of the various tests. This means that there is a risk of 5% that the results will indicate that there is a relationship when there is no relationship.

The null hypothesis states that the variables being tested, has no significant influence on the variable that it is compared against (independent variable). The alternative hypothesis states that at least one of the variables influences the independent variable.

5.1 Test 1: Year and Month versus Price

The overall P-value for the test is 2.2^{-16}. This value is smaller than the significance level of 0.05 and therefor the null hypothesis should be rejected. This means that at least one of the features, Year or Month, has influence with regard to the Price.

Table 7: MANOVA test 1 results

Dependant Variable	P-Value	Analysis
Year	2.2^{-16}	The P-value is smaller than 0.05, therefore the year influences the Price. This is due to the increase in average price from 2021 to 2025 as seen in figure 32.
Month	0.3813	The P-value is larger than 0.5, therefore the month does not influence the price of the products.

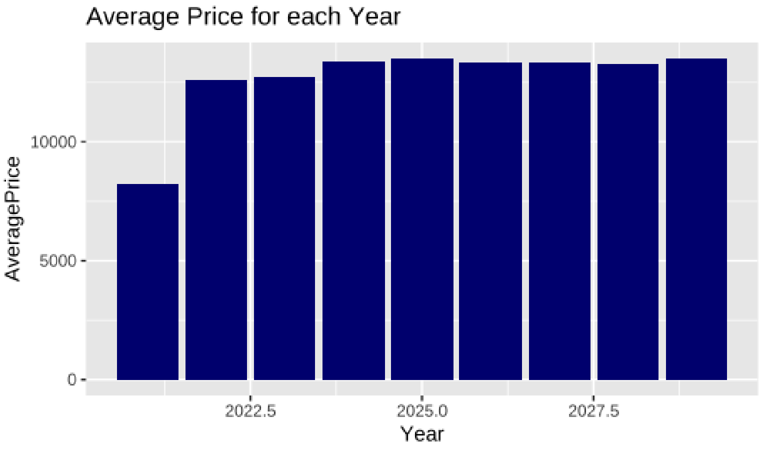


Figure 32: Average price for each year

5.2 Test 2: Delivery time and Price versus Class

The overall P-value for the test is 2.2^{-16}. As the significance level of 0.05 is larger than the P-value, the null hypothesis is rejected. This indicates that the features “Price” or “Delivery time” or both, have an influence on the delivery time.

Table 8: MANOVA test 2 results

Dependant Variable	P-Value	Analysis
Delivery Time	2.2^{-16}	The P-value is smaller than 0.05, therefore the Delivery time influences the Price. It is clear in figure 33 that the distribution for delivery time of each class is significantly different.
Price	2.2^{-16}	The P-value is smaller than 0.5, therefore the Price has influence on the Class and vice versa. Figure 34 is a representation of how the average price varies for each class.

