Quality Assurance 344

John Michael Steyn Laubscher

2022-10-09

Table of Contents

List of Figures

List of Tables

Abstract

The report is split up in to 6 parts each containing a different instruction. The goal of the report is to analyze the given client data of an online business using data analytical, statistical, and manipulative techniques. The information gathered throughout the analysis will be used to analyze and draw conclusions on the performance and situation of the business. This report was done by using R Markdown which was knitted into this Word document.

PART 1: Data Wrangling

The main purpose of part 1 is to make sure the sales data is processed and ready to analyze. The data was initially read into R studios using the read_csv function before performing the pre-processing steps. These steps included removing any "NA" values as well as any negative sales values from the data. A new column called "Index" was added to the data that acts as an index beginning from one and ending at the last entry of the dataset. The original dataset was split into a valid and invalid dataset. The invalid dataset contains 22 entries, which act as an indication of the quality of the original sales data. The code used to split these datasets can be seen in the attached R code file submitted separately.

To make myself familiar with the data I renamed some of the columns and rearranged the columns in a logical manner as can be seen in the representation below. In order to be able to plot and order the data the data was transformed to a numeric form. As part of the pre-processing steps, I thought it would be suitable to order the data according to dates that will be helpful in future parts.

```
Index   X   ID    Year Month Day Age     Class  Price   Delivery.Time  Why.Bought
463   463   47101 2021    01  01  50  Clothing  1030.86     9.0        Recommended
2627  2627  88087 2021    01  01  21  Clothing  428.03      10.0       Recommended
3374  3374  25418 2021    01  01  68 Household  13184.41    48.5       Recommended
5288  5288  13566 2021    01  01  94 Household  7021.90     42.0       Website
8182  8182  84692 2021    01  01  35  Clothing  475.18      9.0        Recommended
9272  9272  46305 2021    01  01  72  Clothing  580.98      8.5        Random
```

PART 2: Descriptive statistics

In part 2 standard descriptive statistics of the valid dataset was used to make myself familiar with the data. Table 1 reveals the number of instances number of features in the valid dataset.

Table 1: Number of Instances and Features

Instances_Features	values
Number of Instances	179978
Number of Features	11

Furthermore, descriptive statistics was used to calculate the average delivery time as well as the average age per product. This helped to better understand a distribution of the sales data set. The high standard deviation of both age and delivery times indicates that the data is spread out more.

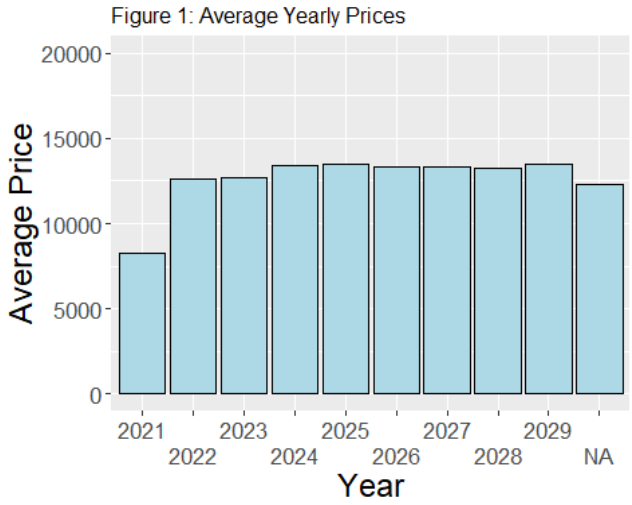
Table 2: Average and Standard Deviation of Age and Delivery Times

Delivery.Time	values
Ave Delivery Time	14.50031
Standard Deviation of Delivery Time	13.95578
Ave Age	54.56552
Standard Deviation of Age	20.38881

Furthermore, a plot of the yearly average prices is used to indicate the effect of inflation on markets as well as the general increase and decrease in average prices. As can be seen in Figure 1, the 2021 average prices were low and stayed relatively constant until 2029. This plot therefore indicated a fairly uniform distribution with the average prices remaining relatively constant throughout the years.

Table 3: Yearly Prices

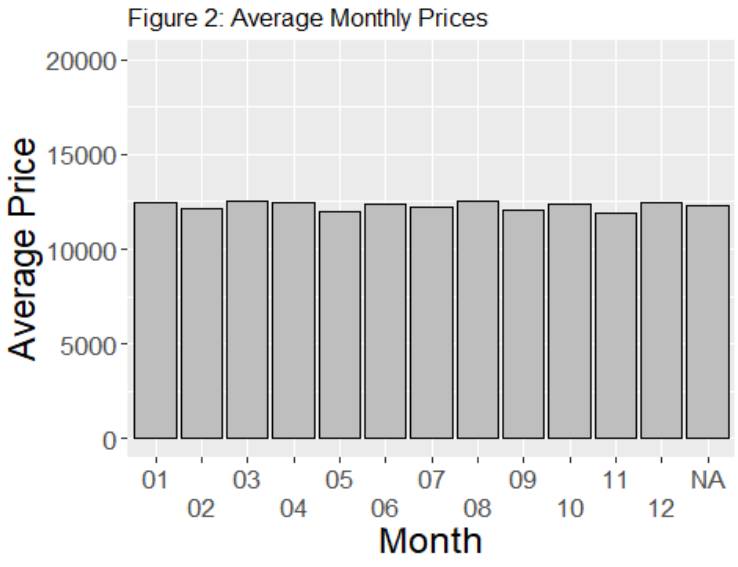
Year	Prices
2021	8236.111
2022	12611.145
2023	12696.081
2024	13418.392
2025	13515.110
2026	13330.417
2027	13307.182
2028	13243.287
2029	13475.226
NA	12311.923



A plot of the average monthly prices was also used to analyze the relationship between monthly prices and to see if sales are higher in certain months as well as when sale amounts peak. The analytical information could be used to help draw conclusions and form hypotheses on when to increase forecasting and when to expect higher sales. Figure 2 reveals how the average prices stay constant throughout 2021 to 2029 as well as a small deviation in prices, once again having a uniform distribution. It is important to note that both the plots in Figures 3 and 4 are not a true representation of the distributions as the means are used in the plots.

