Department of Industrial Engineering, Stellenbosch University

Quality Assurance 344

Project

Prepared by: A van der Merwe, 23552107@sun.ac.za
10-21-2022

Abstract

This report looks at the client data of an online business. The data is cleaned and evaluated to draw conclusions from it.

Data wrangling is done to clean the data. Descriptive statistics draws meaningful information from the data that can be used in further in-depth evaluations. Statistical process control looks at samples of the different classes, where it can be seen which classes' processes should be inspected. Optimisation gives a more in-depth look at which samples of delivery times are out of control. Then, manova tests evaluate certain conclusions drawn from the previous parts. Lastly, system reliability is evaluated.

Contents

Introduction

Client data for an online business is given, and it must be analysed. The statistics and R programming learned in QA 344 is used to evaluate the online business.

Firstly, data wrangling needs to be done to clean the data, then descriptive statistics will be done to analyse the data. After that, statistical process control is used to evaluate if the process is in control and calculations will also be done to optimise the delivery process. Manova tests will be done to evaluate the inconsistencies found in previous parts of the report. Lastly, the reliability of services and products will be evaluated.

Part 1: Data wrangling

Before the data can be used to draw conclusions, the data must be cleaned. To split the data into valid and invalid data, we must evaluate the data first. We split the features into continuous and categorical to evaluate them separately.

Continuous features

The continuous features are age, price, month, day and delivery time. They are evaluated on the minimum, 1st quantile, median, mean, 3rd quantile, maximum and missing values. The table below provides a summary.

Feature	Min.	Q1	Median	Mean	Q3	Max.	Missing values
Age	18.0	38.00	53.00	54.57	70.00	108	0
Price	-588.8	482.31	2,259.63	12,293.7	15,270.7	116,619	17
Month	1.0	4.00	7.00	6.52	10.00	12	0
Day	1.0	8.00	16.00	15.54	23.00	30	0
Delivery time	0.5	3.00	10.00	14.50	18.50	75	0

From the table, 2 problems can be seen. The first is that price has a negative minimum and the second is that price has 17 missing values. Both are classified as invalid data. Therefore, all the negative and missing values should be removed from price.

Categorical features

The categorical features are class and why bought. They are evaluated by looking at the mode, 2nd mode and if there are any missing instances.

Class

The table below shows the number of instances in every class. Gifts are the mode of the class feature and technology is the second mode. There is also no missing features.

Sweets	Household	Gifts	Technology	Luxury	Food	Clothing
21,566	20,067	39,154	36,350	11,869	24,588	26,406

Why bought

The table below shows the number of instances for every reason of buying. Recommended is the mode and website is the 2nd mode. There are also no missing features.

Recommended	Website	Random	Browsing	EMail	Spam
107,000	29,450	13,122	18,995	7,225	4,208

Conclusion

The only problem with the data is that the price feature has negative and missing values. They should be removed to obtain the valid data.

Part 2: Descriptive Statistics

Process capability indices:

These indices are calculated for the technology class items. Since we have no specified limits, these calculations are done on assuming that the USL= 24 hours and LSL= 0 hours. An LSL of 0 is logical because the delivery time (in hours) cannot be below zero and it is the smallest possible time in which a delivery can be made.

