Quality Assurance 344 Project

Jemma Buhrs
23564091
October 2022

Abstract

This report is the discussion of an analysis performed by an industrial engineering on the sales department of a business. The industrial engineer received a data set that contains information relating to the sales of the business in the last few years. The report will make use of and discuss certain quality assurance and statistic techniques, providing explanations where needed.

Table of Contents

List of Figures

1. Introduction

This report is the discussion of an analysis performed by an industrial engineering on the sales department of a business. The industrial engineer was given a data set that contains information relating to the sales of the business in the last few years. The report will make use of and discuss certain quality assurance and statistic techniques, providing explanations where needed. Initially, the data is pre-processed before analysis can begin. For the industrial engineer to grasp an understanding of the data, she will perform a short analysis on the distribution of the data. Further, x-bar and S charts will be initialised and analysed to determine process control abilities and a hypothesis test will be conducted. Lastly, the industrial engineer will conduct statistical problems not related to the sales dataset.

2. Part 1

To pre-process the data, the data needs to be arranged and the missing values need to be omitted. After this, the data is valid and can be used in analysis.

Figure 1 and 2 below are the data sets of the *Valid Data Set* and *Invalid Data Set* respectively.

	ID	X	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought	countn
1	47101	463	50	Clothing	1030.86	2021	1	1	9.0	Recommended	1
2	88087	2627	21	Clothing	428.03	2021	1	1	10.0	Recommended	2
3	25418	3374	68	Household	13184.41	2021	1	1	48.5	Website	3
4	13566	5288	94	Household	7021.90	2021	1	1	42.0	Recommended	4
5	84692	8182	35	Clothing	475.18	2021	1	1	9.0	Recommended	5
6	46305	9272	72	Clothing	580.98	2021	1	1	8.5	Random	6
7	92105	9712	45	Household	6877.00	2021	1	1	43.0	Recommended	7
8	21614	12163	27	Clothing	513.13	2021	1	1	9.5	Recommended	8
9	12174	12195	56	Household	14538.64	2021	1	1	41.5	EMail	9
10	84558	20004	74	Food	255.41	2021	1	1	2.0	Recommended	10
11	15630	20509	32	Clothing	164.56	2021	1	1	9.0	Recommended	11
12	81216	21970	87	Clothing	173.76	2021	1	1	10.0	Recommended	12
13	56240	27161	45	Household	17681.94	2021	1	1	45.5	Website	13

Figure 1 Valid Data Set

	ID	X	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought	countm
1	0	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	2
3	0	0	0	0	0	0	0	0	0	0	3
4	0	0	0	0	0	0	0	0	0	0	4
5	0	0	0	0	0	0	0	0	0	0	5
6	0	0	0	0	0	0	0	0	0	0	6
7	0	0	0	0	0	0	0	0	0	0	7
8	0	0	0	0	0	0	0	0	0	0	8
9	0	0	0	0	0	0	0	0	0	0	9
10	0	0	0	0	0	0	0	0	0	0	10
11	0	0	0	0	0	0	0	0	0	0	11

Figure 2 Invalid Data Set

3. Part 2

3.1. Summary of numerical data

To understand the values within the *Sales* dataset, a summary of the data is obtained for each of the numeric data groups (age, price, year, month, day and delivery time). The summary includes minimum, first quartile, median, mean, third quartile and maximum values.

```
summary(validData$AGE)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
18.00   38.00   53.00   54.57   70.00  108.00
summary(validData$Price)
  Min.   1st Qu.   Median     Mean  3rd Qu.     Max.
-588.8    482.3   2259.6  12293.7  15270.7 116619.0
summary(validData$Year)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2021    2022    2025    2025    2027    2029
summary(validData$Month)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.000   4.000   7.000   6.521  10.000  12.000
summary(validData$Day)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00    8.00   16.00   15.54   23.00   30.00
summary(validData$Delivery.time)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.5     3.0    10.0    14.5    18.5    75.0
```

Figure 3 Summary of Data

This gives the big picture of the dataset. For example, it can be seen that the minimum age was 18 years old, the maximum age was 108 years old with an average age of 54.7 years old. These summaries provide a depiction of the distribution of data within the set.

Below in figures 4 to 9, a visual representation of each numeric data groups is seen. The box plot diagram enables a quick analysis of the distribution of the data to be obtained.

For the *Summary of Age* below, it seen observed that the median age is about 55 years, and that the age distribution range is larger for ages above the median than below the mean.

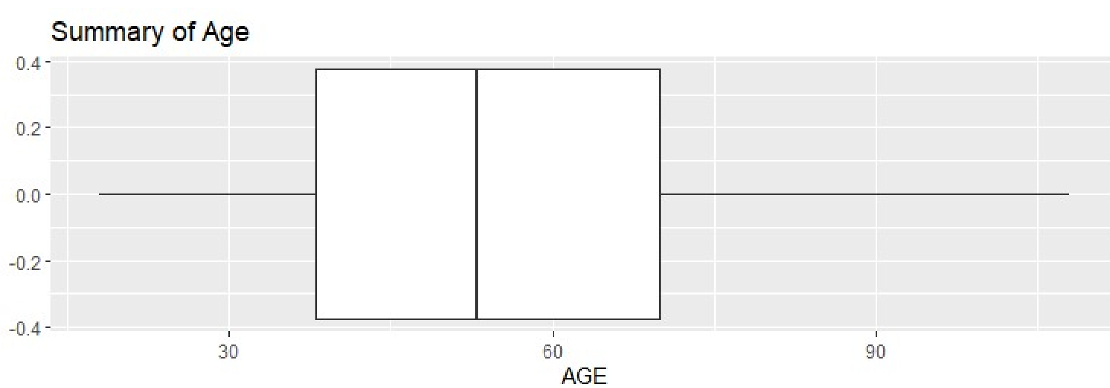

Figure 4 Summary of Age

A similar analysis can be conducted for each of the numeric data groups. For example, in figure 5 below, the *Summary of Price* shows that half of the products are sold for less than about R2500 while the other half of the sales span over a price distribution on R2500 all the way up to the maximum price R116000. This shows that many of the sales are conducted within small bounds of a price difference.