Table 4: Monthly Prices

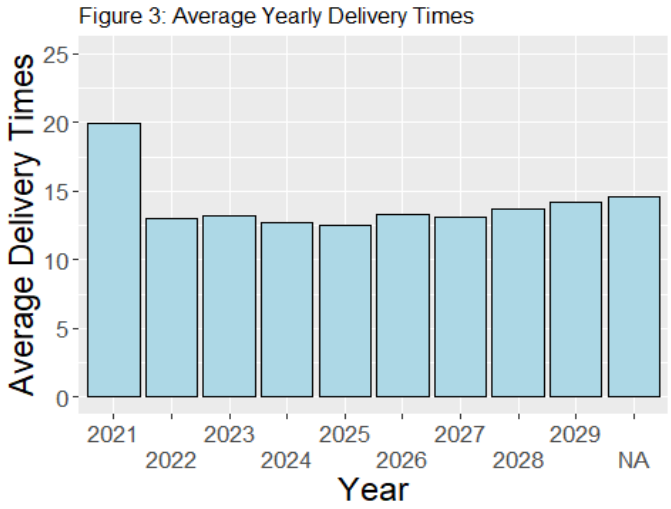
Month	Prices
01	12425.37
02	12170.10
03	12531.76
04	12442.53
05	11951.83
06	12380.76
07	12258.23
08	12569.87
09	12086.68
10	12361.94
11	11902.79
12	12437.34
NA	12311.92



A plot of the average yearly delivery times served very useful as it indicated that the delivery times were fairly high in 2021 and that it improved in the years to come and stayed constant at average of +- 13 hours. This plot in Figure 3 does not offer allot of useful information as we do not know whether the delivery times is allocated to local or international distances making this distribution irrelevant.

Table 5: Yearly Average Delivery times

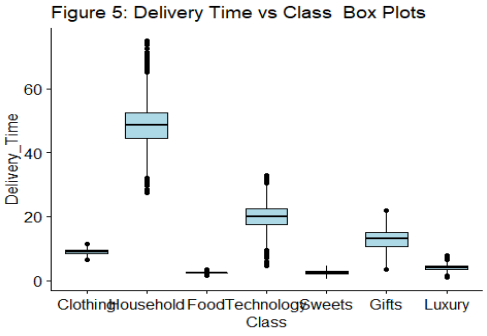
Year	Deliveries
2021	19.91709
2022	13.04408
2023	13.15513
2024	12.72298
2025	12.50804
2026	13.27426
2027	13.14842
2028	13.73842
2029	14.16426
NA	14.54435



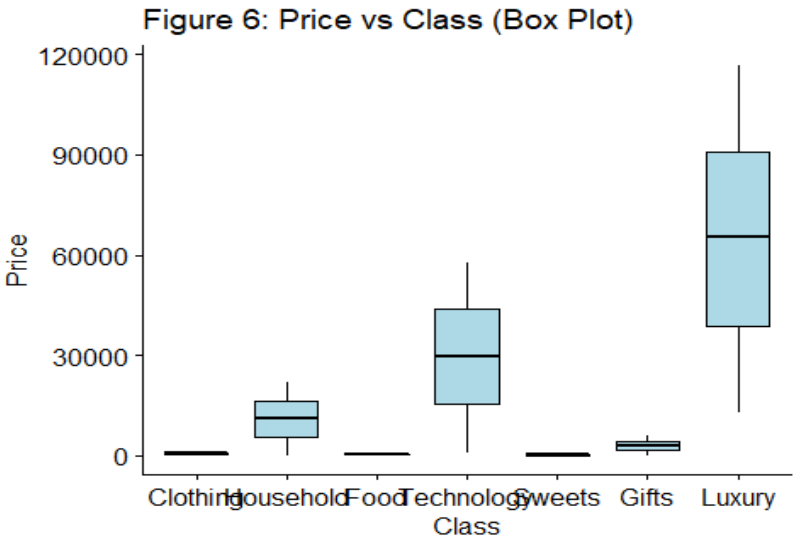
Furthermore, it is, however, crucial to note the relationship between the class of the products sold versus the time to make deliveries. The delivery time and class were plotted against each other as can be seen in figure 4. This plot clearly revealed that household products took the longest to deliver with an average delivery time of more than 45 hours. This makes sense as household product is mostly large or fragile that will require more careful delivery. Products of the class technology, gifts and clothes also has relatively long delivery times as can be seen in figure 4. A box plot is in fact more suitable as can be seen in Figure 5 as it offers allot more useful information regarding the distributions of each individual class. The box plots reveal that clothing, food, sweets and luxury classes has short distributions with small delivery times.