Cp	Cpu	Cpl	Cpk
1.142	0.3797	1.905	0.3797

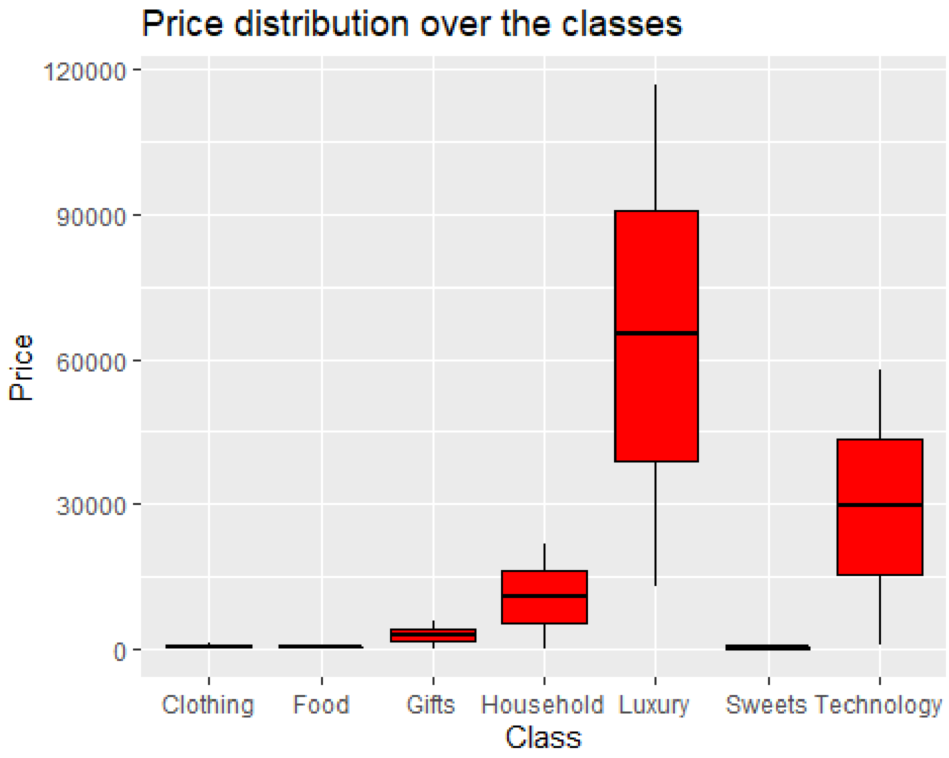
Cp is capability potential, a Cp value between 1 and 1.33 means that the process is barely capable of meeting delivery times below 24 hours. Cpk is process capability index. Cpk is the minimum between Cpl and Cpu. Cp and cpk is used to define the ability of a process to produce a product that meets the requirements. Cpl is a measure of the capability of the process based on its lower limit. Since the lower limit is zero, the Cpl is very large because there are no delivery times below zero. Cpu is process capability base on upper limit. When looking at the data, there is a great amount of delivery times above 24 hours, explaining why the Cpl is 0.3797.

Data Analysis

Firstly, we look at the number of rows and columns in the data frame.