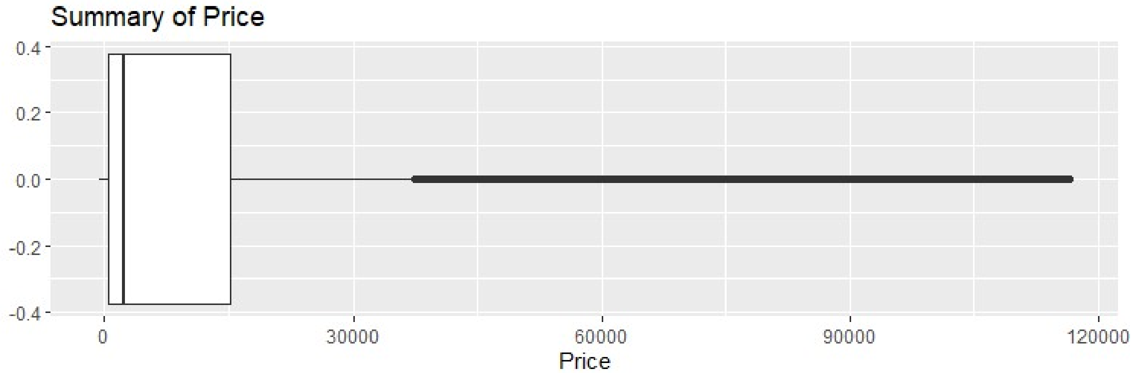

Figure 5 Summary of Price

The *Distribution of Year* below in figure 6 shows that as many sales that occur before the year 2025 occur after the year 2025.

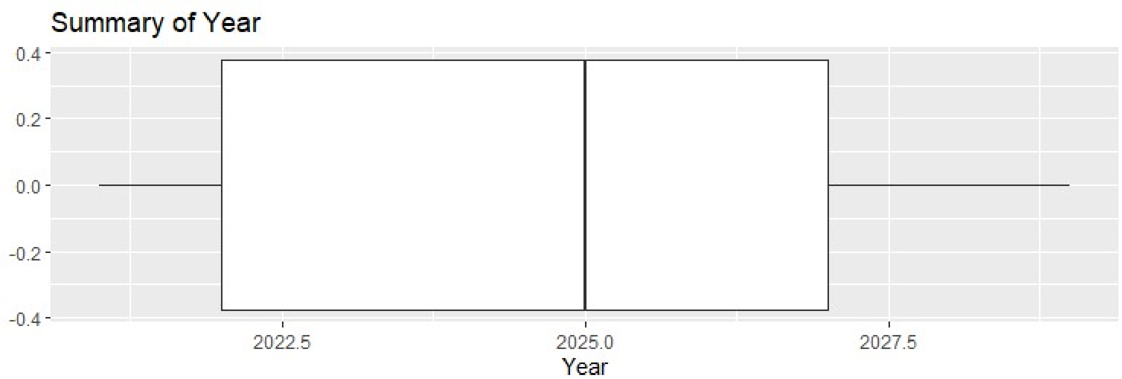

Figure 6 Summary of Year

Below the *Summary of Month* shows a relatively even distirbution of sales in each month of the year.

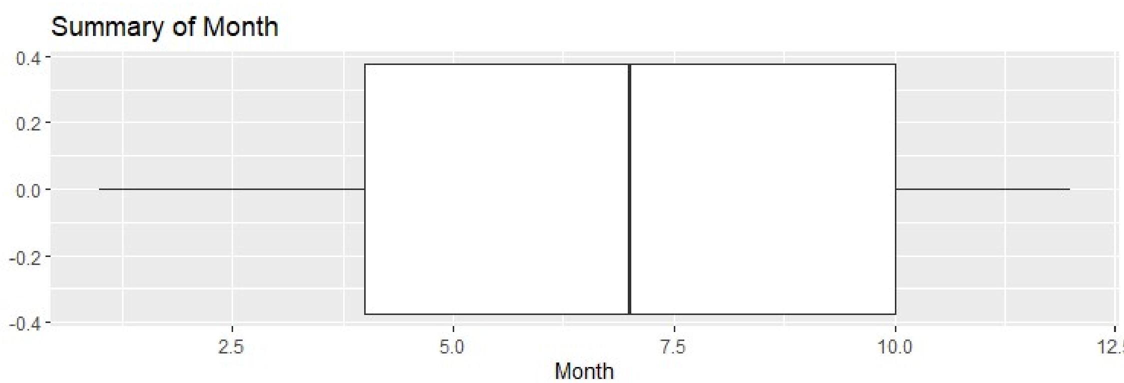

Figure 7 Summary of Month

Similarly, to the month, the *Summary of Day* below shows a relatively stable number of sales on each day of the week.