Table 6: Average Delivery time per Class

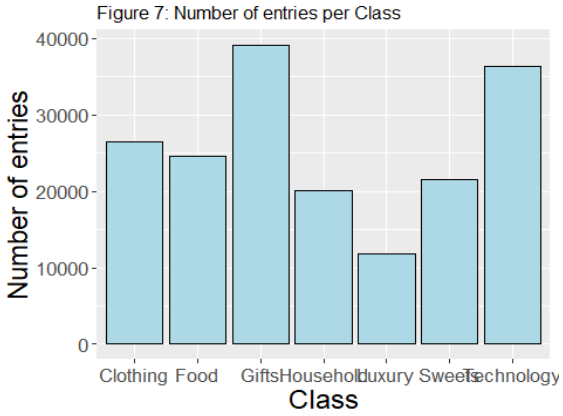
Class	Time
Clothing	8.999527
Food	2.502014
Gifts	12.890546
Household	48.719561
Luxury	3.971520
Sweets	2.501206
Technology	20.010950



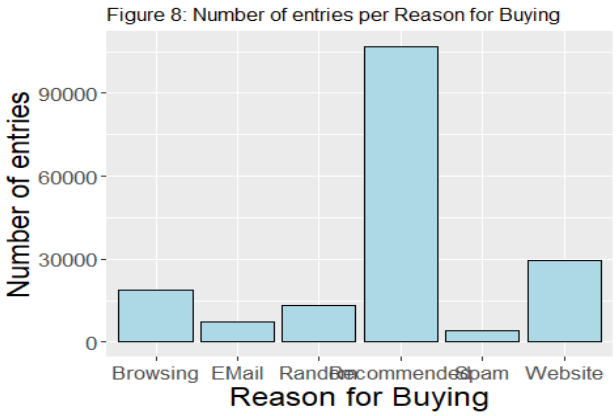
A box plot of the different classes versus the price per class served to be informative with regards to understanding the data further. Figure 6 revealed how luxury, Technology and household products are sold for higher prices than the other classes. Furthermore, the plot in Figure 4 is also a logical conclusion to make as these types of products are more expensive in general.



The Figure 7 plot indicates which class is sold the most throughout the years. It is therefore evident that products of the class gifts and technology are the most popular products. The plot reveals that the luxury items is sold the least. It can be concluded that products with high average prices such as luxury products are sold less often than products with lower prices.



In order to understand the advertising aspects of the company the below plot in Figure 8 allows an analysis of the reasons why a customer bought a certain product. It can therefore be concluded that the company's main source of advertising is by word of mouth. In other words, the product is mostly bought by customers because the product was recommended by someone else. Word of mouth or recommended based sales is a very good and cheap form of advertisement as it cost the company nothing and is done by the customers themself.



Process Capability indices for the technology class delivery times was calculated using the below formulas. Throughout the calculations both the upper specification limit (USL) and the lower specification limit (LSL) was assumed to be 24 and 0 hours respectively. It is important to note and understand why the LSL is assumed to be zero. It is an intuitive and logical concept as it is impossible for and delivery time to be less than zero, therefore causing the lower limit to be zero. The results of the below calculations are displayed and rounder to 3 significant figures in Table 7 for representation purposes. Note that the symbol "σ" represents standard deviation of delivery times in the technology class.

Process Capability (Cp): $Cp = \frac{(USL - LSL)}{6\sigma}$

Upper Process Capability (Cpu): $Cpu = \frac{(USL - Mean)}{3\sigma}$

Lower Process Capability (Cpl): $Cpl = \frac{(Mean - LSL)}{3\sigma}$

Process Capability index (Cpk): $Cpk = \min(Cpu, Cpl)$

Table 7: Process Capability indices

Process_Capability_Indices	**values**
Standard Deviation	3.500
Mean	20.000
USL (Assumed)	24.000
LSL (Assumed)	0.000
Cp	1.140
Cpu	0.381
Cpl	1.900
Cpk	0.381

It has been proven that a process capability above 1.33 has a high relative precision whereas a process capability below 1 has a low relative precision. It can therefore be concluded that with a Cp of 1.14, the process has a medium relative precision. Therefore it can be concluded that the technology delivery time distribution has a relatively good process capability.

PART 3: Statistical Process Control

3.1: Constructing Control Charts

For part 3 I decided to re-read the data into R Studios again and to pre-process the data as requested. I started by splitting the data into different data sets for each unique class. This allowed me to create a list containing each individual class dataset that will be used to create the X and S-Charts. The initialization of the two charts can be seen in the attached R code file that is submitted separately. After creating two empty X and S-Charts a big for loop containing two sub loops was used to calculate and read the respective values into the

respective data frames. The below two X and S-Chart contains the tabulated results of the initial 30 samples each with 15 sales.