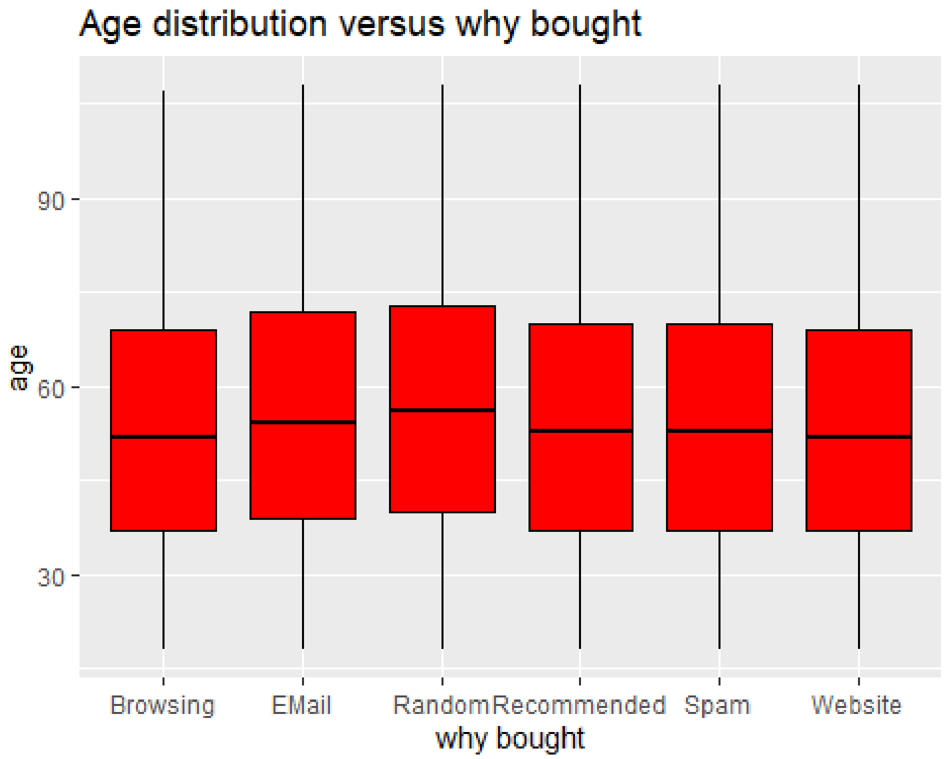
Nr. of rows	Nr. of columns
179,978	10

Categorical features

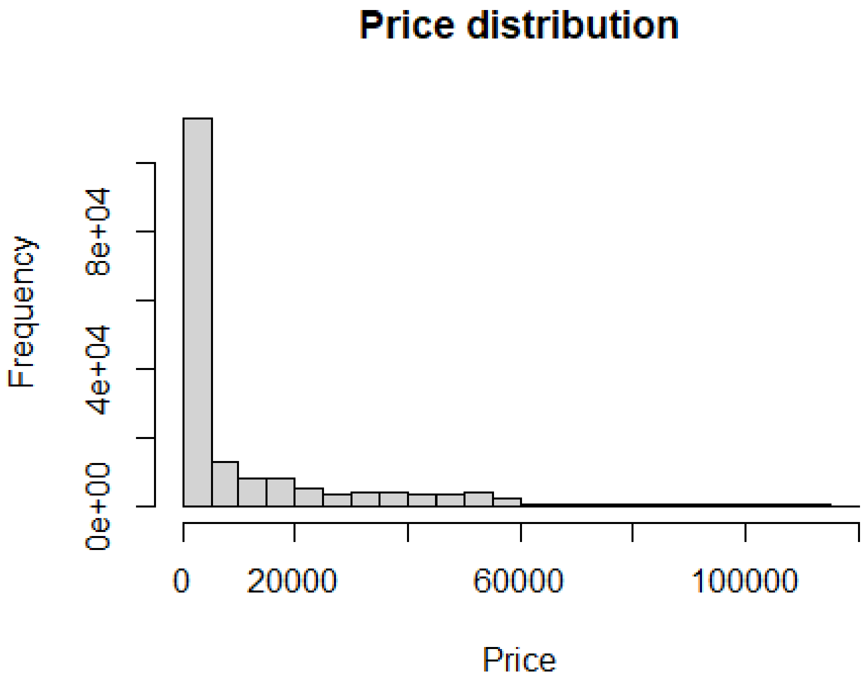


For Technology, luxury, household, clothing, and gifts classes the distribution is normal. Food and sweets are skewed to the right.

Luxury has the greatest distribution (the longest box plot), and the highest prices. It is followed by technology, class, and gifts (in highest distribution and prices).