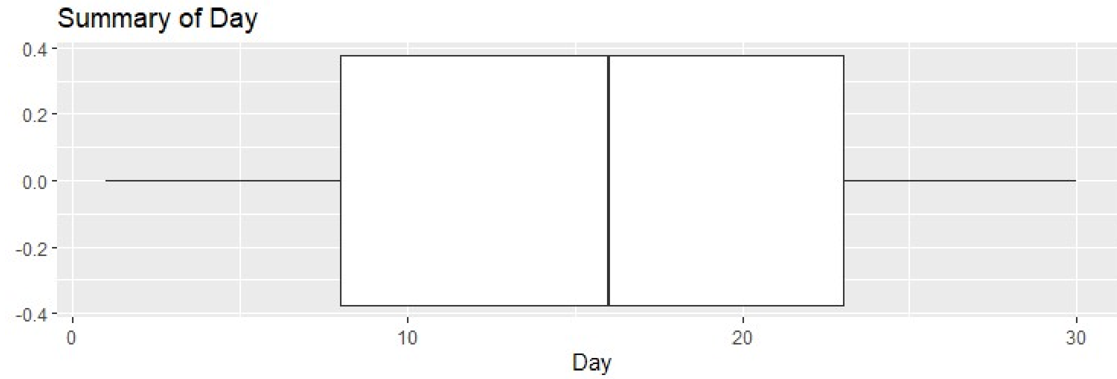

Figure 8 Summary of Day

In the *Summary of Delivery Time* in figure 9 it is observed that three quarters of the sales are delivered within twenty hours while one quarter of sales are delivered between twenty and seventy-five hours.

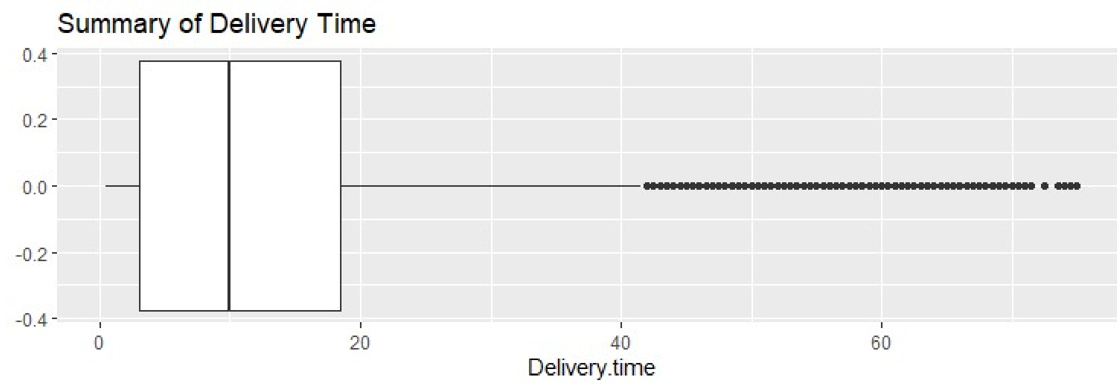

Figure 9 Summary of Delivery Time

3.2. Correlation of data groups

To determine which data groups within the data set correlate with each other an analysis is conducted. Multiple plots were created to aid in the analysis and understanding of the data.

3.2.1. Distribution of each numeric data group

Figures 10 to 15 visually represent the distribution of the data groups. The *Distribution of Age* shows that majority of the products are bought by individuals between the ages of 30 and 90. This agrees with the age summary which shows that the first quartile is 38 years, and the third quartile is 70 years. The *Distribution of Price* shows that although the maximum price is R116619, there is a large portion of the data values below R30000. The *Distribution of Products Bought Per Year* shows that although more products were bought in 2021, the number of products bought thereafter remains relatively stable. Similarly, the *Distribution of Products Bought Per Month* and *Distribution of Products Bought Per Day* shows a stable demand of products. Lastly, the *Distribution of Delivery Time* varies drastically. From this analysis is it decided that a further investigation will be conducted on Price and Delivery Time as these two data groups are less stable.