Table 8: X-Chart

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	9.390601	9.250401	9.110200	8.970000	8.829800	8.689599	8.549399
Household	50.126859	48.938647	47.750435	46.562222	45.374010	44.185797	42.997585
Food	2.702226	2.631484	2.560742	2.490000	2.419258	2.348516	2.277774
Technology	22.888932	22.050770	21.212607	20.374444	19.536282	18.698119	17.859957
Sweets	2.883225	2.748076	2.612927	2.477778	2.342629	2.207479	2.072330
Gifts	9.451412	9.087978	8.724545	8.361111	7.997678	7.634244	7.270811
Luxury	5.468973	5.224501	4.980028	4.735556	4.491083	4.246611	4.002138

Table 9: S-Chart

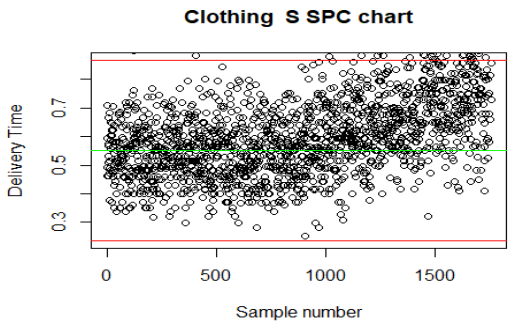
Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	0.5661217	0.4259213	0.2857210	0.1455206	0.0053202	-0.1348801	-0.2750805
Household	4.7468725	3.5586602	2.3704478	1.1822355	-0.0059769	-1.1941892	-2.3824015
Food	0.2772340	0.2064920	0.1357501	0.0650081	-0.0057339	-0.0764758	-0.1472178
Technology	3.4535377	2.6153751	1.7772125	0.9390500	0.1008874	-0.7372752	-1.5754378
Sweets	0.5520133	0.4168641	0.2817149	0.1465656	0.0114164	-0.1237328	-0.2588821
Gifts	1.4721776	1.1087441	0.7453106	0.3818771	0.0184436	-0.3449899	-0.7084233
Luxury	1.0194840	0.7750115	0.5305389	0.2860664	0.0415938	-0.2028787	-0.4473512

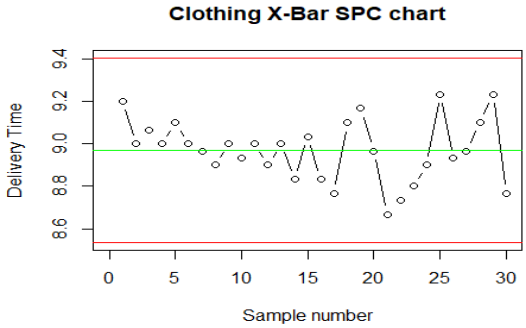
3.2: Continue drawing samples of 15 delivery times

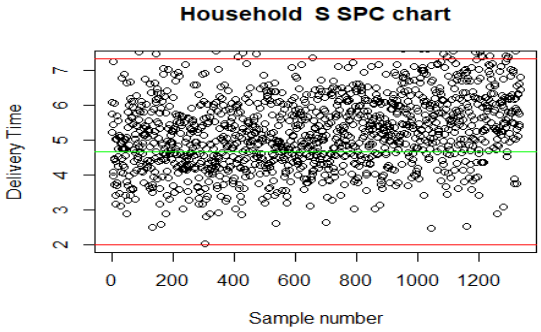
Statistical process control (SPC) was used to sample the rest of the data. The data was then stored in two separate datasets namely an “X_List” and a “S_List” as can be seen in the separate R code which will be further used in Part 4.

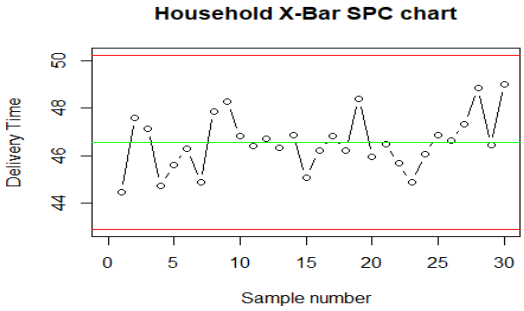
PART 4: Optimizing the delivery processes

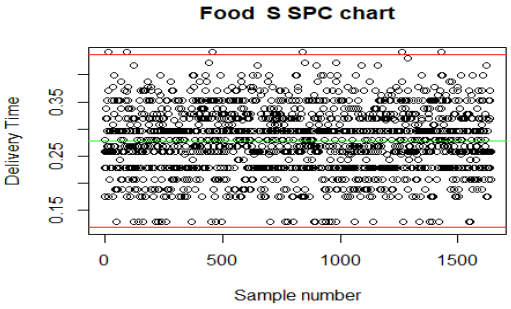
In part 4.1 the x-bar SPC chart and the s SPC chart was plotted for each unique class in order to give a visual representation of the distribution of each class. With regards to analysis, these plots served to be very insightful as you can analyse each class individually. You can identify the outliers, means, distributions as well as upper and lower limits.