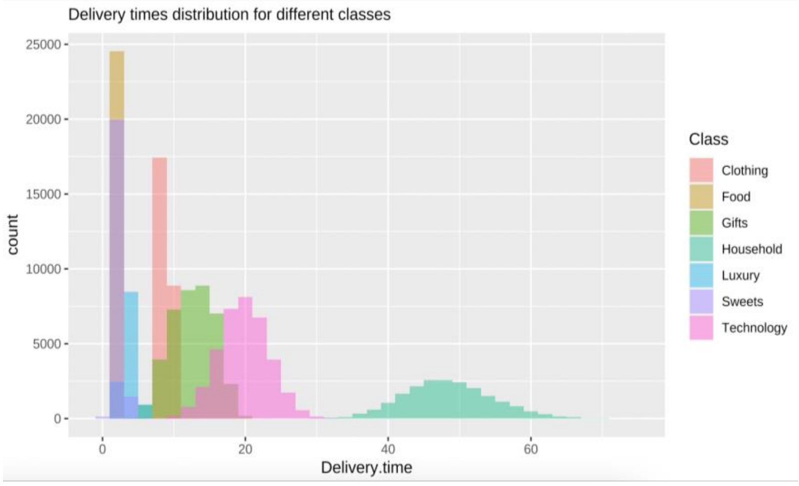


Figure 34: Distribution of delivery times for each class

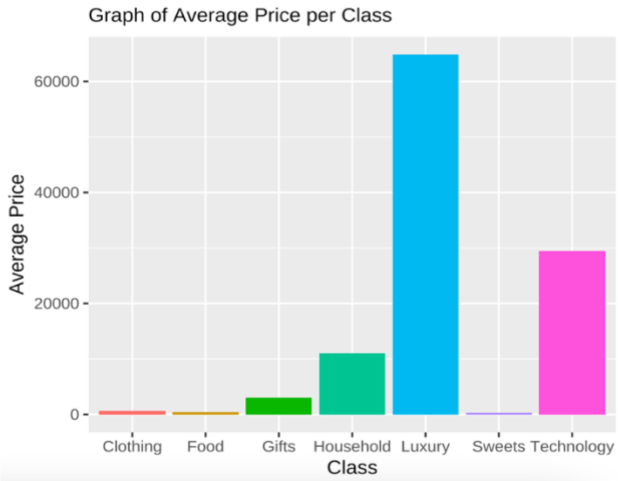


Figure 33: Average Price for each Class

6. Reliability of the service and products

6.1. Reliability of Food deliveries: Lafrideradora

The reliability the subsidiary company, Lafrideradora, had to be analysed. Lafrideradora manufactures parts that are used to keep products cool during transit. Information for the calculations was gained from problem 6 and 7 on page 363 in the 11th edition of *Managing for Quality and Performance Excellence* by *J. Evans* and *W.M. Lindsay*.

6.1.1. Taguchi Loss function of a refrigerator part - Problem 6

- $Thickness = 0.06\ \pm 0.04$
- $Scrap\ Costs\ = \$45$

$$Taguchi\ Loss\ Function:$$

$$L(y) = k(y - T)^2$$

$$45\ = k(0.04)^2$$

$$k = 28\ 125$$

$$\therefore L(y) = 28\ 125(y - 0.06)^2$$

6.1.2. Further analysis – Problem 7

a) $Thickness = 0.06\ \pm 0.04$

$Scrap\ Costs\ = \$35$

$$Taguchi\ Loss\ Function:$$

$$L(y) = k(y - T)^2$$

$$35\ = k(0.04)^2$$

$$k = 21\ 875$$

$$\therefore L(y) = 21\ 875(y - 0.06)^2$$

b) $Process\ deviation = 0.027\ cm$

$$Taguchi\ Loss: L(y) = k(y - T)^2$$

$$L\ = 21\ 875(0.027)^2$$

$$\therefore L = 15.95$$

6.2. Reliability of technology manufacturing: Magnaplex

The reliability of another subsidiary, Magnaplex, had to be analysed. The management team wanted to investigate the impact of using identical machines for backup in the manufacturing process. Information regarding the Magnaplex process was gained from problem 27, Chapter 7 in the 11th edition of *Managing for Quality and Performance Excellence* by *J. Evans* and *W.M. Lindsay*.

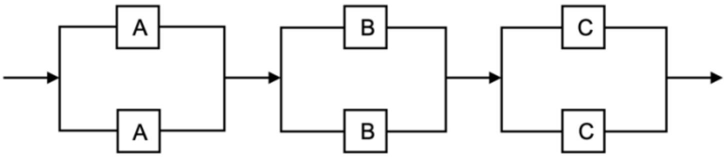


Figure 35: Production system reliability block diagram

Machines	Reliability
A	0.85
B	0.92
C	0.90

Table 9: Machine reliability

a) System reliability when only one machine is working:

$$Reliability = R_A \times R_B \times R_C$$
$$= 0.85 \times 0.92 \times 0.9$$
$$= 0.7038$$

If only one machine is working in each block, the probability that none of the 3 machines will fail is 70.38 %

b) System reliability when both machines is working:
The block of machines will only fail if both machines in a single block fail. The combined reliability is calculated as follow: $Combined\ Reliability\ =\ 1 - P(both\ fail)$.

