ECSA Graduate Attributes Project

Quality Assurance 344

Due Date: 21 October 2022

Name: Johan Hendrik Olivier **Lecturer: Dr Van Schalkwyk**

Student Number: 23729961

BSc Wood and Wood Product Science

Table of Contents

List of Tables

List of Figures

Introduction

An online business has approached me and they need help analysing their sales data to see if there are some relationships between certain aspects which could help them to improve the business sales and Delivery service.

R studio will be used to analyse the data. R studio is a platform where you can create open-source software for data science, technical communication and scientific research for free (RStudio, 2022).

In this report, the analysed data will be presented together with graphs, plots and tables which will help with further improvements of the company.

Part 1: Data Wrangling

Data wrangling is also known as the process of cleaning up data (Stobierski, 2021). It is the transformation of raw data into more usable data (Stobierski, 2021). Data wrangling must be the first step to ensure that accurate analysis takes place.

The original data

The data file "*salesTable2022.csv*" were provided and used for the data analysis. After inserting the data into R studio it was stored in a data file named "DataSales". The data set contains 180 000 rows and 10 columns.

```
> head(DataSales)
  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
```

Table 1: Head of DataSales

```
> tail(DataSales)
             X     ID AGE      Class     Price Year Month Day Delivery.time  Why.Bought
179995 179995  49178  82       Food    505.88 2024     2  20           2.5     Website
179996 179996  65414  31      Gifts   3147.66 2026     2   1          13.0 Recommended
179997 179997  57864  34      Gifts   1111.36 2023     6   4          10.0 Recommended
179998 179998  48301  77      Gifts   3943.92 2028     4  29          17.0     Website
179999 179999  96502  56     Sweets    243.00 2023     5  26           2.0     Website
180000 180000  71587  53  Household  15362.39 2021     8  22          43.5     Website
```

Table 2: Tail of DataSales

Invalid Data

The next step was data wrangling. In this case, it was removing all the rows which contain "NA" from the data set "*DataSales*" and storing them in a new data set named "*Invalid*". A new column was inserted into the Invalid data sets named "*primary_Invalid*" to rearrange the data from 1 to **m**. There were 17 rows containing NA data of which all came from the "Price" Column.

```
> head(Invalid)
      primary_Invalid     X    ID AGE      Class Price Year Month Day Delivery.time  Why.Bought
12345               1 12345 18973  93      Gifts    NA 2026     6  11          15.5     Website
16321               2 16321 81959  43 Technology    NA 2029     9   6          22.0 Recommended
19541               3 19541 71169  42 Technology    NA 2025     1  19          20.5 Recommended
19999               4 19999 67228  89      Gifts    NA 2026     2   4          15.0 Recommended
23456               5 23456 88622  71       Food    NA 2027     4  18           2.5      Random
34567               6 34567 18748  48   Clothing    NA 2021     4   9           8.0 Recommended
```

Table 3: Head of Invalid data set

```
> tail(Invalid)
      primary_Invalid     X    ID AGE      Class Price Year Month Day Delivery.time  Why.Bought
87654              12  87654 40983  33       Food    NA 2024     8  27           2.0 Recommended
98765              13  98765 64288  25   Clothing    NA 2021     1  24           8.5    Browsing
144444             14 144444 70761  70       Food    NA 2027     9  28           2.5 Recommended
155555             15 155555 33583  56      Gifts    NA 2022    12   9          10.0 Recommended
166666             16 166666 60188  37 Technology    NA 2024    10   9          21.5     Website
177777             17 177777 68698  30       Food    NA 2023     8  14           2.5 Recommended
```

Table 4: Tail of Invalid data set

Valid Data

All the remaining data were stored in a new data set named "Valid". A new column was inserted into the valid data sets named "*primary_valid*" to rearrange the data from 1 to **n**. There were 179 983 rows that contained valid data. This data set was then used for further data analysis.

```
> head(Valid)
  primary_Valid X     ID AGE      Class    Price Year Month Day Delivery.time  Why.Bought
1             1 1  19966  54     Sweets   246.21 2021     7   3           1.5 Recommended
2             2 2  34006  36  Household  1708.21 2026     4   1          58.5     Website
3             3 3  62566  41      Gifts  4050.53 2027     8  10          15.5 Recommended
4             4 4  70731  48 Technology 41843.21 2029    10  22          27.0 Recommended
5             5 5  92178  76  Household 19215.01 2027    11  26          61.5 Recommended
6             6 6  50586  78      Gifts  4929.82 2027     4  24          14.5      Random
```

Table 5: Head of Valid data set

```
> tail(Valid)
       primary_Valid      X     ID AGE      Class    Price Year Month Day Delivery.time  Why.Bought
179995        179978 179995 49178  82       Food   505.88 2024     2  20           2.5     Website
179996        179979 179996 65414  31      Gifts  3147.66 2026     2   1          13.0 Recommended
179997        179980 179997 57864  34      Gifts  1111.36 2023     6   4          10.0 Recommended
179998        179981 179998 48301  77      Gifts  3943.92 2028     4  29          17.0     Website
179999        179982 179999 96502  56     Sweets   243.00 2023     5  26           2.0     Website
180000        179983 180000 71587  53  Household 15362.39 2021     8  22          43.5     Website
```

Table 6: Tail of Valid data set

Part 2: Descriptive Statistics

The “Valid” data set now contains 179 983 rows and 11 Columns. Only 8 columns can be analysed in a descriptive statistic way.

AGE

This is the age of the client who bought the product.

```
> summary(valid$AGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   38.00   53.00   54.57   70.00  108.00
```

Table 7: Summary of the AGE column

There are 90 age classes from the youngest which is 18 years old to the oldest which is 108 years old. The average age of a client who purchases a product is 54.57 years. With a standard deviation of 20.38906 years.

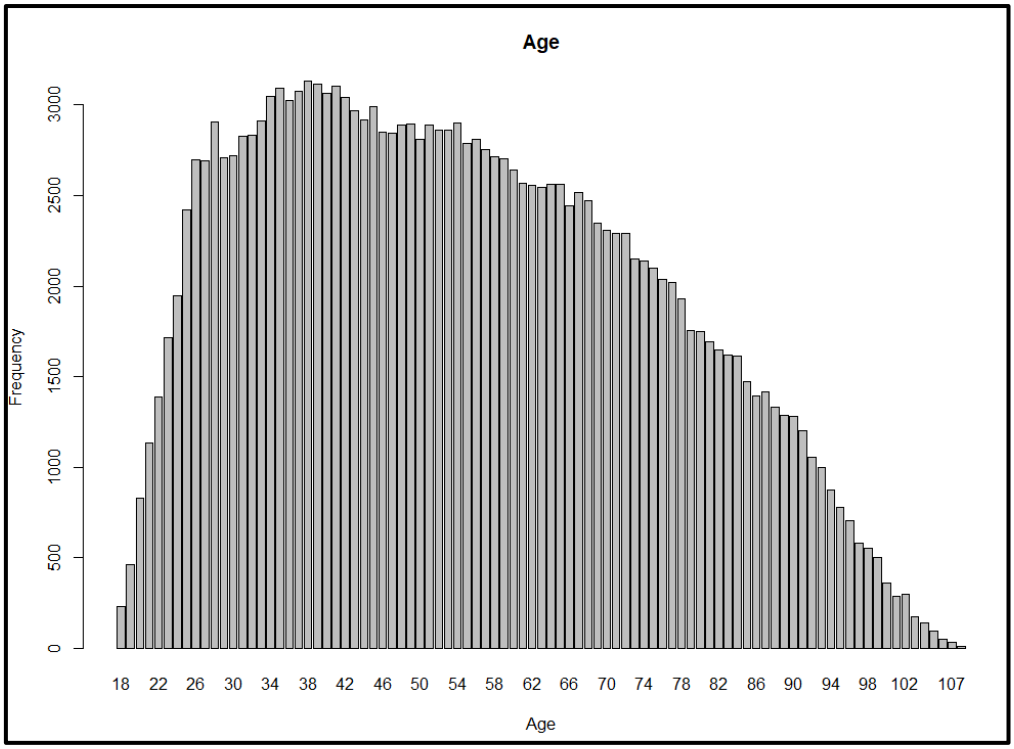

Figure 1: A graph that shows how much each age class appears

When looking at figure 1 one can see that the data is skewed to the left which means that more younger age people buy from the company. The tidyvers package was used to determine the age that appears the most and the least. According to the data, the age group that bought the most products were the age group 38 years and the count was 3130 people. The age group that bought the least amount of products was the oldest age group, age group 108 with a count of 9.

The average amount of products bought per age group was 1977.835 with a standard deviation of 978.0363.

Class

The “Class” column is the category that the product sold falls under.

There are 7 different categories in the class column:

- Clothing
- Food
- Gifts
- Household
- Luxury
- Sweets
- Technology

```
> table(valid$class)

  Clothing       Food      Gifts  Household     Luxury     Sweets Technology
     26403      24583      39149      20067      11869      21565      36347
```

Table 8: Summary of the class column

Table 8 shows how many sales occurred in each Class.