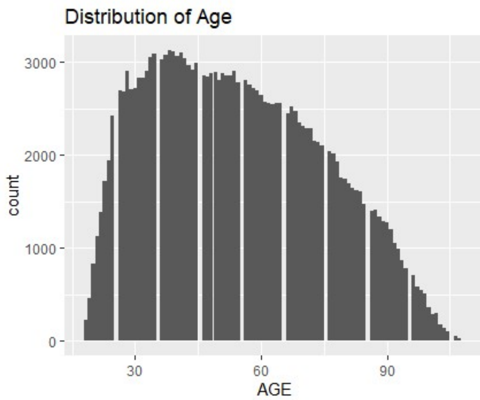

Figure 10 Distribution of Age

Figure 11 Distribution of Price

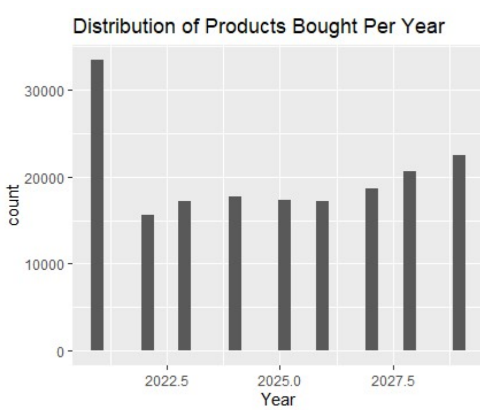

Figure 12 Distribution of Products Bought Per Year

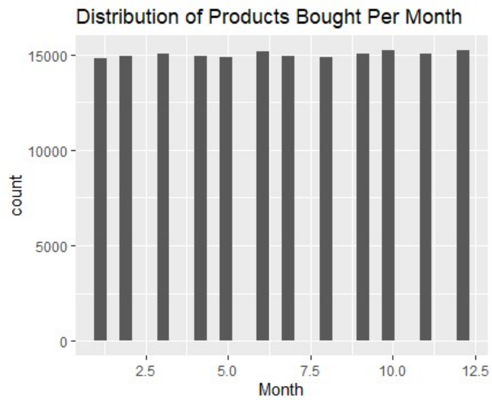

Figure 13 Distribution of Products Bought Per Month

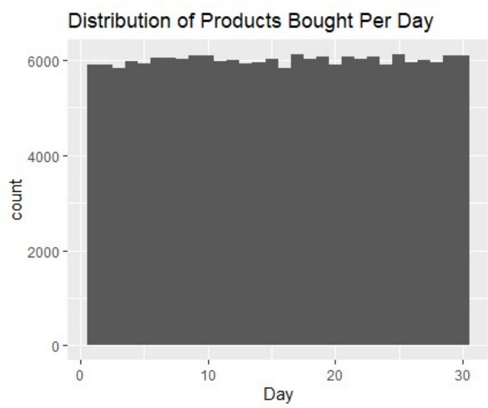

Figure 14 Distribution of Products Bought Per Day

Figure 15 Distribution of Delivery Time

3.2.2. Corelation of Data Groups

3.2.2.1. Products Bought per Month by Class

Figure 16 shows the number of products bought per month for each of the class products. This shows that the demand for products is stable throughout the year.

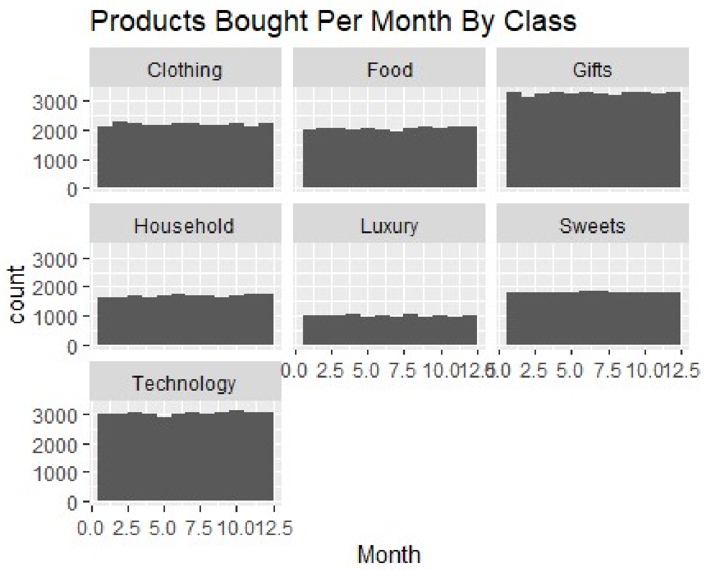

Figure 16 Products Bought Per Month by Class

3.2.2.2. Products Bought per Age by Class

Figure 17 shows the number of products bought per age for each class of product. This can be used to determine which age groups are more likely to purchase which class of product. This can enable certain business strategies such as marketing. For example, under the *Luxury* plot, a spike is seen for customers around the age of forty years old. This can enable marketing strategies to be aimed at this target audience.

Figure 17 Products Bought Per Age by Class

3.2.2.3. Products Bought per Price by Reason for Purchase

Figure 18 shows the range of price paid for products with regards to the reason for purchase. This can be used to determine which reason causes customers to spend less/more money when purchasing a product. It is observed that for less expensive products, customers are likely to make a purchase if the item has been recommended to them.

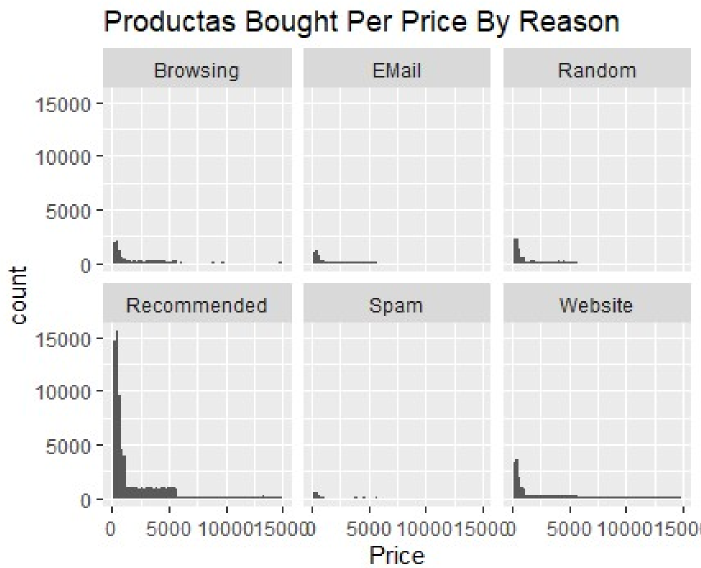

Figure 18 Products Bought Per Price by Reason

3.2.2.4. Delivery Time for Products by Class

Figure 19 begins to bring more sense to the drastically unstable *Distribution of Delivery Time* graph seen earlier. This graph shows the distribution of delivery time for products with relation to the class of product bought. Figure 19 demonstrates that similar delivery times of items within the same class are observed.