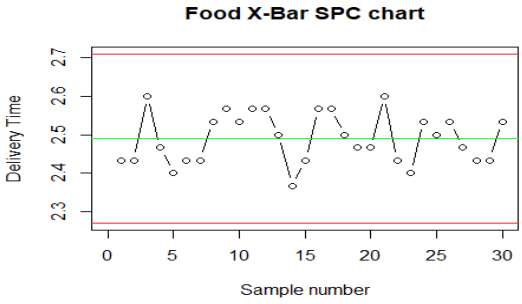


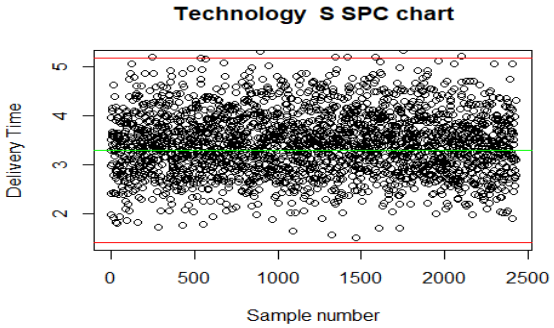


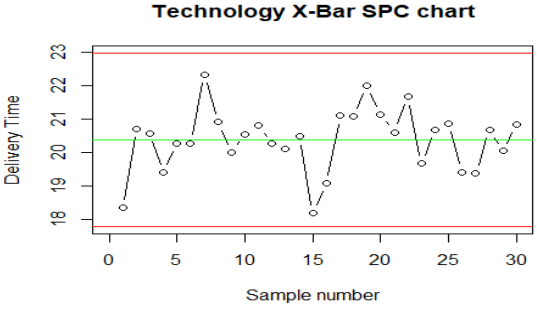


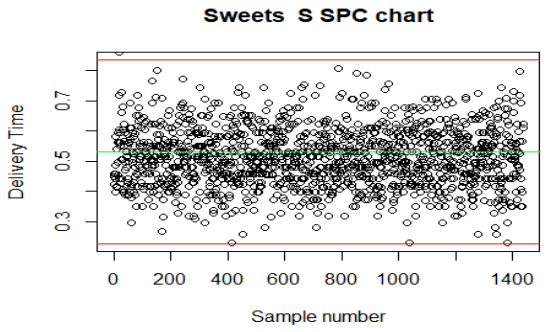


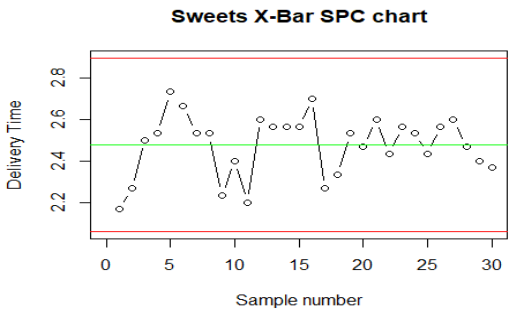


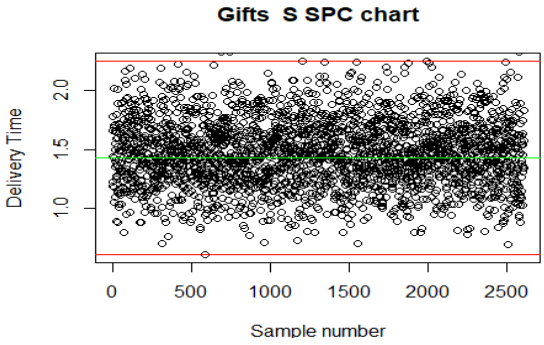


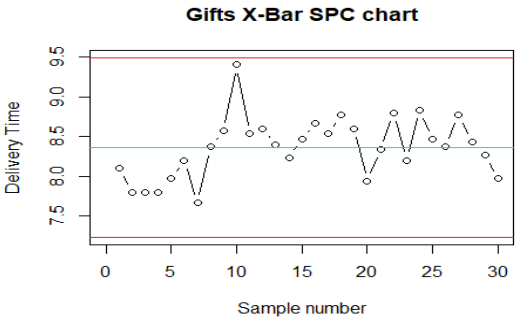


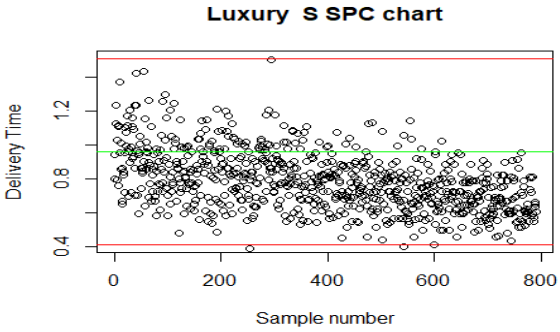


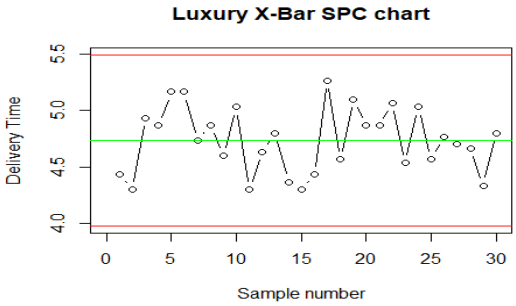

Sweets S SPC chart
Delivery Time
0.3
0.5
0.7
0
200
400
600
800
1000
1400
Sample number


Sweets X-Bar SPC chart
Delivery Time
2.2
2.4
2.6
2.8
0
5
10
15
20
25
30
Sample number


Gifts S SPC chart
Delivery Time
1.0
1.5
2.0
0
500
1000
1500
2000
2500
Sample number


Gifts X-Bar SPC chart
Delivery Time
7.5
8.0
8.5
9.0
9.5
0
5
10
15
20
25
30
Sample number


Luxury S SPC chart
Delivery Time
0.4
0.8
1.2
0
200
400
600
800
Sample number


Luxury X-Bar SPC chart
Delivery Time
4.0
4.5
5.0
5.5
0
5
10
15
20
25
30
Sample number

4.1 (A)

Assuming that my X-Chart is correctly calculated the following table represents the sample means outside the outer control limits. Note that an entry containing NA means that there were less than 6 entries outside the control limits.