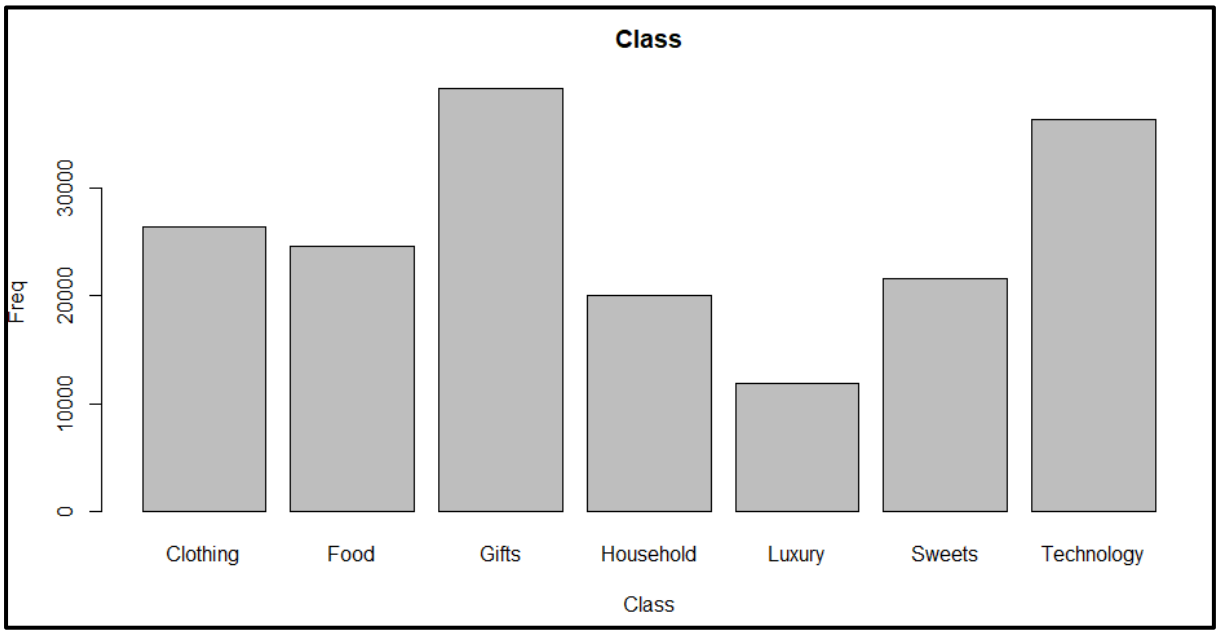

Figure 2: Graph to show how much each class appears

In figure 2 one can see that the most appeared category is Gifts with a count of 39 149 and the least appeared category is Luxury with a count of 11 869. There is an average of 25 711.86 products sold per category with a Standard deviation of 9452.518.

Price

The “Price” column is the price of the product bought.

```
> summary(valid$Price)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-588.8    482.3   2259.6  12293.7  15270.7 116619.0
```

Table 9: Summary of the Price column

The Average price for an item sold is R 12 293.70 with a standard deviation of R 20 888.97. The min price for an item is in the negatives which possible means it was a refund on a product.

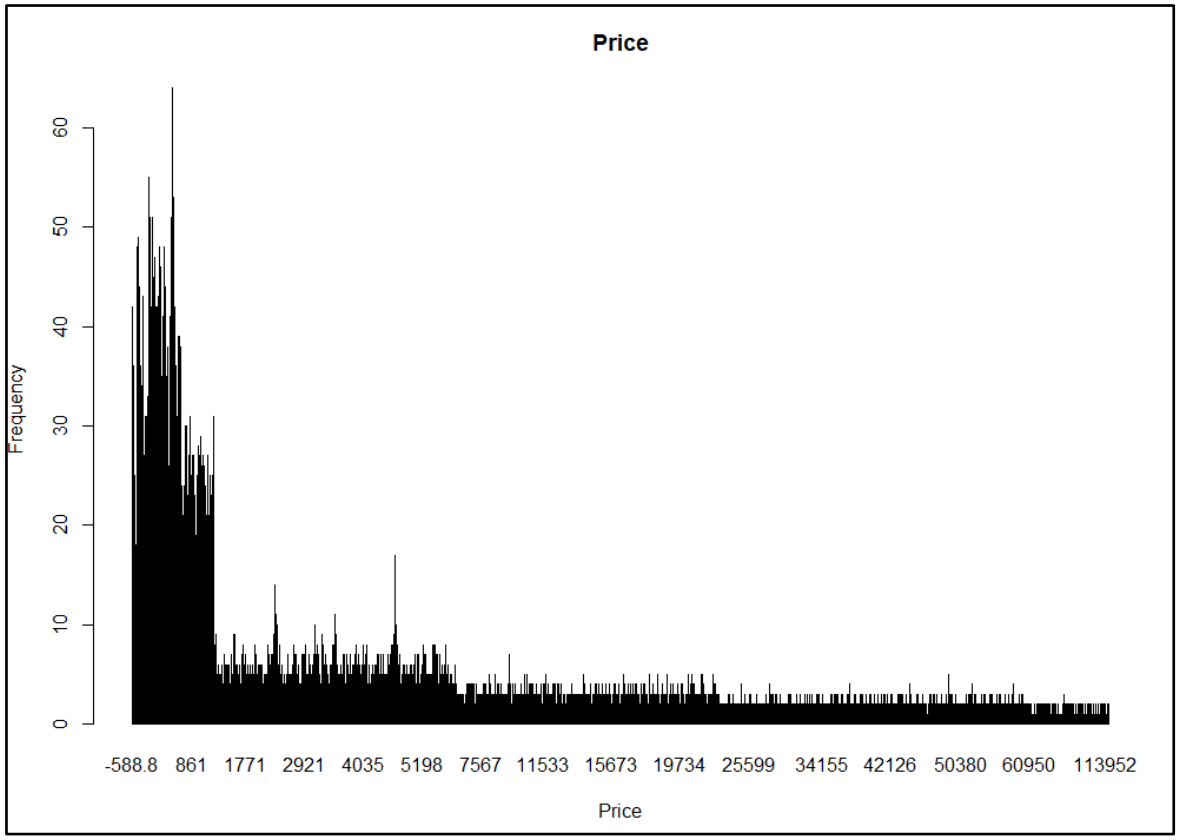

Figure 3: Graph that shows how much each price appears

In figure 3 one can see the price graph is skewed to the left which means that there are more sales in the lower price range. The tidyvers package was used to determine the category in the class column that appears the most and the least. The most appeared price was R 567 with a count of 64 and the least appeared price was R116 619.00 with a count of one this price was also the max price.

The average count of prices in a price range is 2.283092 with a Standard deviation of 4.126521.

Year

The year column is the year when the product was bought.

```
> summary(valid$Year)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2021    2022    2025    2025    2027    2029
```

Table 10: Summary of the Year column

In table 10 one can see that the years when products were bought stretch over 8 years from 2021 to 2029.

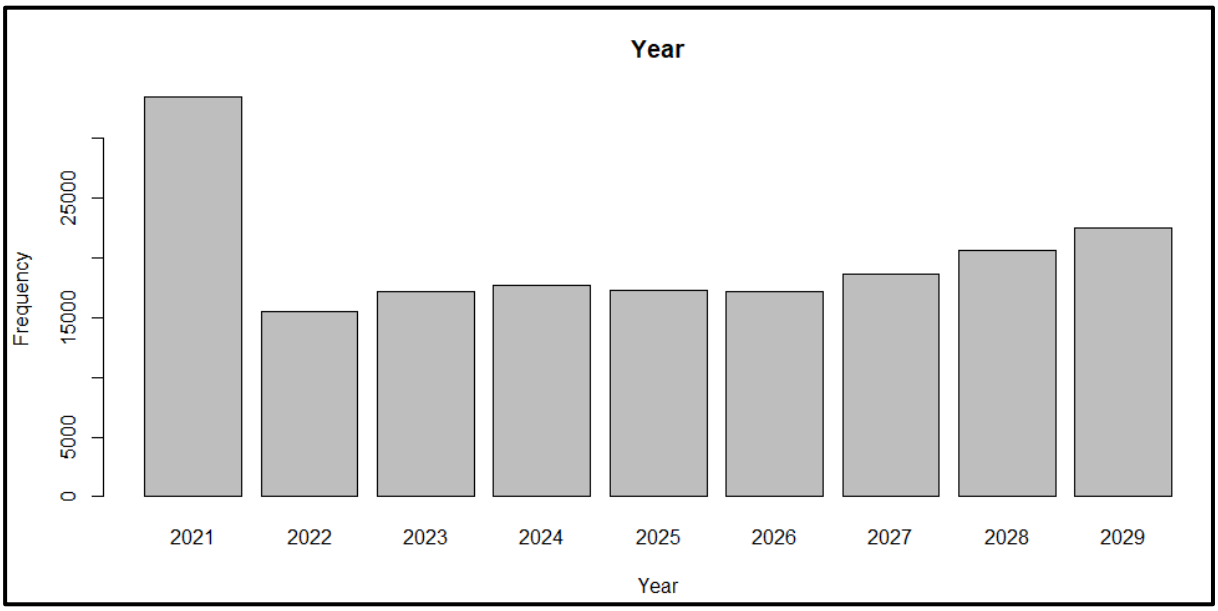

Figure 4: Graph to show how much sales appear in each year

Figure 4 shows that the year where the most sales occurred was 2021 with a count of 33 443 and the year where the least sales occurred was 2022 with a count of 15 547.

The average amount of sales per year is 19 998.11 with a standard deviation of 5453.947

Month

The "Month" column is the month when the products were bought.

```
> summary(valid$Month)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   4.000   7.000   6.521  10.000  12.000
```

Table 11: Summary of the Month column

The months where products were bought stretch from 1 to 12 which is January to December.