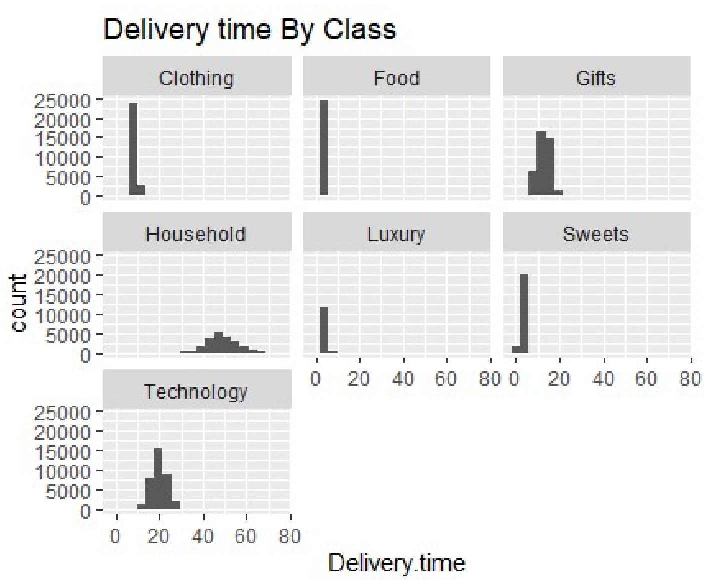

Figure 19 Delivery Time by Class

3.3. Calculations

Capability potential (Cp) and Capability performance (Cpk) represent the ability of a process to achieve its specifications. The Cp ratio implies that the spread of data fits into the control limits 1.142207 times. A Cpk ratio represents the proximity of the process mean to the required specification. A Cpk of 0.90472 is not considered a capable value. This shows that the process has many non-conforming delivery times. To improve the Cpk, the process should be redesigned to achieve a more optimal value. Figure 20 displays the necessary process capability indices.

```
      sd       cpu       cp     cpl       cpk
3.501993 0.3796933 1.142207 1.90472 0.3796933
```

Figure 20 Process Capacity Indices

4. Part 3

X-bar and S charts are a valuable tool to ensure processes are maintained within required specifications. The respective charts provide information about how well processes are performing and whether the process needs to be redesigned, maintained, or is running sufficiently.

4.1. Initializing x-bar and s charts

To initialise the x-bar and S charts for each class the first four hundred and fifteen delivery times for each class in split into thirty samples of fifteen units each. Using the *qcc* packaged offered on RStudio, the x-bar chart for Technology is produced below. The x-bar chart gives the limits of where all samples of the Technology delivery time mean should lie between. The centre line (CL) for the Technology x-bar chart is 20.37444 and the upper control limit (UCL) and lower control limit (LCL) is calculated from given formulas.

$$UCL = \bar{\bar{x}} - 0.789 * \bar{S}$$

Figure 21 Formula UCL x-bar chart

$$LCL = \bar{\bar{x}} + 0.789 * \bar{S}$$

Figure 22 Formula LCL x-bar chart

Figure 21 is the formula used to calculate the UCL for an x-bar chart and figure 22 is the formula used to calculate the LCL for an x-bar chart. The value 0.789 is a constant used for samples of size fifteen, this remains constant for every class.

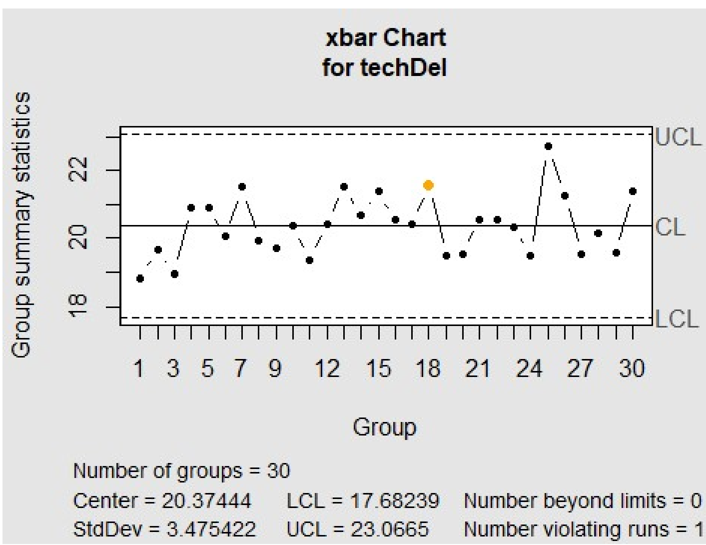

Figure 23 x-bar chart for Technology

$$UCL = 0.428 * \bar{S}$$

Figure 24 Formula UCL S chart

$$UCL = 1.572 * \bar{S}$$

Figure 25 Formula LCL S chart

Figure 24 is the formula used to calculate the UCL for an S chart and figure 25 is the formula used to calculate the LCL for an S chart. The values 0.428 and 1.572 are constants used for samples of size fifteen, this remains constant for every class.

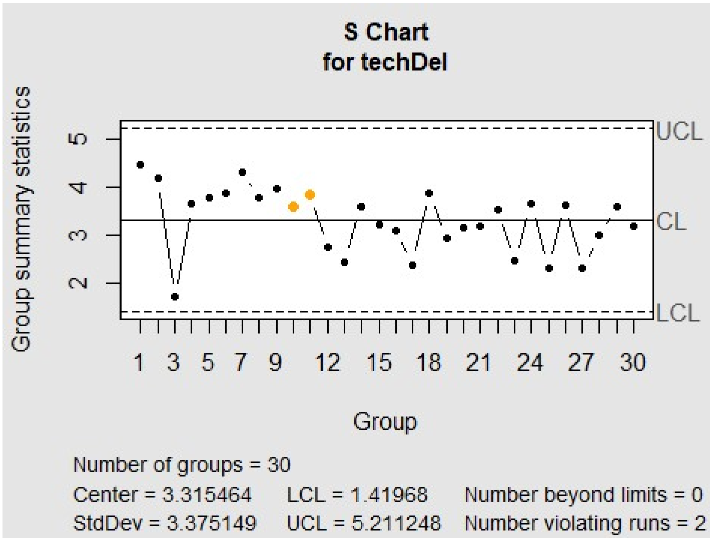

Figure 26 s chart for Technology

The process of developing the x-bar and S charts is repeated for each class. The results of the UCL, CL and LCL calculations are seen below in figures 27 and 28. The UCL and LCL limits lie six sigma's apart of each other. U1Sigma represents one sigma value above the CL, similarly, the L2Sigma represents two sigma values below the CL.