Table 10: Sample Means Outside Outer Control Limits

Class	First	Second	Third	Third Last	Second Last	Last	Total number
Clothing	9.433333	9.400000	9.400000	9.466667	9.400000	8.433333	26
Household	42.233333	42.466667	42.500000	57.366667	54.566667	55.800000	406
Food	2.266667	2.733333	2.733333	NA	NA	2.733333	4
Technology	17.500000	17.833333	17.800000	17.600000	17.466667	22.900000	23
Sweets	2.900000	2.066667	3.000000	2.933333	2.900000	2.066667	6
Gifts	10.233333	9.600000	9.900000	16.033333	16.933333	15.700000	2296
Luxury	4.000000	4.000000	3.966667	3.400000	3.266667	3.600000	457

4.1(B)

4.2: Type I Error

Estimate the likelihood of making a Type I Error for A and B A

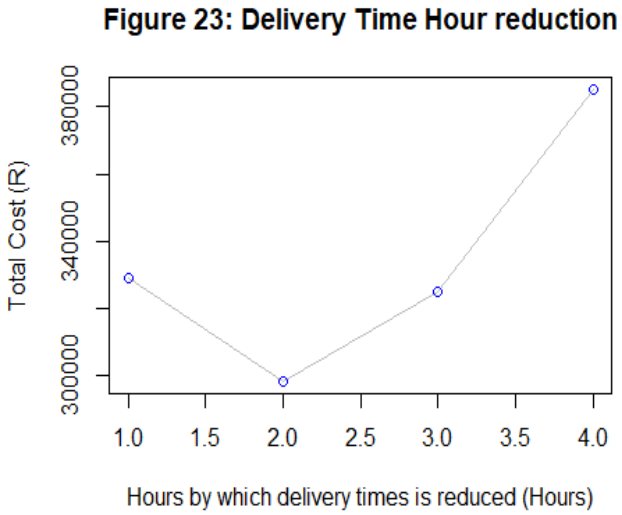
A Type I Error is defined as the probability of rejecting Ho given that the Ho is true. In other word it is a false-positive.

4.3: Centering the Delivery Process

To find the lowest total cost of delivering technology items brute force was initially used to find “x”. “x” is defined as the number of hours the current delivery time is reduced in order to find the lowest possible cost. This was calculated by a for loop for four reduction possibilities. The total costs for each loop were tabulated in Table 11. The mean hours per delivery was included in the table to serve analysts with the complete picture. The total costs for each reduction possibility were plotted in Figure 23 to give a graphical representation of the optimal solution.

Table 11: Delivery Time Cost Analysis

Comment	Values
Initial nr of too slow deliveries	1356
Current Additional Costs (R)	446124
Initial Mean Delivery time	20
Total Costs - 1 hour Reduction (R)	329066
Mean - 1 hour reduction	19
Total Costs - 2 hour Reduction (R)	298206
Mean - 2 hour reduction	18
Total Costs - 3 hour Reduction (R)	324921
Mean - 3 hour reduction	17
Total Costs - 4 hour Reduction (R)	385194
Mean - 4 hour reduction	16



From Table 9 and Figure 23 it is evident that a 2-hour reduction of each delivery time will result in an optimal lowest total cost of R 298 206 whereas the initial additional costs was at R 446124 before any reduction. Furthermore, the plot in Figure 23 reveals similar characteristics to that of the Taguchi loss function. Similar to the Taguchi loss function the most satisfied point is located on the bend of the parabola which when the delivery times is reduced by 2 hours.

4.4: Type II Error

A Type II Error is defined as the probability of accepting Ho given that the Ho is false. In other word it is a false-negative.

PART 5: DOE and MANOVA

5.1 & 5.2

With the use of MANOVA the two plots in Figure 24 & 25 helped to determine and analyse the service and reliability of the individual classes. The main objective of the MANOVA test is to see and compare which feature differs from which. It is set up to predict whether the class of a product has an effect or influence on its price or delivery time. It is however expected that the class of a product does indeed have an influence on the delivery time and price of a product.

```
               Df Pillai approx F num Df den Df    Pr(>F)
Class           6 1.6796   157243     12 359908 < 2.2e-16 ***
Residuals 179954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Below are the results to see which feature differs. The below summary of results reveals that both the Delivery Time and Price features differ, this can be seen by the small P values of both features. It is crucial to note that a p-value below 0.05 indicates that there is a significant effect of treatment on outcome. As can be seen in the below summary of results the p value is smaller than 0. 05 (p-value << 0.05) in both cases. Therefore, we can conclude that the class has an significant influence on both the delivery time and the price of the sold item.

```
 Response Delivery.time :
               Df   Sum Sq Mean Sq F value    Pr(>F)
Class           6 33456906 5576151  629515 < 2.2e-16 ***
Residuals   179954  1594005       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 5.7156e+13 9.5259e+12   80224 < 2.2e-16 ***
Residuals   179954 2.1368e+13 1.1874e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The following two plots in Figure 24 and 25 is a graphical representation of the above test showing the comparison between the different behaviors. It includes a plot for delivery time versus price as well as a delivery time versus class box plot.