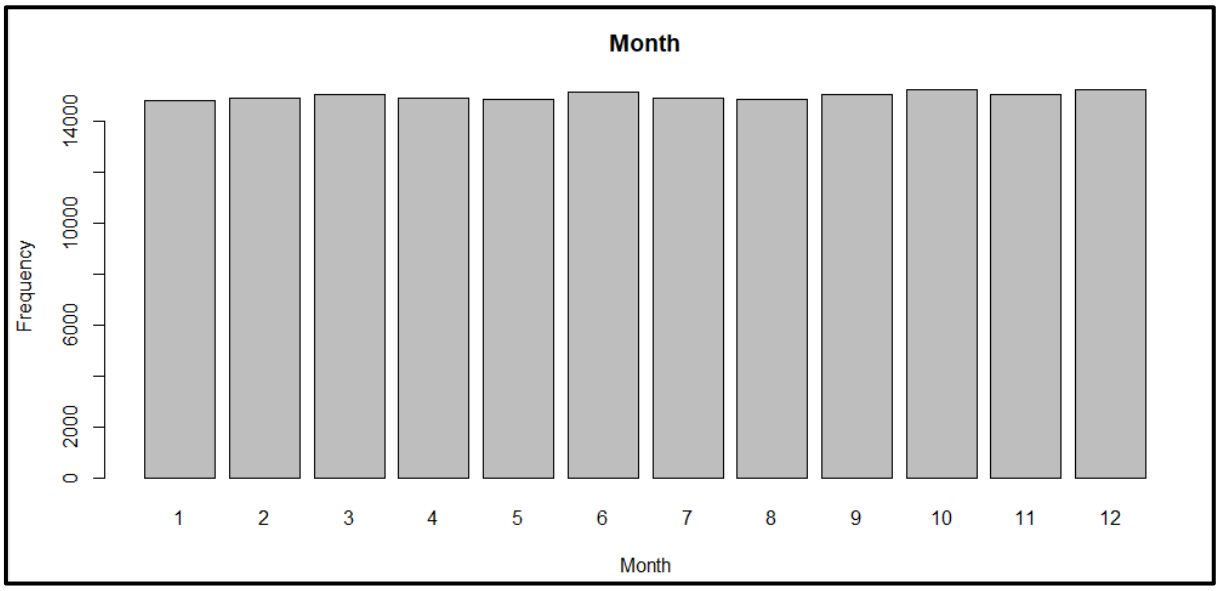

Figure 5: Graph to show how much sales occur in each month

In figure 5 one can see that the sales are very evenly spread over the 12 months but the most sales occurred in Month 12 (December) with a count of 15 226 and the least amount of sales occurred in Month 1 (January) with a count of 14 799. In December a lot of people buy gifts thus it makes sense that December have the highest sales and in January it is after the holiday thus people tend to spend less money.

The average amount of sales in a month is 14 998.58 with a standard deviation of 4.126521

Day

The "Day" column is the day in a month on which a product was bought.

```
> summary(valid$Day)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    8.00   16.00   15.54   23.00   30.00
```

Table 12: Summary of the column Day

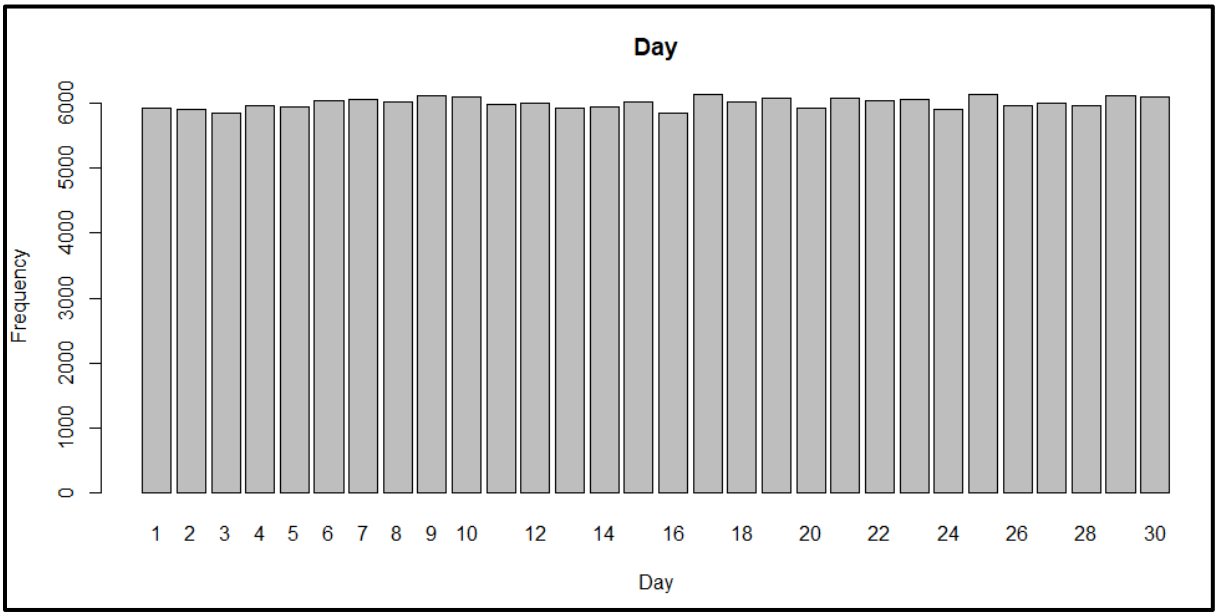

Figure 6: Graph that shows the amount of sales that occur on a given day in a month

In figure 6 one can see once again, same as the "Month" column, the average sales per day are spread very evenly. The 17th day of the month is the day on which the most sales occurred with a count of 6126 and the day on which the least sales occurred was the 16th day of the month with a count of 5839.

The average sales on a given day in a month are 5999.433 with a standard deviation of 83.11549.

Delivery time

The "Delivery.time" column is the amount of time it took to deliver a product in hours.

```
> summary(valid$Delivery.time)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.5     3.0    10.0    14.5    18.5    75.0
```

Table 13: Summary of the Delivery.time column

In table 13 one can see the average time it took for a delivery of a product to arrive was 14.5 hours with a standard deviation of 13.95608. The range stretched from 0.5 hours to 75 hours for a delivery.

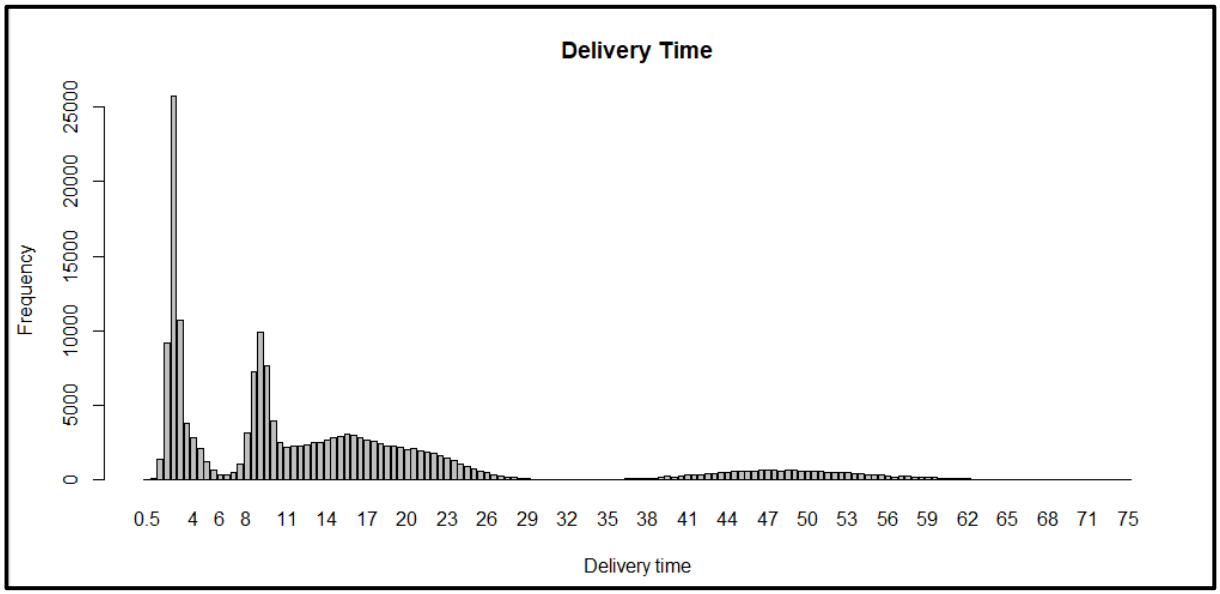

Figure 7: Graph that shows how much each delivery time appears

Figure 7 is skewed to the left which means the delivery time tends to be fast. When looking at the graph one can see that most deliveries happen within a day, 24 hours from the placement of the order. The Tidyverse package was used to determine the delivery time that appears the most as well as the delivery time that appears the least. The most appeared delivery time is 2.5 hours with a count of 25724 and the least appeared delivery time is 75 hours with a count of one.

The average sales per delivery time are 1216.101 with a standard deviation of 2702.559.

Why Bought

The "Why.Bought" column is the reason why the item was bought by the client.

There are 6 different categories in the "Why.Bought" column:

- Browsing
- Email
- Random
- Recommended
- Spam
- Website

```
> table(valid$Why.Bought)

   Browsing       EMail      Random Recommended        Spam     Website
      18994        7225       13121      106988        4208       29447
```

Table 14: Summary of the Why.bought column

Table 14 shows how many sales occurred in each why bought category.