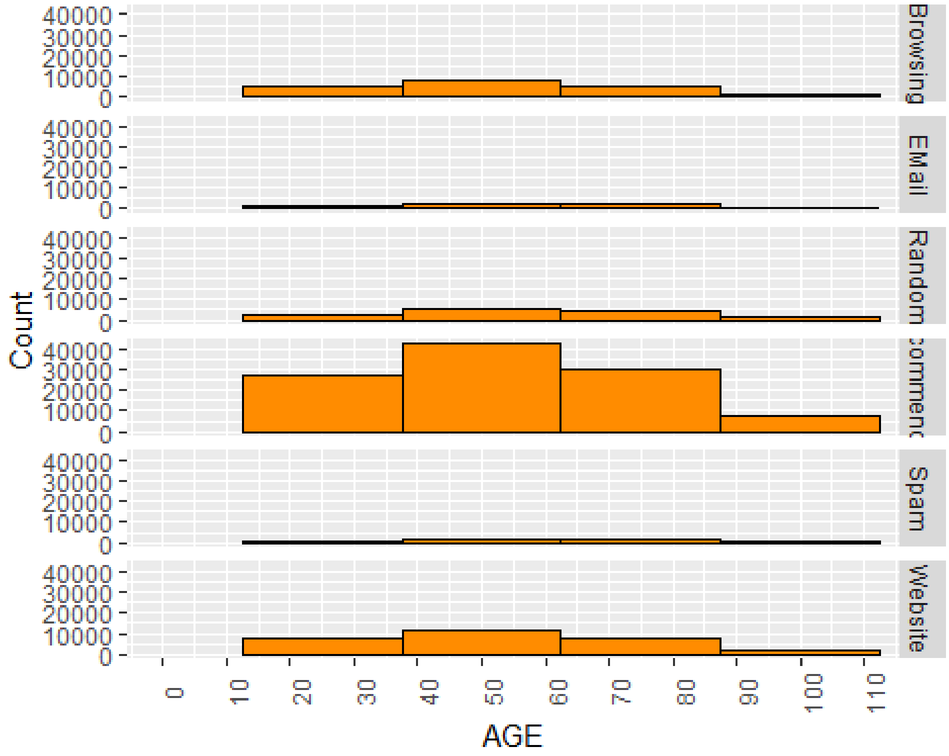
We would think that younger people would buy more because of browsing or websites, but in this distribution, it can be seen that the age is distributed the same over all reasons for buying.



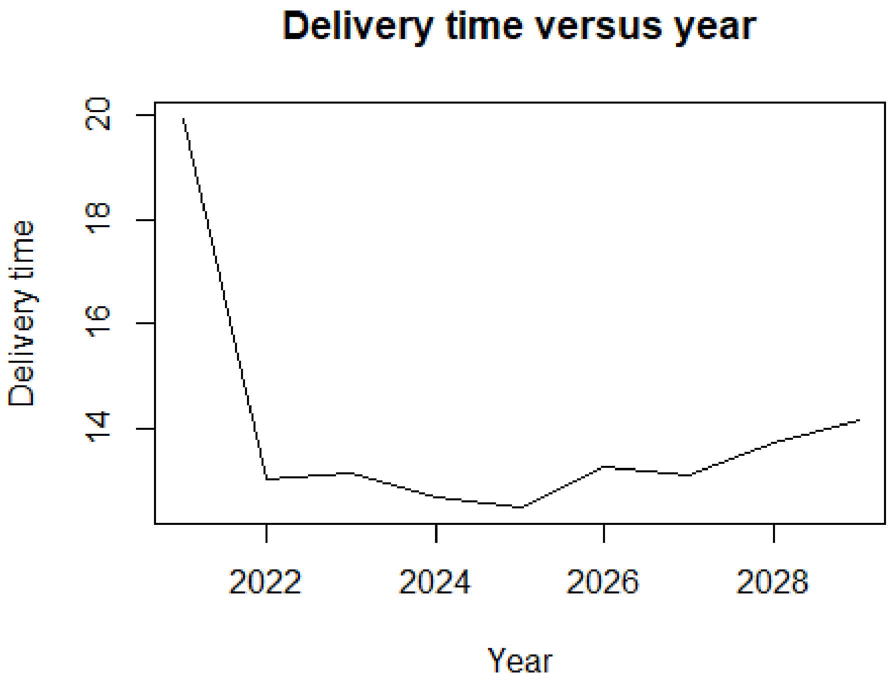
This histogram shows that a lot of items are priced below R 5000 and that there are very few items that are priced above R 60 000. When evaluating those two extremes:

It is seen that all the sales of Sweets, food, and clothing fall below the R 5 000 threshold. As expected, all the items priced above R 60 000 are luxury items.