Figure 24: Delivery Time vs Price box plots

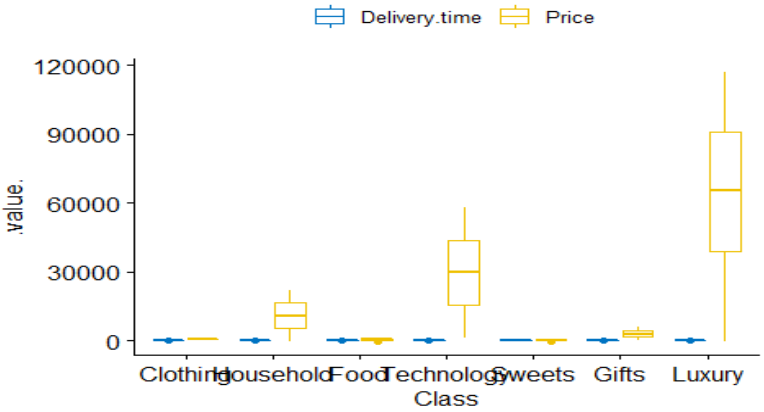
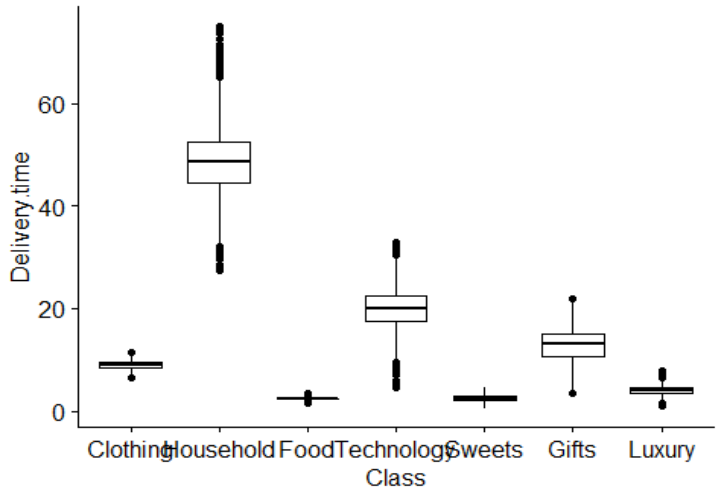


Figure 25: Delivery vs Class box plots



It can be concluded that household, technology, and gifts classes have bigger distributions than the other classes. With regards to service, it is evident that it takes longer to deliver products that are larger or more sensitive (fragile). It makes sense that household products and technology items are often bigger or fragile which will take longer to deliver regardless of the distance. When looking closely at the food class one can note its small distribution and quick delivery times. Small distributions and low delivery time is proof of the good reliability of delivery times.

PART 6

6.1: Problem 6 & 7

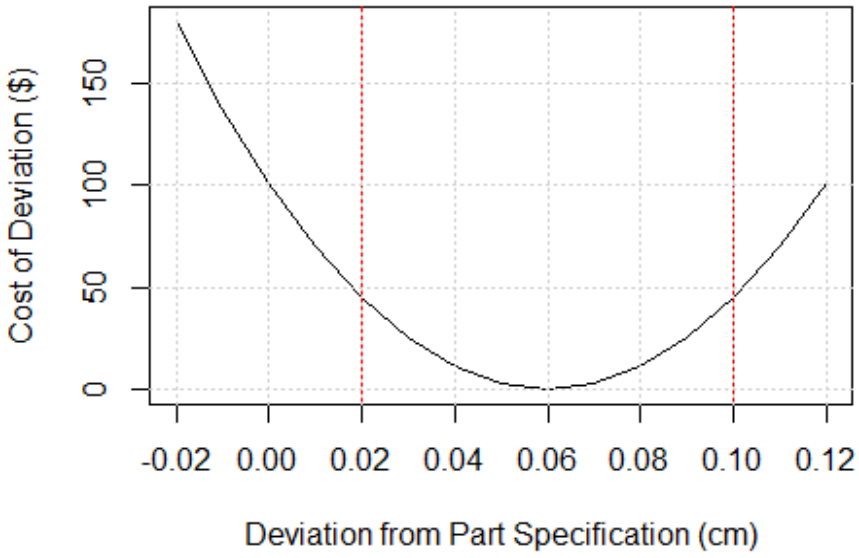
Problem 6

In problem 6 & 7 the Taguchi loss function is determined and calculated as can be seen in the separate R file containing all the R code. The Taguchi loss function is a graphical depiction of loss to describe a phenomenon affecting the value of products produced by a company. The graphs below represent how a slight increase in variation within specification limits can lead to an exponential increase in customer dissatisfaction. The loss function therefore emphasizes the need for incorporating quality and reliability at the design stage, prior to production. The formula for the Tagichi loss function is the following:

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

The loss due to performance variation is proportional to the square of the deviation of the performance characteristics from its nominal value.

Figure 26: Taguchi's loss functon (Problem 6)