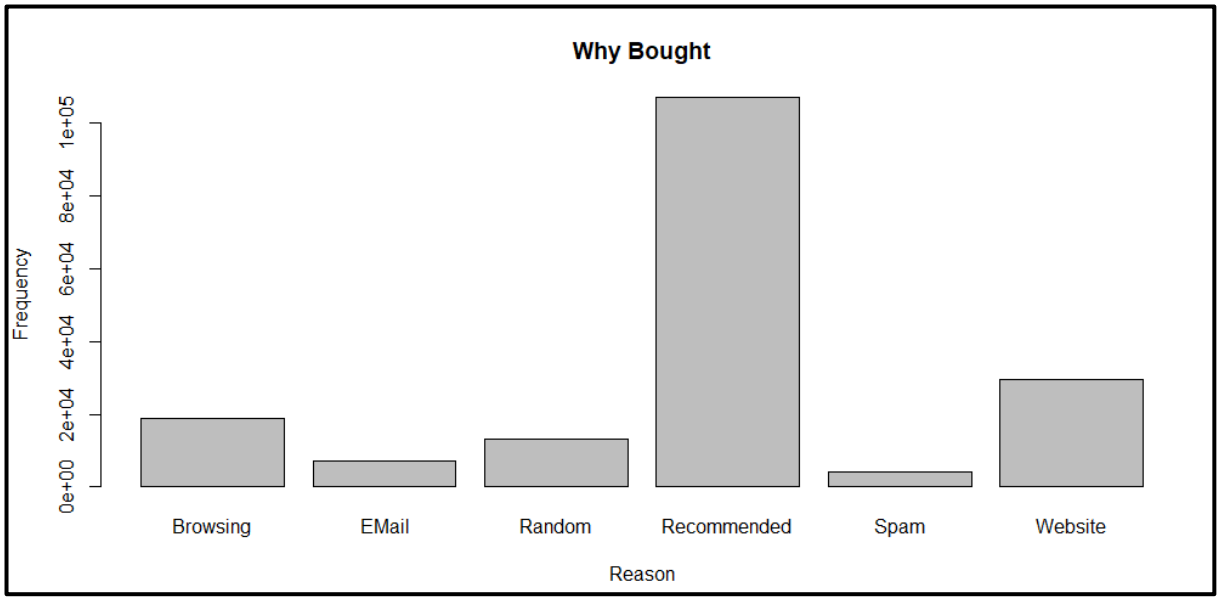

Figure 8: Graph that shows the sales per category in the Why.bought column

In figure 8 one can see that the "Recommended" category is the most with a count of 106 988 sales and the category with the least amount of sales is "Spam" with a count of 4208.

The average sales per category in the "Why. Bought" column is 29 997.17 with a standard deviation of 38 774.78.

Process capability

The Upper Specification Limit (USL) was given as 24 and the Lower specification limit (LSL) was given as 0. These limits are the target ± the accepted deviation. The USL is 24 because they are 24 hours a day available for delivery and the LSL is 0 because it is the min time deliveries can take.

The process capabilities of the delivery time from the Technology class were calculated.

The delivery times of the technology class have an average of 20.01095 and a standard deviation of 3.501993.

The formulas used to calculate the Process capabilities are as follows:

Capability potential: $C_P = (USL - LSL)/6\sigma$

Capability potential refers to the data spread and width of the data range (Chitranshi, 2022).

Capability based on Upper limit: $C_{PU} = (USL - \mu)/3\sigma$

The Capability potential focuses on the Upper specification limit of the data spread (Chitranshi, 2022).

Capability based on lower limit: $C_{PL} = (\mu - LSL)/3\sigma$

The Capability potential focuses on the lower specification limit of the data spread (Chitranshi, 2022).

Capability performance: $C_{PK} = \min(C_{PL}, C_{PU})$

The Cpk also give the process capability but refers to the data points around and near the mean and how close the process mean is to the target value (Chitranshi, 2022).

```
#-------------Process Capability-------------#

process_capa <- subset(valid,valid$class == "Technology")
process_capa_del_time <- process_capa$Delivery.time

sd_process_capa_del_time <- sd(process_capa_del_time)
mean_process_capa_del_time <- mean(process_capa_del_time)
summary(process_capa_del_time)
mean_process_capa_del_time
sd_process_capa_del_time

USL <- 24 #given
LSL <- 0  #given

Cp <- (USL - LSL)/(6*sd_process_capa_del_time)
Cp
Cpu <- (USL - mean_process_capa_del_time)/(3*sd_process_capa_del_time)
Cpu
Cpl <- (mean_process_capa_del_time - LSL)/(3*sd_process_capa_del_time)
Cpl
Cpk <- min(Cpl,Cpu)
Cpk
```

Figure 9: Code for calculating Process Capability

Cp	Cpu	Cpl	Cpk
1.142207	0.3796933	1.90472	0.3796933

Table 15: Process Capability answers

In table 15 one can see the process is capable because the Cp is larger than one but only by 0.142207 thus it is barely capable. The company is not completely capable of delivering products in the time specified when it's not perfectly centred. The standard deviation of the delivery time of the technology category should be made smaller to make the process more efficient.

The Cpk is 0.7625135 smaller than the Cp thus the process is not centred. The process should be moved to fit the target better or one should reduce the variation and spread of the data.

Part 3: Statistical process control

X&s-charts are used to evaluate the process mean and the process standard deviation over a period of time respectively (Hesing, 2022). These charts can be used when there are large sample sizes and it helps to understand the spread of the data (Hesing, 2022).

Re-ordering of data

The "Valid" data set was ordered chronologically by year then month then day. This re-ordering made the oldest valid data appear first. If there were rows that had the same values, it was ordered by the "primary_Valid" column. This data was stored in a new data set called "order_Data".

Initialising the X&s-Chart

From this re-ordered data, subsets of every category under the class column were put into samples of 15 and stored in a new data set. For the control chart, only 30 of these samples were used and that was a total of 450 values.

Control Limits

The following control limits were generated for the X&s-charts evaluation of the Delivery time column.

- Upper Control Limit: Line which indicates if processes are out of control on the upper limit.
- U2Sigma
- U1Sigma

  } Devides the space between the centre line and Upper control limit into 3 equal parts.
- Centre line: Process Average line (The average of the 30 samples)
- L1Sigma
- L2Sigma

  } Devides the space between the centre line and Upper control limit into 3 equal parts.
- Lower control Limit: Line which indicates if processes are out of control on the lower limit.

X-Chart

Control limits were generated for the samples created.

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	23.03203	22.14617	21.26031	20.37444	19.48858	18.60272	17.71686
Clothing	9.41003	9.26335	9.11668	8.97000	8.82332	8.67665	8.52997
Household	50.32763	49.07250	47.81736	46.56222	45.30709	44.05195	42.79681
Luxury	5.50692	5.24980	4.99268	4.73556	4.47843	4.22131	3.96419
Food	2.71405	2.63936	2.56468	2.49000	2.41532	2.34064	2.26595
Gifts	9.50748	9.12536	8.74323	8.36111	7.97899	7.59687	7.21474
Sweets	2.90551	2.76293	2.62035	2.47778	2.33520	2.19263	2.05005

Table 16: Control limits for the X-Chart of the delivery time column

Process control graphs for X-Chart

Technology

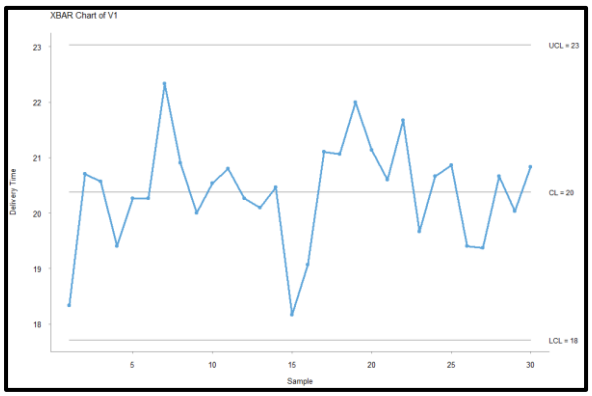

Figure 10: X-Chart for Technology

In figure 10 all the samples are between the Upper Control Limit and Lower Contol Limit and there is also no Out-of-Control Signal thus samples don't have to be removed and the Centre Line and Control Limits do not need to be re-calculated. The sample range can be evaluated accurately.

Clothing

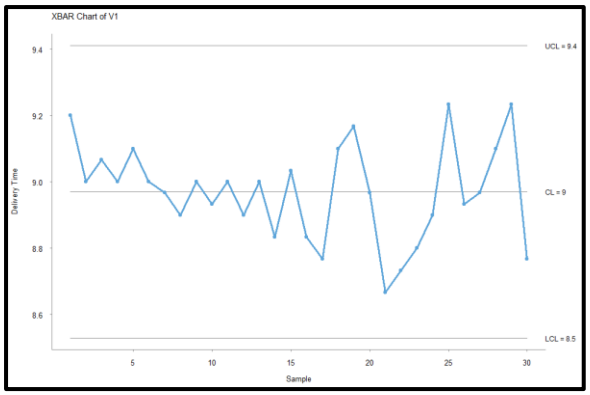

Figure 11: X-Chart for Clothing

In Figure 11 all the samples are between the Upper Control Limit and Lower Contol Limit and there is also no Out-of-Control Signal thus samples don't have to be removed and the Centre Line and Control Limits do not need to be re-calculated. The range can be evaluated accurately.

Household

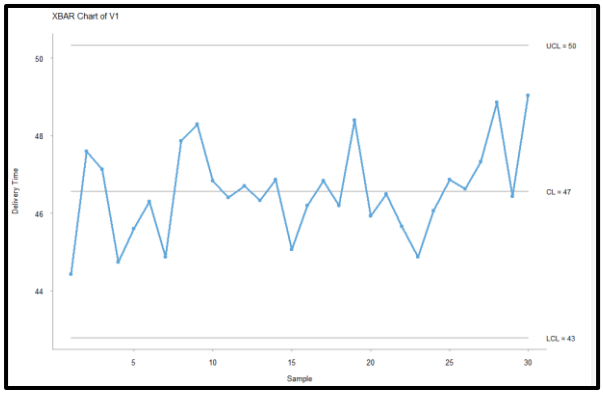

Figure 12: X-Chart for Household

In Figure 12 all the samples are between the Upper Control Limit and Lower Contol Limit and there is also no Out-of-Control Signal thus samples don't have to be removed and the Centre Line and Control Limits do not need to be re-calculated. The range can be evaluated accurately.