Class	**UCL**	**U2Sigma**	**U1Sigma**	**CL**	**L1Sigma**	**L2Sigma**	**LCL**
Technology	22.99034	22.11837	21.24641	20.37444	19.50247	18.63051	17.75854
Clothing	9.4066	9.261067	9.115533	8.97	8.824467	8.678933	8.5334
Household	50.27781	49.03928	47.80075	46.56222	45.32369	44.08516	42.84663
Luxury	5.511178	5.252637	4.994097	4.735556	4.477015	4.218475	3.959934
Food	2.704311	2.632874	2.561437	2.49	2.418563	2.347126	2.275689
Gift	9.491443	9.114666	8.737888	8.361111	7.984334	7.607556	7.230779
Sweets	2.901901	2.760526	2.619152	2.477778	2.336404	2.19503	2.053655

Figure 27 x-bar Chart

Class	**UCL**	**U2Sigma**	**U1Sigma**	**CL**	**L1Sigma**	**L2Sigma**	**LCL**
Technology	5.211909	4.579761	3.947612	3.315464	2.683316	2.051167	1.419019
Clothing	0.869881	0.764373	0.658866	0.553359	0.447852	0.342345	0.236838
Household	7.402932	6.505036	5.60714	4.709244	3.811348	2.913452	2.015556
Luxury	1.545345	1.357911	1.170478	0.983044	0.79561	0.608177	0.420743
Food	0.426993	0.375203	0.323414	0.271624	0.219834	0.168045	0.116255
Gift	2.252069	1.978917	1.705766	1.432614	1.159462	0.886311	0.613159
Sweets	0.84502	0.742528	0.640036	0.537545	0.435053	0.332561	0.230069

Figure 28 s Chart

5. Part 4

5.1. Part 4.1

A

To perform statistical process control, one needs to continually track the process to ensure measures are taken when the points on x-bar chart exceed the set bounds.

In figure 29 below, the total number of points above the upper control limit (UCL) and below the lower control limit (LCL) of the x-bar chart for all classes in the data set as well as the first three and last three data points and their location in the data set that are outside of the control limits are observed.

Class	Total number of data points outside the control limits		First three data points outside the control limits [location; delivery time]	Last three data points outside the control limits
	>UCL	<LCL		
Technology	1	13	[248; 20.5] [318; 13] [474; 20.5]	[1597; 19] [1689; 15.5] [1718; 19]
Clothing	12	1	[129; 9.5] [157; 9] [316; 9]	[1259; 9.5] [1296; 10.5] [1475; 10]
Household	172	0	[2; 45.5] [14; 54.5] [25; 43]	[1304; 46.5] [1305; 51] [1307; 46]
Gifts	2579	0	[1; 12.63333] [2; 12.86667] [3; 12.90]	[498; 12.16667] [499; 13.10000 [500; 12.13333]
Sweets	3	1	[354; 1.5] [336; 3.5]	[1016; 3.5] [1051; 2.5]
Food	1	6	[519; 2.266667] [597; 2.266667] [734; 2.266667}	[921; 2.266667] [1131; 2.733333] [1158; 2.266667]
Luxury	0	389	[5; 3.866667] [6; 3.60] [7; 3.80]	[751; 3.833333] [753; 3.933333] [759; 3.933333]

Figure 29 Points Outside Control Limits

Technology, Clothing, Sweets and *Food* classes show relatively stable processes. Even though a few points are outside of the control limits, the processes are within control and are running sufficiently. However, the *Household* and *Gifts* classes show that all points are above the UCL and the *Luxury* class shows that all the points are below the LCL.

Figures 30 and 31 show the delivery time for *Household* and *Gifts* respectively from the first purchase to the last purchase in the dataset. It is observed that overtime, there is an increase in the delivery time. Since the first four hundred and fifty data points are used to initialise the x-bar charts, it makes sense that many of the points, after the initial four hundred and fifty, are above the UCL. To counter the unstable increase of delivery time, the process should be redesigned and closely monitored to attempt to reduce delivery time.