$$P(both\ fail) = (1 - reliability)^2$$

The probability that both machines won't fail:

$$R_{AA} = 1 - (1 - 0.85)^2 = 0.9775$$
$$R_{BB} = 1 - (1 - 0.92)^2 = 0.9936$$
$$R_{CC} = 1 - (1 - 0.9)^2 = 0.99$$

$$Reliability = R_{AA} \times R_{BB} \times R_{CC}$$
$$= 0.9775 \times 0.9936 \times 0.99$$
$$= 0.9615$$

$$Improvement = \frac{0.9615 - 0.7038}{0.7038} \times 100 = 36.62\%$$

There is an improvement of 36.62% in the reliability of the production process when two machines per block is used instead of one. The probability that the product will go through the production process smoothly is 96.15%

6.3. Reliability of Drivers and Delivery Vehicles

The reliability of delivery vehicles and drivers were determined using the following information that was given: The company has 20 delivery vehicles. Nineteen of these vehicles have to operate to achieve reliable service. Apart from this, there are 21 drivers who work eight hour shifts each day. The amount of days that a certain amount of vehicles and driver were available in the past 1560 days can be found in table 10.

Table 10: Amount of days that a specific amount of vehicles and drivers is available

Vehicles available	Days	Drivers available	Days
20	190	20	95
19	22	19	6
18	3	18	1
17	1	-	-

The number of days that the company will have reliable delivery when they own 20 delivery vehicles was calculated.

$$Reliability(vehicles) = R(v) = 0.9904183$$
$$Reliability(drivers) = R(d) = 0.9979039$$

Reliability of drivers and vehicles combined:

$$Total\ Reliability = R(v) \times R(d) = 0.9883423$$

The expected amount of days with reliable delivery in a year:

$$Reliable\ days = Total\ Reliability \times 365$$
$$= 0.9883423 \times 365 = 360.74\ days$$

The same calculations were done for a situation where the company has 21 vehicles and 21 drivers.

$$Reliability(vehicles) = R(v) = 0.9999836$$
$$Reliability(drivers) = R(d) = 0.9979039$$

Reliability of drivers and vehicles combined:

$$Total\ Reliability = R(v) \times R(d) = 0.9978875$$

The expected amount of days with reliable delivery in a year:

$$Reliable\ days = 0.9978875 \times 365 = 364.23\ days$$

It is clear that the addition of one more vehicle, while keeping the number of drivers constant, improves the days of reliable delivery by 3.48 days. It is now in the hands of the managers to decide whether to invest in a new delivery vehicle, now knowing the effect that it will have on delivery reliability.

Conclusion

The dataset of an online sales business was prepared by removing all the invalid instances from the data, including missing values and negative price values. Descriptive statistics was applied to all the valid instances to determine how the features are distributed. Relationships between the features were identified, for example that household products have the longest delivery time.

Statistical process control was performed on the data and it was found that the delivery process of clothing, gifts, luxury and household products are unstable and out of control. Corrective measures should be taken to improve these processes. The delivery process was further investigated by applying certain rules to the dataset and measuring the performance and errors of the different classes. A recommendation was made that if the average delivery time of technology items could be reduced by 3 hours, the optimal delivery process would be achieved.

MANOVA tests were performed to determine the relationships between different features an in both tests it was clear that the features do influence each other and they are not completely independent. Lastly, it was important to calculate the reliability of the current products and services in this business. Management was left with a decision regarding the investment in more delivery vehicles.

The processes of this online business is not flawless and therefore this analysis should be used to improve the quality of their services and products. Possible problems with the current processes was clearly stated in the report and should not be ignored.

References

Bhandari, P., 2021. *Type I & Type II Errors | Differences, Examples, Visualizations.* [Online]
Available at: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/
[Accessed 7 October 2022].

Chetty, P. & Jain, R., 2021. *Interpreting MANOVA test with more than one dependent variable.* [Online]
Available at: https://www.projectguru.in/interpreting-manova-test-with-more-than-one-dependent-variable/
[Accessed 15 October 2022].

Hessing, T., n.d. *Statistical Process Control (SPC).* [Online]
Available at: https://sixsigmastudyguide.com/statistical-process-control-spc/
[Accessed 1 October 2022].

PQSystems, 2022. *6.3. Interpret Cp and Cpk.* [Online]
Available at:
https://www.pqsystems.com/qualityadvisor/DataAnalysisTools/capability_4.6.3.php#:~:text=The%20Cp%20and%20Cpk%20indices,values%20will%20be%20the%20same.
[Accessed 25 September 2022].