Luxury

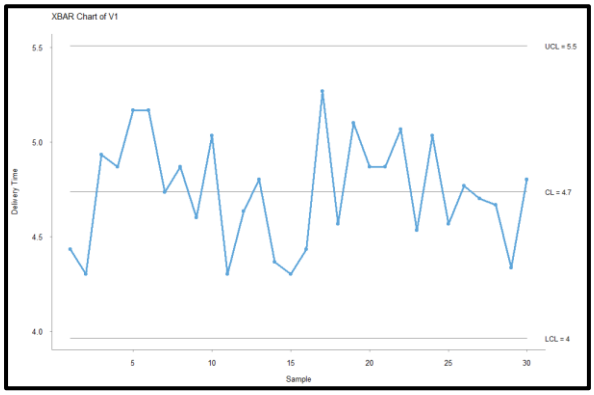

Figure 13: X-Chart for Luxury

In figure 13 all the samples are between the Upper Control Limit and Lower Contol Limit and there is also no Out-of-Control Signal thus samples don't have to be removed and the Centre Line and Control Limits do not need to be re-calculated. The range can be evaluated accurately.

Food

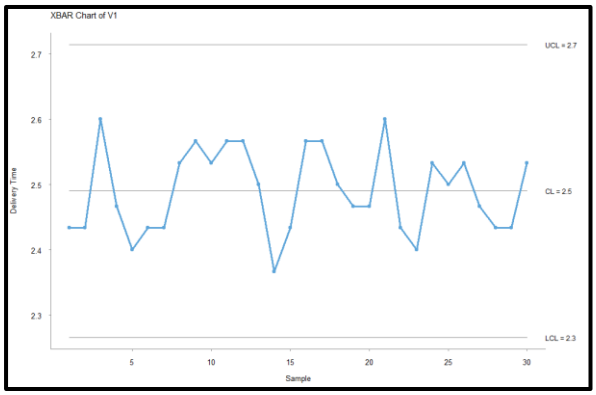

Figure 14: X-Chart for Food

In figure 14 all the samples are between the Upper Control Limit and Lower Contol Limit and there is also no Out-of-Control Signal thus samples don't have to be removed and the Centre Line and Control Limits do not need to be re-calculated. The range can be evaluated accurately.

Gifts

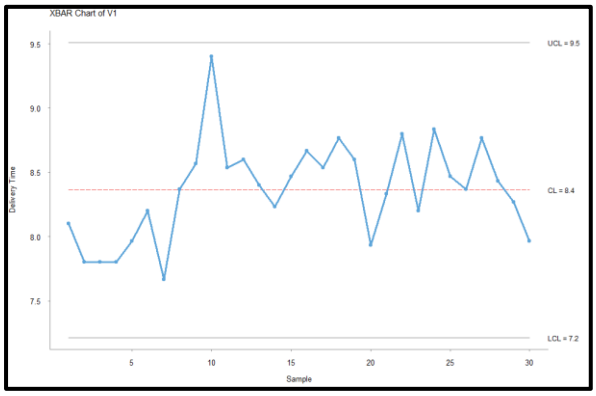

Figure 15: X-Chart for Gifts

Figure 15 shows that there is an Out-of-Control signal. There are 7 consecutive points below the centre line thus the process is also out of control thus some samples need to be removed and the Centre Line and Control Limits need to be re-calculated.

Sweets

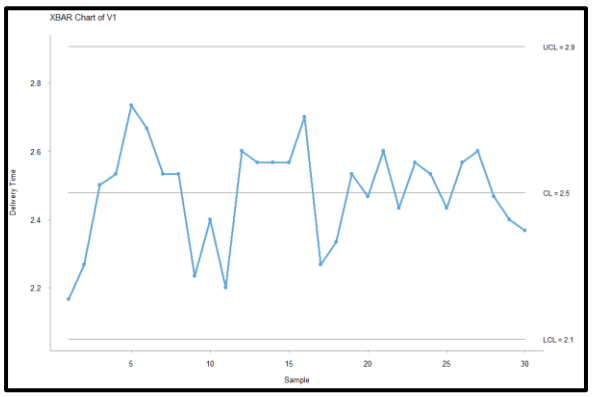

Figure 16: X-Chart for Sweets

In figure 16 all the samples are between the Upper Control Limit and Lower Contol Limit and there is also no Out-of-Control Signal thus samples don't have to be removed and the Centre Line and Control Limits do not need to be re-calculated. The range can be evaluated accurately.

S-Chart

Control limits were generated for the samples created.

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.29496	4.65274	4.01051	3.36829	2.72607	2.08385	1.44163
Clothing	0.87671	0.77037	0.66404	0.55770	0.45137	0.34503	0.23870
Household	7.50219	6.59225	5.68232	4.77238	3.86245	2.95251	2.04258
Luxury	1.53686	1.35046	1.16405	0.97765	0.79124	0.60484	0.41843
Food	0.44639	0.39225	0.33811	0.28396	0.22982	0.17568	0.12154
Gifts	2.28402	2.00699	1.72997	1.45294	1.17591	0.89888	0.62186
Sweets	0.85221	0.74884	0.64548	0.54212	0.43875	0.33539	0.23203

Table 17:: Control limits for the S-Chart of the delivery time column

Process control graphs for S-Chart

Technology

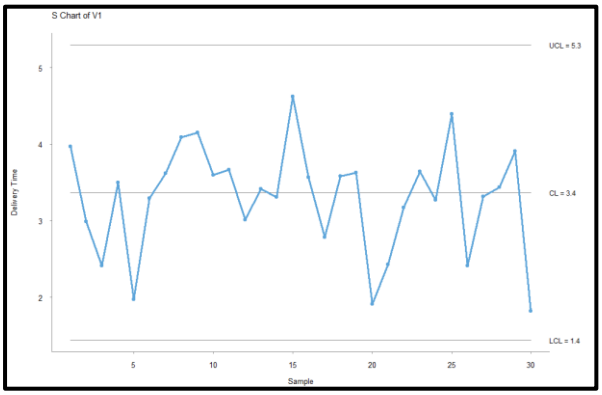

Figure 17: S-Chart for Technology

In figure 17 all the samples are between the Upper Control Limit and Lower Contol Limit and there is also no Out-of-Control Signal thus samples don't have to be removed and the Centre Line and Control Limits do not need to be re-calculated. The range can be evaluated accurately.

Clothing

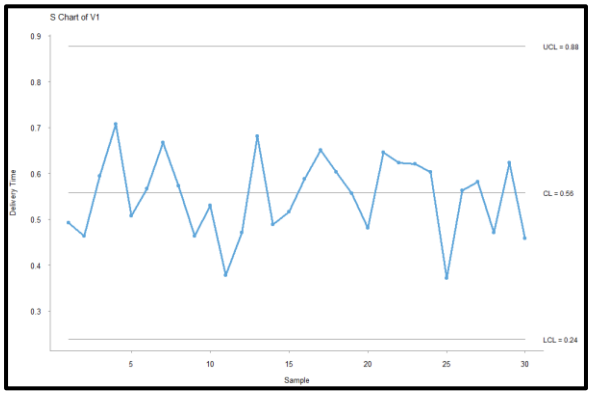

Figure 18: S-Chart for Clothing

In figure 18 all the samples are between the Upper Control Limit and Lower Contol Limit and there is also no Out-of-Control Signal thus samples don't have to be removed and the Centre Line and Control Limits do not need to be re-calculated. The range can be evaluated accurately.

Household

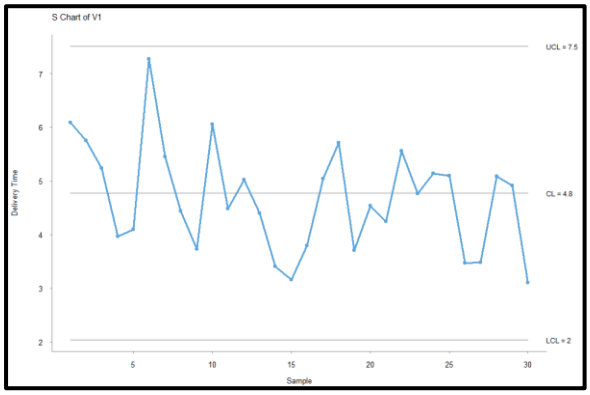

Figure 19: S-Chart for Household

In figure 19 all the samples are between the Upper Control Limit and Lower Contol Limit and there is also no Out-of-Control Signal thus samples don't have to be removed and the Centre Line and Control Limits do not need to be re-calculated. The range can be evaluated accurately.

Luxury

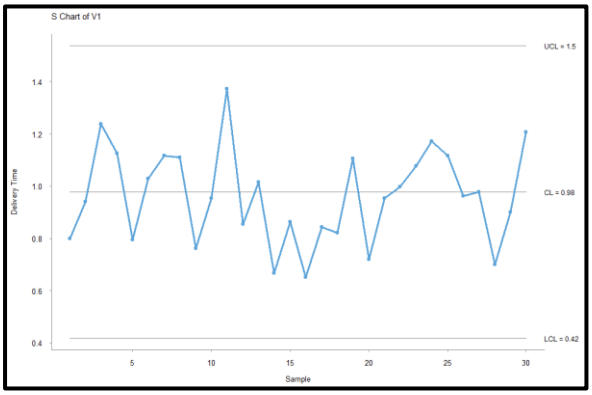

Figure 20: S-Chart for Luxury

In figure 20 all the samples are between the Upper Control Limit and Lower Contol Limit and there is also no Out-of-Control Signal thus samples don't have to be removed and the Centre Line and Control Limits do not need to be re-calculated. The range can be evaluated accurately.