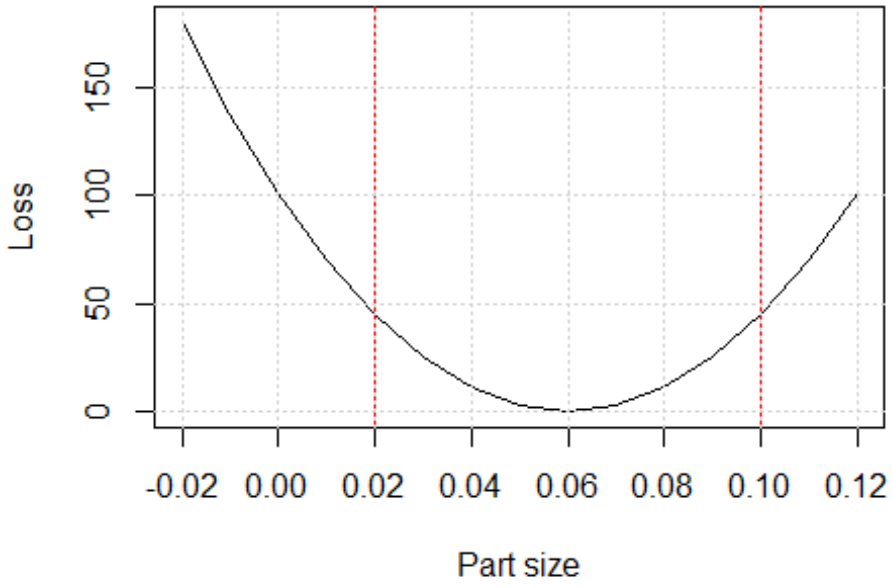


The plot in Figure 12 above indicates the acceptable range for the specifications of a part to deviates to still ensure customer satisfaction. This range is calculated to be between 0.02 to 0.10 cm. Due to the relatively big acceptable specification range the reliability of the product is expected to be quite high as it should be produced according to specifications.

Problem 7.A

Figure 13 & 14 both represent the acceptable part size specification ranges of a part to be between 0.02 and 0.10 cm which is a relatively large rage leaving some room for error. However, the acceptable specification range for both Figures 26 and 27 will still require manufacturing to be precise and accurate. This will lead to the reliability of the product to be similar to that of problem 6.

Figure 27: Taguchi's loss functon (Problem 7.A)



Problem 7.B

The target is reduced to 0.027 cm, leaving the Taguchi loss function to be

$$L(x) = 21875(0.027)^2 = 15.95$$

6.2: Problem 27

27.A

To calculate the system reliability for the system, assuming one machine at each stage the following formula was used:

$$System\ Reliability = P(A) * P(B) * P(B) = (0.85)(0.9)(0.90) = 0.7038$$

With one machine at each stage, it means the three machines are in series with each other. This can serve as a problem as if one machine broke the other machine will not be able to work as well.

27.B

Two machines in parallel at each stage increase the system reliability drastically. The following formula calculates the total reliability:

$$\begin{aligned} System\ Reliability &= (1-(1-P(A))\text{^}2) * (1-(1-P(B))\text{^}2) * (1-\big(1-P(C)\big)\ \text{^} \\ &= (1-(0.15)\text{^}2) * (1-(0.08))\text{^}2) * (1-(0.10))\text{^}2) \\ &= 0.9615316 \end{aligned}$$

With the machines parallel to each other it ensures that if one machine is to break down, the other machines will still be able to function, ensuring reliability and continuity. Therefore, it is evident that reliability increases from 0.704 to 0.962. Reliability therefore increased by 36.6 %.

6.3: Vehicles

For this problem I followed a logical approach and did not consider the fact that it is a binomial distribution. Therefore, the following method was followed:

Probability of vehicles still providing a reliable service:

$$P(Vehicles) = \frac{(1560-3-1)}{1560}$$

Probability of drivers still providing a reliable service:

$$P(Drivers) = \frac{(1560-1)}{1560}$$

$$Total\ Reliability = P(Vehicles)P(Drivers)$$

$$\begin{aligned} Reliable\ Days\ per\ Year &= (Total\ Reliability)\ (365) \\ &= (0.9967965)(365) \\ &= 363.8307 \\ &= 363\ days \quad \text{(Rounded down)} \end{aligned}$$

Therefore, it can be concluded that there will be 363 days of reliable service per year.

Conclusion

In conclusion, the company can look to promote the sales of luxury and technology items as their sales prices are the highest and that will increase the company's overall revenue. Decreasing each delivery time by 2 hours will be the best managerial decision as it will result in an optimally lowest total cost of R 298 206, saving the company a lot of money in the long run. Although this decision will initially serve the company extra additional costs the business will save on late delivery-related costs. Furthermore, the company can look to increase advertisements by website as most of the sales are currently being advertised by word of mouth (recommended). In general, the performance of the company is relatively good and will continue to perform for the foreseeable future.

References

Taguchi Loss Function, Lean Manufacturing and Six Sigma Definitions. Available at: https://www.leansixsigmadefinition.com/glossary/taguchi-loss-function/ (Accessed: October 20, 2022).

Team, V.C. (2022) How do you calculate type 1 error and type 2 error probabilities?, How do you calculate type 1 error and type 2 error class 11 maths CBSE. Available at: https://www.vedantu.com/question-answer/calculate-type-1-error-and-type-2-error-class-11-maths-cbse-6010bf52962db9208d20f4e5 (Accessed: October 20, 2022).

Type II error in hypothesis testing with R programming (2021) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/type-ii-error-in-hypothesis-testing-with-r-programming/. (Accessed: October 20, 2022).

Type II error in hypothesis testing with R programming (2021) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/type-ii-error-in-hypothesis-testing-with-r-programming/. (Accessed: October 20, 2022).

Banerjee, A. et al. (2009) Hypothesis testing, type I and type II errors, Industrial psychiatry journal. Medknow Publications. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996198/#:~:text=A%20type%20I%20error%20(false,actually%20false%20in%20the%20population. (Accessed: October 20, 2022).

Korstanje, J. (2021) Manova, Medium. Towards Data Science. Available at: https://towardsdatascience.com/manova-97e675a96158#:~:text=The%20test%20statistic%20in%20MANOVA,effect%20of%20treatment%20on%20outcome. (Accessed: October 20, 2022).