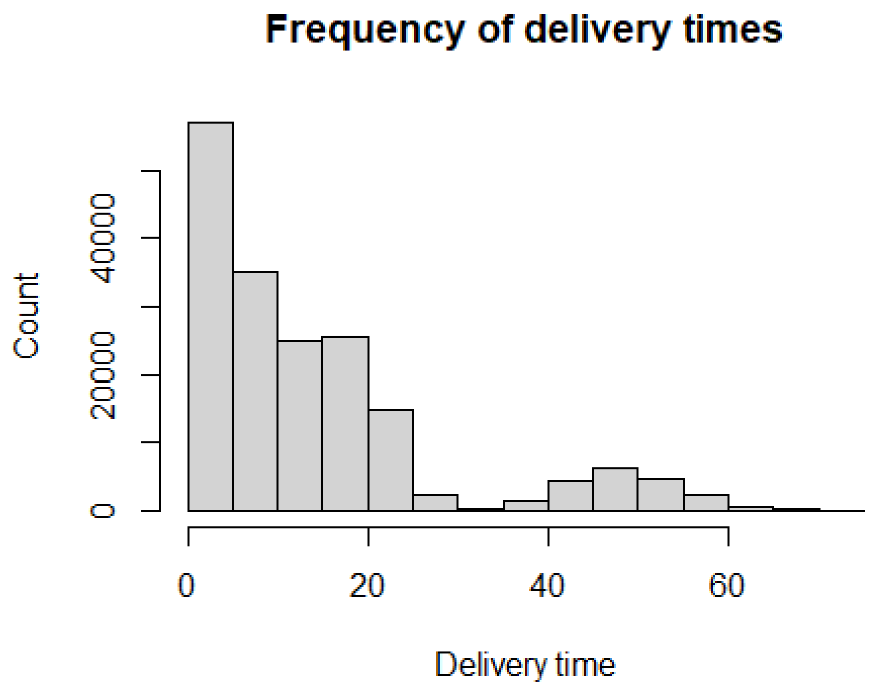
Continuous features:



All the histograms above show the reason for buying versus the age. All the graphs have the same distribution, excepts spam and random. This implies that the reason for buying does not vary with age.



Delivery time decreases significantly from 2021 to 2022 and then becomes mostly converged with small fluctuations from 2022 until 2029.



The delivery times have a bimodal distribution, indicating that there are 2 different groups. The frequency of the deliveries between 40 and 60 hours are far less than the number of deliveries between 0 and 30 hours. The first peak is skewed to the left, indicating that there are more deliveries that take less time.

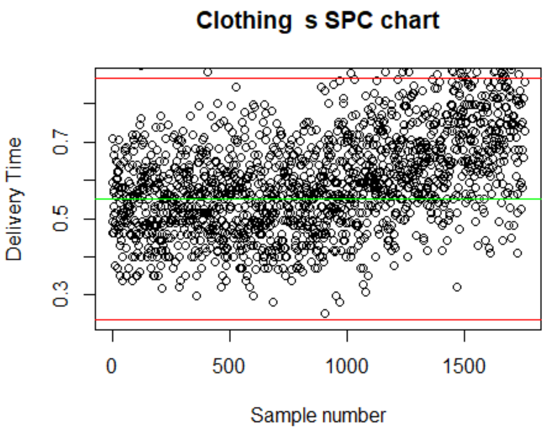
Part 3: Statistical process control

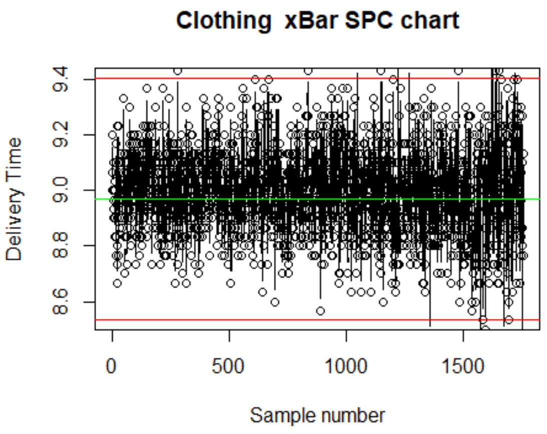
30 samples of 15 instances each are drawn from each class of sales to find the variables for statistical process control. These variables are based on the delivery times. The first table is for x-charts and based on the mean and the second table for s-charts, based on standard deviation.