Food

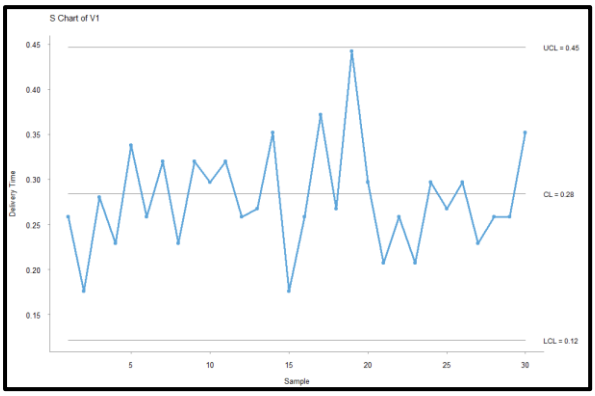

Figure 21: S-Chart for Food

In figure 21 all the samples are between the Upper Control Limit and Lower Contol Limit and there is also no Out-of-Control Signal thus samples don't have to be removed and the Centre Line and Control Limits do not need to be re-calculated. The range can be evaluated accurately.

Gifts

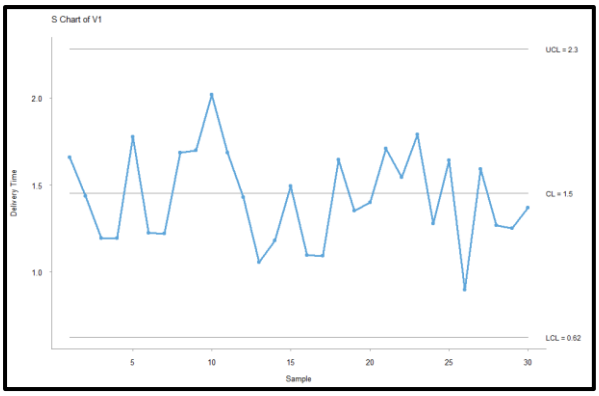

Figure 22: S-Chart for Gifts

In figure 22 all the samples are between the Upper Control Limit and Lower Contol Limit and there is also no Out-of-Control Signal thus samples don't have to be removed and the Centre Line and Control Limits do not need to be re-calculated. The range can be evaluated accurately.

Sweets

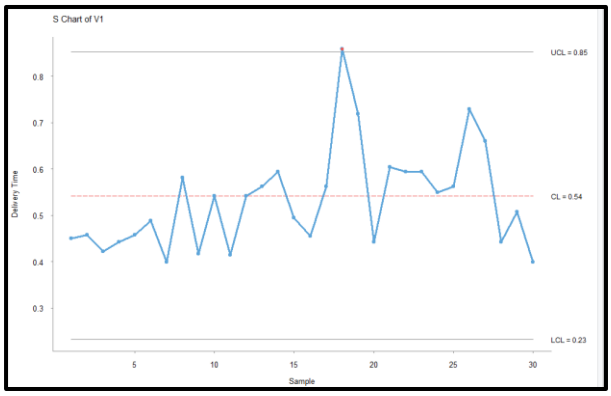

Figure 23: S-Chart for Sweets

Figure 23 shows that there is an Out-of-Control signal. There are 7 consecutive points below the centre line and also one sample above the Upper Control Limit thus the process is out of control and some samples need to be removed and the Centre Line and Control Limits need to be re-calculated.

Part 4: Optimising the delivery processes

In part 4 a new set of samples were created and used to evaluate the process control for the X&s-Charts. All the data from the delivery times for each class was put into Samples of 15. Samples **1:EndOfSamples** was evaluated for each class.

4.1 A) Sample means outside of the outer control limit

Class	# Identified	1st	2nd	3rd	3rd Last	2nd Last	Last
Technology	5	1122	1501	1858	1858	1872	2009
Clothing	4	1222	1525	1677	1525	1677	1724
Household	267	43	48	50	1335	1336	1337
Luxury	127	3	4	5	784	785	790
Food	3	75	432	633	75	432	633
Gifts	1939	2	3	4	2607	2608	2609
Sweets	4	1104	1243	1294	1243	1294	1403

Table 18: Sample means outside the control limits

In table 18 one can see that the classes “Technology”, “Clothing”, “Food”, and “Gifts” have very few samples outside the control limits and could thus be viewed as under control. This means there are not many samples which deviate from the mean delivery time for these classes. The Classes “HouseHold”, “Luxury”, and “Gifts” have a lot of samples outside the control limits and are thus not under control. This mean there are a lot of samples that deviate from the mean delivery time and thus the delivery process is not very reliable and constant. The First 3 samples of the Classes “HouseHold”, “Luxury”, and “Gifts” are grouped close to each other and the last 3 samples are also grouped close to each other. This could mean that there are certain times, could be days, months or years on end when the process is out of control and deviates from the mean.

4.1 B) The most consecutive samples between 0.4 and -0.3 sigma control limits

Class	Most consecutive samples	Ending sample number of the most consecutive samples
Technology	23	612
Clothing	28	308
Household	28	451
Luxury	20	32 & 477
Food	20	927 & 1254
Gifts	17	1063
Sweets	19	325

Table 19: Most consecutive samples between 0.4 and -0.3 Sigma control limits

4.2 Likelihood to make a Type I error

A Type I error is also called a producer's risk or manufacturer's risk. This is when one investigates the process because it gave signals that the process is out of control but the process is stable.

H_0: "The process is in control and centred on the centreline calculated using the first 30 samples"

H_1: "The process is not in control and has moved from the centreline or has increased or decreased in variation"

For A)

The control limit is 3 Sigma

There is an upper and lower control limit thus the function should be multiplied by 2

$$p(Type\ I\ error) = \big(1 - pnorm(3)\big) * 2$$

$$= \ 0.2699796\%$$

There is only a 0.2699796% chance to classify a sample as outside the upper and lower control limits when it is in fact in control.

For B)

The control limits of 0.4 sigma and -0.3 sigma were used to determine the likelihood to make a Type I error.

$$p(Type\ I\ error) = \big(pnorm(0.4)\big) - (pnorm(-0.3))$$

$$= \ 27.33332\%$$

There is a 27.33332% chance to classify a sample as outside these control limits of 0.4 sigma and -0.3 sigma when it is in fact in control.

4.3 Optimising delivery time for Technology

If the delivery time of an item ordered from the technology column is slower than 26 hours the company will lose R 329 per item late per hour. It only costs R 2.5 per item per hour to reduce the average delivery time by an hour.

Figure 24: Lowest possible cost for Technology class

The average delivery time for an order in the technology class is 20.01095 hours. When looking at figure 24, one can see that the delivery time should be decreased by 3 hours to optimise it. This means the average delivery time for an order in the technology column should be around 17 hours. An Average delivery time of 17 hours will minimise the cost to R 340 870 from the initial cost of R 758 674.

4.4 Estimate the likelihood of making a Type II Error

A Type II error is also called a Consumer's risk. A type II error is when the process is unstable (outside the upper and lower control limits) but there is no indication that it is unstable and one thinks it is stable. All the samples were used for this calculation.

UCL for Technology = 22.77679

LCL for Technology = 17.24511

Standard deviation (σ) = 0.9219219

Mean (μ)= 23

$$p(Type\ II\ error) = pnorm(UCL, \mu\ , \sigma) - pnorm(LCL, \pi\ , \sigma)$$

$$= 40.435\%$$

Part 5: DOE and MANOVA

Two columns from the data set namely “Class” and “Why.Bought” was used to do a MANOVA. The columns that were used for the analysis were “AGE”, “Price” and “Delivery.time.”

Class

H_0 : The Information in the Column “Class” is not of significants.