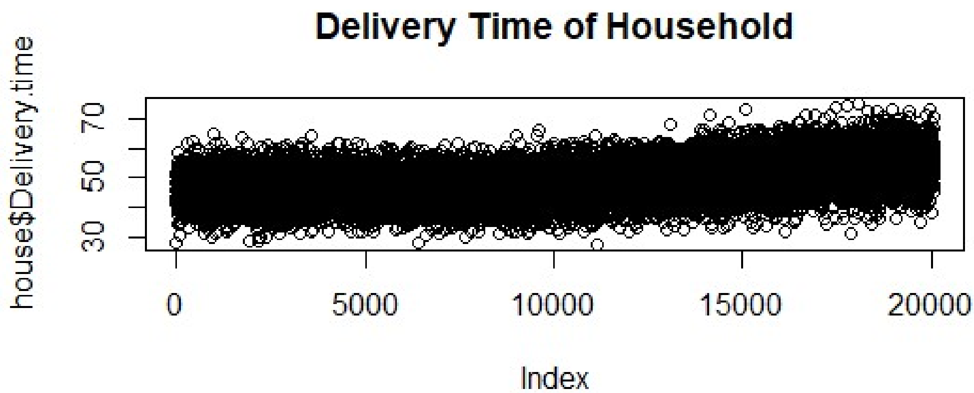

Figure 30 Delivery Time of Household

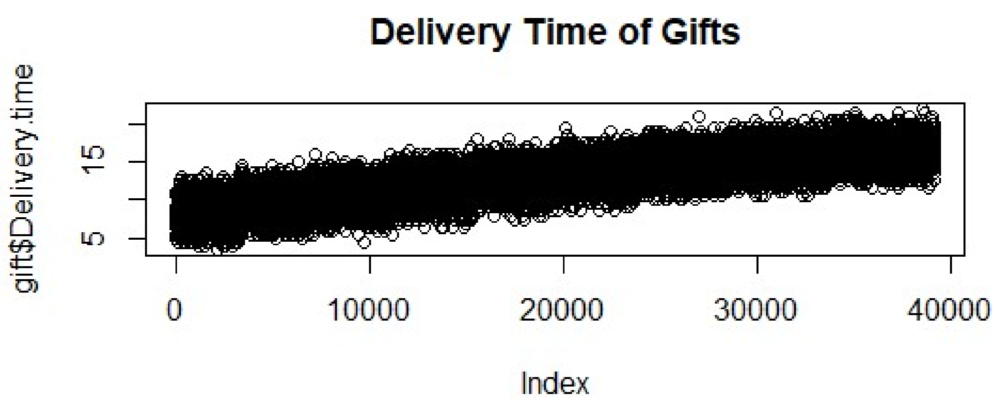

Figure 31 Delivery Time of Gifts

In figure 32 is the delivery time for *Luxury* from the first purchase to the last purchase in the dataset. It is observed that the delivery time decreased since the initialisation of the x-bar chart. Although this appears to be a positive aspect, the business could be overspending on delivery costs. Thus, an investigation should be taken into the process to assess possible improvements and cost reductions.

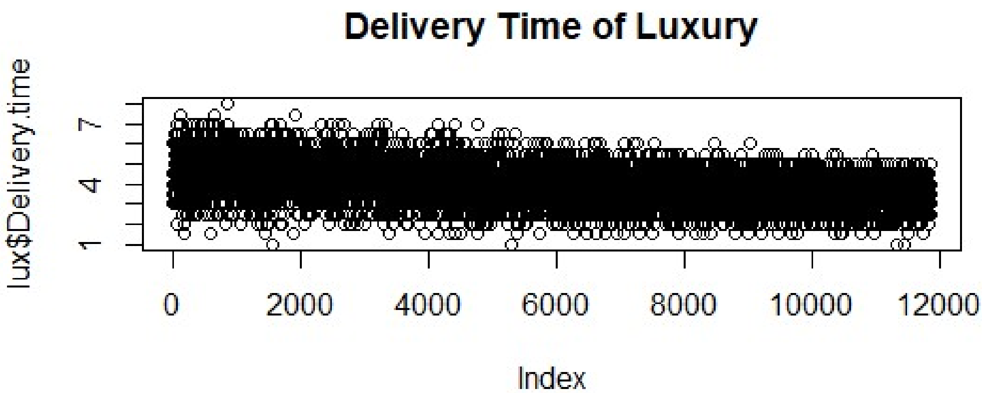

Figure 32 Delivery Time of Luxury

B

In this scenario the S chart is used to determine how many sample standard deviations consecutively lie within -0.3 and +0.4 sigma values of the CL for each class. This shows the stability of the delivery times per class and their concentration to the CL. Figure 33 displays the respective class names, the last index of the most consecutive group and the maximum number of consecutive groups. *Technology, Clothing* and *Sweets* shows the most number consecutive points. This gives an indication that the respective processes are stable. The classes *Household, Luxury* and *Gifts* have the least number of consecutive points. This can be expected after the discussion in part A about their change in delivery time.

	class.names	index	max.consec
1	TECH	2318	157
2	CLOTH	1699	92
3	HOUSE	1284	10
4	LUX	753	34
5	FOOD	1595	87
6	GIFT	NaN	1
7	SWEET	1366	101

Figure 33 Consecutive points in S chart

5.2. Part 4.2

A manufacturer's error, also known as a Type I error, is made when the manufacture performs an investigation into a process, but the process is running well. This results in cost and time wasted and should be avoided. The probability of such an error occurring is the same for any process that is occurring, the probability is 0.2699796%.

5.3. Part 4.3

The business is charged R329 for every hour of every item of class *Technology* that is delivered late. Further, it costs the business R2.5 to decrease the delivery time by one hour for every item. Currently, it costs the business R758674 for items being delivered late. Figure 34 shows the change in cost for every hour, x, that is decreased from delivery time. Therefore, the minimal cost for the business occurs at a decrease of 6 hours and a cost of R1391.

Figure 34 Cost of decreasing delivery time by x-hours

5.4. Part 4.4

A customer's error, also known as a Type II error, is made when there is an error in the process, but it is not detected and thus, goes unnoticed. The probability of a Type II error made for delivery time for class *Technology* is 49.54249% given the process average moves to 23 hours.

6. Part 5

A multivariate analysis of variance (MANOVA) is used when there are at least two dependent variables needed to consider between the several independent variables. (Radečić, 2022)

In this scenario,

H0: the class price and delivery time mean for all class do not significantly differ

H1: at least one of the classes price and delivery time mean differ from the other classes