H_1 : The Information in the Column “Class” is of significants.

```
> summary.aov(MANOVA1)
 Response Delivery.time :
               Df   Sum Sq Mean Sq F value    Pr(>F)
Class           6 33461034 5576839  629489 < 2.2e-16 ***
Residuals  179976  1594464       9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Price :
               Df     Sum Sq    Mean Sq F value    Pr(>F)
Class           6 5.7165e+13 9.5275e+12   80238 < 2.2e-16 ***
Residuals  179976 2.1370e+13 1.1874e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response AGE :
               Df   Sum Sq Mean Sq F value    Pr(>F)
Class           6  8423110 1403852  3805.2 < 2.2e-16 ***
Residuals  179976 66397874     369
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(MANOVA1)
               Df Pillai approx F num Df den Df    Pr(>F)
Class           6 1.7577    42438     18 539928 < 2.2e-16 ***
Residuals  179976
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 20: Summary of MANOVA for the Class column

In Statistical software $2.2e^{-16}$ is used to define a very small number. The Pr(>F) value in table 20 which says < $2.2e^{-16}$ means there is a significant result. This means the p-value is even smaller than $2.2e^{-16}$ and thus p < 0,05 which is the typical threshold (Morphist, 2019). This means that the Class Column does matter and we reject the H_0.

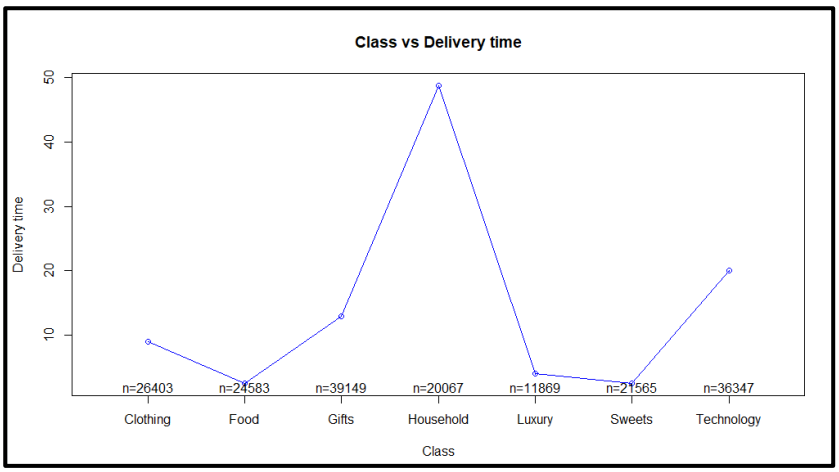

Figure 25: Graph of the mean Delivery time for each Class

In figure 25 one can see there is a correlation between the Delivery time and some categories in the “Class” Column. The mean delivery time for the household class is by far the highest, this could be because the household items are large in size, for example, furniture, and they will take a long time to pack and deliver it. On the other hand, sweets and food have the lowest mean delivery time because it is easy to pack and deliver.

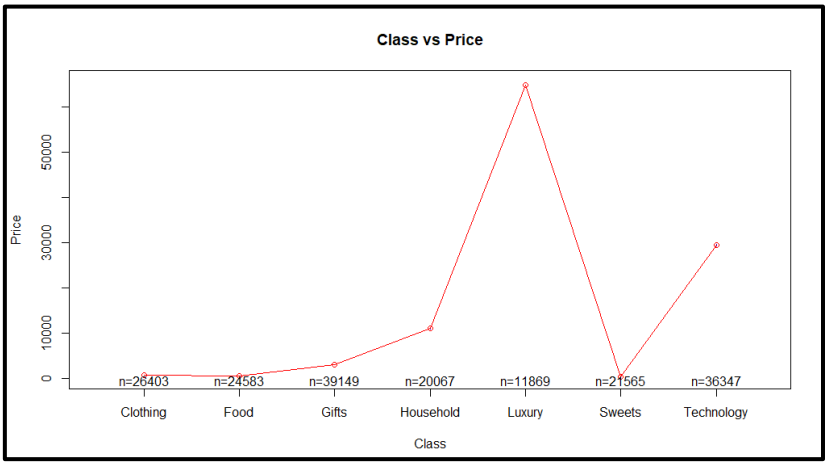

Figure 26: Graph of the mean Price for each Class

In figure 26 one can see there is a correlation between the Price and some categories in the "Class" Column. The mean price of the Class Luxury is the highest, which makes sense because luxury items tend to be more expensive. The second highest Price mean is the Class Technology. Technology also tends to be expensive. The classes with the lowest mean price are Clothing and food. These items tend to be inexpensive compared to the other classes as your average prices in these classes will be below R 1000.

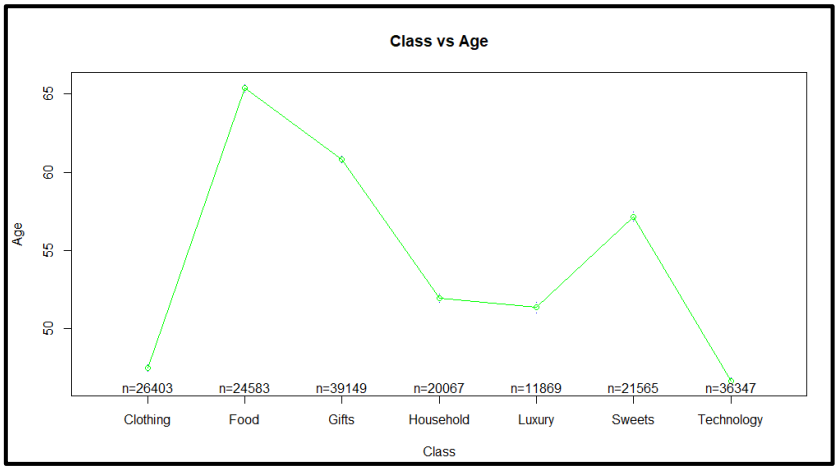

Figure 27: Graph of the mean Age for each class

In figure 27 one can see there is a correlation between Age and some categories in the "Class" Column. The Food class have the highest mean Age, this could be due to the older population that doesn't want to go out and do physical shopping as it is easier to order online and have it delivered. The classes with the lowest mean age are Technology and Clothing. It is usually the younger population that buys Technology items and gadgets. It is usually the younger population that also buys clothing online as the older population would rather go into the store to fit clothing before buying.

Why Bought

H_0 : The Information in the Column "Why.Bought" is not of significants.

H_1 : The Information in the Column "Why.Bought" is of significants.

```
> summary.aov(MANOVA2)
 Response Delivery.time :
                Df   Sum Sq Mean Sq F value    Pr(>F)
Why.Bought       5   783297  156659  822.68 < 2.2e-16 ***
Residuals   179977 34272201     190
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response Price :
                Df     Sum Sq    Mean Sq F value    Pr(>F)
Why.Bought       5 1.5744e+12 3.1487e+11  736.35 < 2.2e-16 ***
Residuals   179977 7.6961e+13 4.2761e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response AGE :
                Df   Sum Sq Mean Sq F value    Pr(>F)
Why.Bought       5   106800 21360.0  51.453 < 2.2e-16 ***
Residuals   179977 74714183   415.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(MANOVA2)
               Df   Pillai approx F num Df den Df    Pr(>F)
Why.Bought      5 0.044145   537.59     15 539931 < 2.2e-16 ***
Residuals  179977
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Table 21 : Summary of MANOVA for Why Bought Column

In Statistical software $2.2e^{-16}$ is used to define a very small number. The Pr(>F) value in table 21 which says < $2.2e^{-16}$ means there is a significant result. This means the p-value is even smaller than $2.2e^{-16}$ and thus p < 0,05 which is the typical threshold (Morphist, 2019). This means that the "Why.Bought" Column does matter and we reject the H_0.

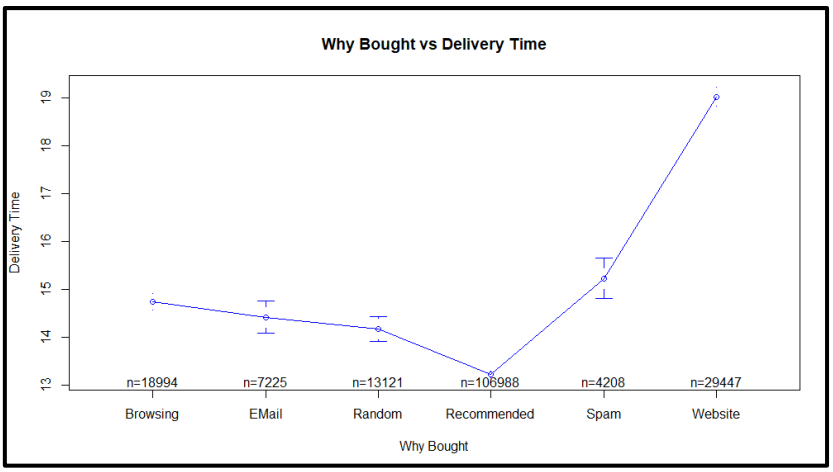

Figure 28: Graph of mean Delivery time for each reason in Why bought Column

In figure 28 one can see there is a correlation between the Delivery time and some reasons in the "Why.Bought" Column. The Highest mean delivery time is the why bought reason Website. It can be that most people that use the website buy household items and thus have a very high mean delivery time. The fastest mean delivery time is the recommended reason in the Why Bought column. It may be that people share the news when there are specials on certain foods because food also has the fastest delivery time.

Figure 29 : Graph of the mean Price for each reason in Why bought Column

In figure 29 one can see there is a correlation between the "Price" and some reasons in the "Why.Bought" Column. The mean highest price comes out of the Browsing column. It can be that most people that buy items when browsing buy luxury items because the luxury class also have the highest mean price. The lowest mean price is the Random reason in the Why Bought column. It may be that people don't know what the random column means and then just choose another column.

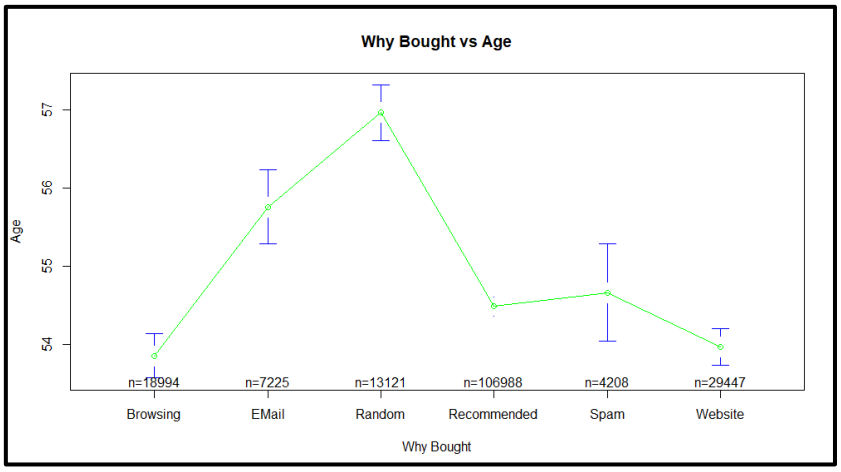

Figure 30 : Graph of the mean Age for each Reason in Why bought Column

In figure 30 one can see there is a correlation between Age and some reasons in the "Why.Bought" Column. The highest mean age is the Random reason in the Why bought Column. It may be that the older population does not want to waste time choosing a reason and just choose Random. All the means are very close to each other and only range from around the age of 54 to 57.

Part 6: Reliability of the service and products.

6.1 **Problem 6:** A blueprint specification for the thickness of a refrigerator part at Cool Food, Inc. is 0.06 ± 0.04 centimetres (cm). It costs $45 to scrap a part that is outside the specifications. Determine the Taguchi loss function for this situation.

Taguchi loss Function

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

$$k = \frac{45}{(0.04)^2}$$

$$= 28125$$

$$L(y) = 28125(y - 0{,}06)^2$$

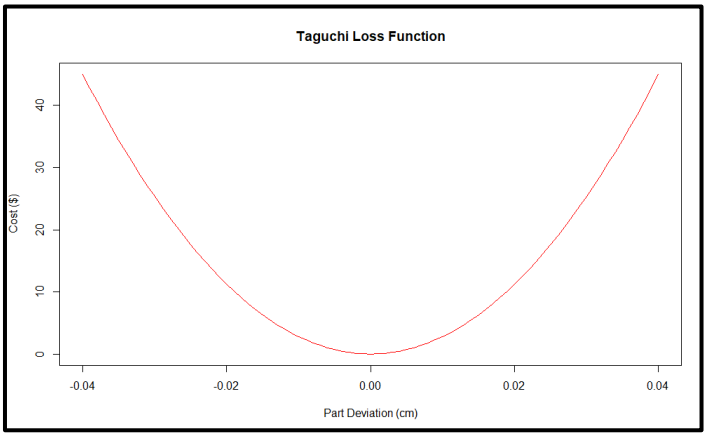

Figure 31 : Taguchi loss function graph for Problem 6

Figure 31 shows the Taguchi loss function graph for problem 6.

Problem 7: A team was formed to study the refrigerator part at Cool Food, Inc. described in Problem 6. While continuing to work to find the root cause of the scrap, they found a way to reduce the scrap cost to $35 per part.

a. Determine the Taguchi loss function for this situation.

Taguchi loss Function

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

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

$$= 21875$$

$$L(y) = 21875(y - 0{,}06)^2$$

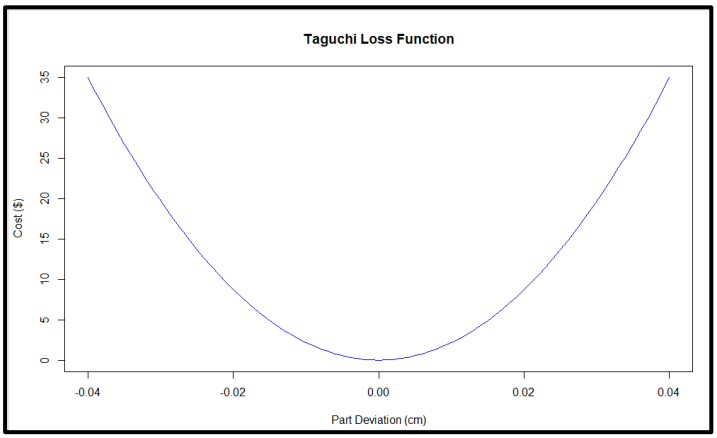

Figure 32: Taguchi Loss Function graph for Problem 7a

In figure 32 one can see when the scrap cost gets reduced it reduces the Taguchi loss function a lot which is a good thing because it will decrease your overall losses.

b. If the process deviation from the target can be reduced to 0.027 cm, what is the Taguchi loss?

$$Target = (y - m)$$

$$Target = 0.027$$

$$Taguchi\ Loss = 21875(0.027)^2$$

$$Taguchi\ Loss = 15.95\ \$\ per\ part$$

When inserting the target value into the Loss function one can see that there is a loss of 15.95$ per part that need to be scraped.

6.2 Problem 27: Magnaplex, Inc. has a complex manufacturing process, with three operations that are performed in series. Because of the nature of the process, machines frequently fall out of adjustment and must be repaired. To keep the system going, two identical machines are used at each stage; thus, if one fails, the other can be used while the first is repaired (see accompanying figure).

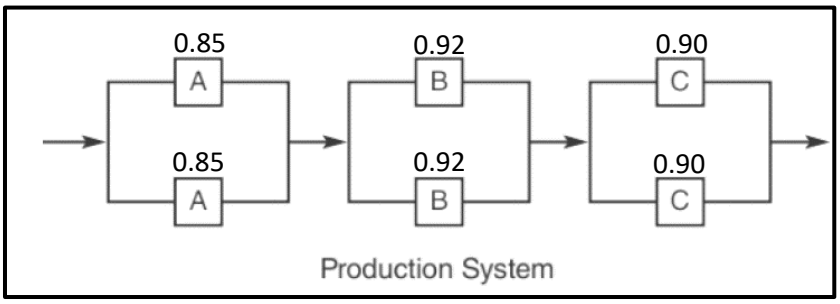

Figure 33: Production System for Magnaplex Inc. manufacturing process

Machine	Reliability
A	0.85
B	0.92
C	0.90

a. Analyze the system reliability, assuming only one machine at each stage (all the backup machines are out of operation).

$$P(A)\ \times P(B) \times P(C) = 0.85\ \times 0.92\ \times 0.90$$

$$= 70.38\%$$

b. How much is the reliability improved by having two machines at each stage?

$$P(System\ Reliability) = (1-(A')^2) \times (1-(B')^2) \times (1-(A')^2)$$
$$= (1-(1-0.85)^2) \times (1-(1-0.92)^2) \times (1-(1-0.90)^2)$$
$$= 0.9615315$$
$$= 96.15\%$$

$$System\ Improvement = 96.15\% - 70.38\%$$
$$= 25.77\%$$

6.3 Vehicle availability over the past 1560 days:

$$P(y=17) = \frac{1}{1560}$$
$$P(y=18) = \frac{3}{1560} = \frac{1}{520}$$
$$P(y=19) = \frac{22}{1560} = \frac{11}{780}$$
$$P(y=20) = \frac{190}{1560} = \frac{19}{156}$$
$$P(y=21) = \frac{(1560-190-22-3-1)}{1560} = \frac{56}{65}$$

Driver availability over the past 1560 days:

$$P(z=18) = \frac{1}{1560}$$
$$P(z=19) = \frac{6}{1560} = \frac{1}{260}$$
$$P(z=20) = \frac{95}{1560} = \frac{19}{312}$$
$$P(z=21) = \frac{(1560-95-6-1)}{1560} = \frac{243}{260}$$

We take the reliability requirements of the vehicles and drivers as 20 because 20 vehicles should be available to give reliable service. Thus we can use the following probability to ensure there is always more than or equal to 20 drivers and vehicles available.

$$P(\geq 20\ drivers\ and\ vehicles)$$
$$= \big(P(y = 20) \times P(z = 20)\big) + \big(P(y = 20) \times P(z = 21)\big)$$
$$+ \big(P(y = 21) \times P(z = 20)\big) + \ \big(P(y = 21) \times P(z = 21)\big)$$

$$= \left(\frac{19}{156} \times \frac{19}{312}\right) + \left(\frac{19}{156} \times \frac{243}{260}\right) + \left(\frac{56}{65} \times \frac{19}{312}\right) + \ \left(\frac{56}{65} \times \frac{243}{260}\right)$$

$$= 0.9789209402$$

The number of reliable days:

$$P(\geq 20\ drivers\ and\ vehicles)\ \times 365 = 357.3061\ day\ per\ year$$

There are only about 8 days in a year where the delivery service of the company will be unreliable. Which makes the company's delivery service fairly reliable.

When looking at the reliability of the vehicles and drivers respective the results are as follows:

The Probability of there always being more than or equal to 20 vehicles are:

$$P(y \geq 20) = \frac{56}{65} + \frac{19}{156} = 0.9833$$

The Probability of there always being more than or equal to 20 drivers are:

$$P(z \geq 20) = \frac{243}{260} + \frac{19}{312} = 0.9955$$

One can see the vehicles are less reliable than the drivers thus it will benefit the company more to increase the number of vehicles to 22. When increasing the vehicles to 22 you assume that the reliability of the vehicles is high enough and thus one will only look at the max achievable reliability of the drivers which will be 0.9955. This will give you 0.9955 x 365 = 363.3572 reliable days per year. Thus it will improve the reliability from 357 days to 363 days (6 days).
When increasing the number of vehicles by one yields an improvement of 1.67% in the reliability of the delivery service.

Conclusion

After analysing the company one can see that the most sales come from the Gifts and Technology class thus one can say the company should focus on these areas more but the classes that bring in the most revenue are the luxury and Technology classes thus the company should focus on them. The company could expand these ranges by looking at more possible products that could be sold.

The company's delivery times are very scattered around the centre lines and they could try to get them more stable but delivery time is very class dependent. The distance from the warehouse also plays a massive role in terms of delivery time. It is only the “gifts” and “sweets” classes which seem mildly out of control. The company should try and minimise their average delivery time from 20 hours to 17 hours to minimise the cost and to avoid making losses. The “Household” class have the longest average delivery time because it tends to be big items. If the company invest in some more shipment workers or some heavy-duty machinery they could decrease this average time and satisfy the customers even more. The company should increase their number of vehicles to 22 as this would increase the reliability of the delivery service.

The ”Browsing” reason in the “Why.Bought” column generated the most revenue. Thus if the company could invest a bit into their website to make the browsing experience more enjoyable it could ramp up sales even more. The “Random” reason generated less revenue, this could be because people tend to not know what it means and don’t choose it.

References

Chitranshi, U. (2022, October 05). *GreyCampus*. Retrieved from https://www.greycampus.com/blog/quality-management/how-to-measure-process-capability-and-process-performance

Hesing, T. (2022, 10 06). *Six Sigma Study Guide*. Retrieved from Six Sigma Study Guide. com: https://sixsigmastudyguide.com/x-bar-s-chart/

Morphist. (2019, April 12). *Cross Validated*. Retrieved from Stats Stack exchange: http://www.stats.stackexchange.com

Stobierski, T. (2021). *A Beginner's Guide to Data & Analytics.* US: Harvard Business School.