The results of the MANOVA are seen in figure 35 below. It is seen that the p-value is close to zero, thus the null hypothesis is rejected. Therefore, at least one of the classes price and delivery time mean differs from the other classes.

```
                   Df Pillai approx F num Df den Df    Pr(>F)    
validData$Class      6 1.6796   157265     12 359952 < 2.2e-16 ***
Residuals       179976                                           
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 1 :
                    Df     Sum Sq    Mean Sq F value    Pr(>F)    
validData$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 2 :
                    Df   Sum Sq Mean Sq F value    Pr(>F)    
validData$Class      6 33461034 5576839  629489 < 2.2e-16 ***
Residuals       179976  1594464       9                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 35 MANOVA table

Figure 36 and 37 show the boxplots of the price and delivery time respectively per class. The distribution in price of classes *Luxury* and *Technology* is larger than that of the other classes. Similarly, the distribution in delivery time of classes *Gifts, Household* and *Technology* is larger than that of the other classes. Therefore, this agrees with the hypothesis test conducted above as at least one of the classes price and delivery time mean differs from the other classes.

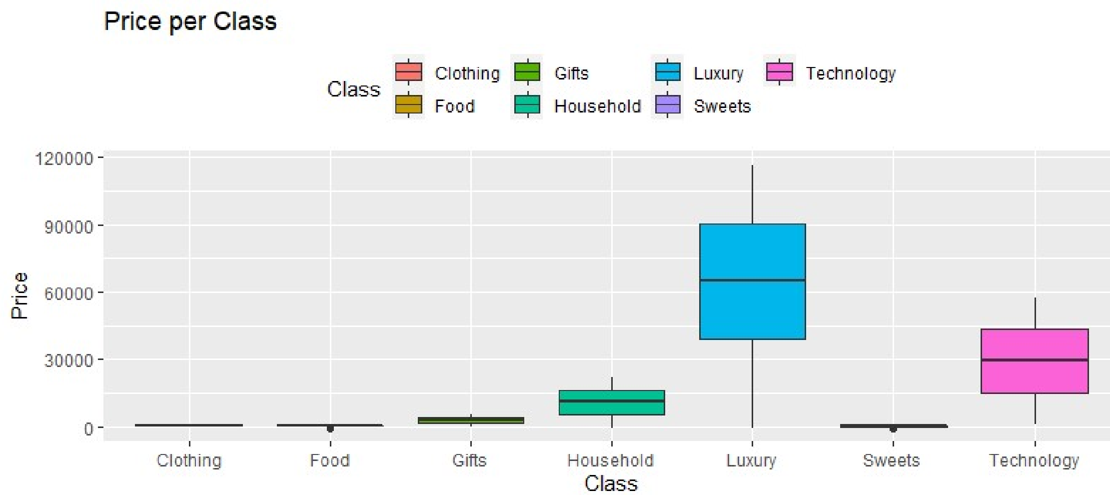

Figure 36 Price per class boxplot

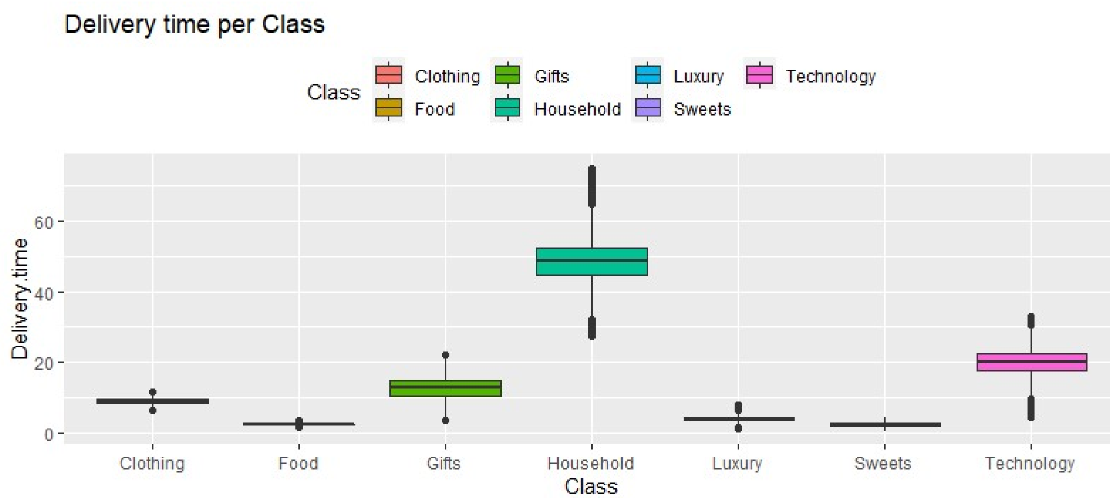

Figure 37 Delivery time per class boxplot

7. Part 6

7.1. Question 6+7 pg. 363

The Taguchi loss function represents the cost of not producing a perfect product/service. The process is not sufficient within the control limits, as it cost of poor quality is experienced.

The Taguchi loss function obtained is,

$$L = k * (y - m)^2 \qquad where, k = 24489.8$$

After the cost of the part is changed, the Taguchi loss function is,

$$L = k * (y - m)^2 \qquad where, k = 20408.16$$

If the thickness of the part is 0.027cm, the loss is 3.44898

7.2. Question 27 pg. 365

The reliability of the system failing when only one of each machine is used is 0.7038 while the reliability of the system failing when two of each machine are placed in parallel is 0.965028. this is a reliability improvement of 0.261228.

7.3. Vehicle and Driver Reliability

The company can expect a vehicle reliability of 0.4420258 when they have 21 vehicles. Therefore, 161.34 days of the year the company will have vehicle reliability. This improves to a reliability of 0.6399049 and 233.57 days of the year by adding one vehicle. The driver reliability, with 21 drivers, is 0.8484614 or 309.69 days of the year.

8. Conclusion

This report discussed the analysis performed by an industrial engineering on the sales department of a business. The industrial engineer was given a data set that contains information relating to the sales of the business in the last few years. The report made use of and discussed certain quality assurance and statistic techniques. Firstly, the data was pre-processed. For the industrial engineer to grasp an understanding of the data, she performed a short analysis on the distribution of the data. Further, x-bar and S charts were initialised and analysed to determine process control abilities and a hypothesis test was conducted. Lastly, the industrial engineer completed statistical problems not related to the sales dataset.

9. References

Radečić, D., 2022. *MANOVA in R – How To Implement and Interpret One-Way MANOVA | R-bloggers*. [online] R-bloggers. Available at: <https://www.r-bloggers.com/2022/01/manova-in-r-how-to-implement-and-interpret-one-way-manova/> [Accessed 14 October 2022].