class	UCL	U1 SIGMA	U2 SIGMA	CL	L1 SIGMA	L2 SIGMA	LCL
Clothing	9.40	9.11	9.04	8.97	8.83	8.90	8.54
Food	2.71	2.56	2.53	2.49	2.42	2.45	2.27
Luxury	5.49	4.99	4.86	4.74	4.48	4.61	3.98
Technology	22.97	21.24	20.81	20.37	19.51	19.94	17.77
Gifts	9.49	8.74	8.55	8.36	7.99	8.17	7.23
Household	50.25	47.79	47.18	46.56	45.33	45.95	42.88
Sweets	2.90	2.62	2.55	2.48	2.34	2.41	2.06

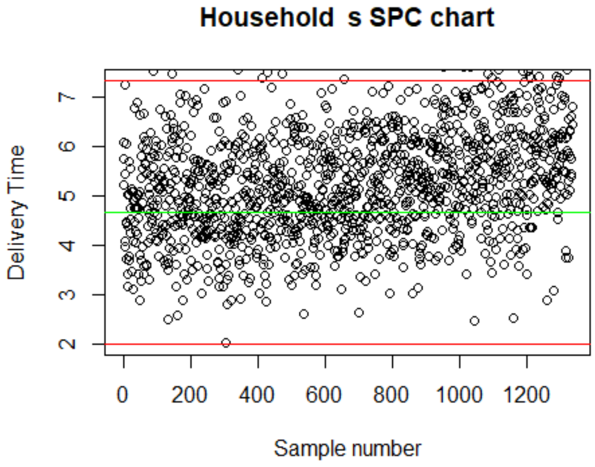
class	UCL	U1SIGMA	U2SIGMA	CL	L1SIGMA	L2SIGMA	LCL
Clothing	0.867	0.656	0.604	0.551	0.446	0.499	0.236
Food	0.437	0.331	0.305	0.278	0.225	0.252	0.119
Luxury	1.51	1.14	1.05	0.961	0.778	0.870	0.411
Technology	5.18	3.92	3.61	3.30	2.67	2.98	1.41
Gifts	2.25	1.70	1.57	1.429	1.16	1.29	0.612
Household	7.34	5.56	5.12	4.67	3.78	4.23	1.9999
Sweets	0.84	0.633	0.582	0.531	0.430	0.481	0.227

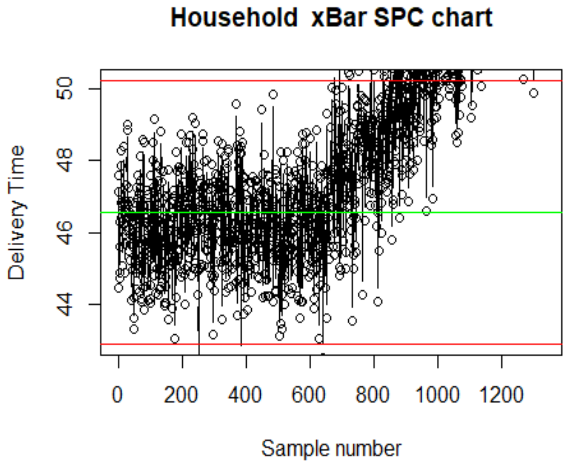
The sample standard deviation means of the delivery times are plotted for every class on the s SPC charts below and the sample means of the delivery times are plotted on the x-Bar SPC charts below.



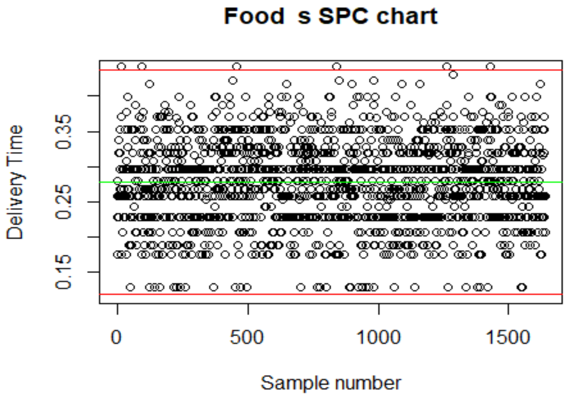


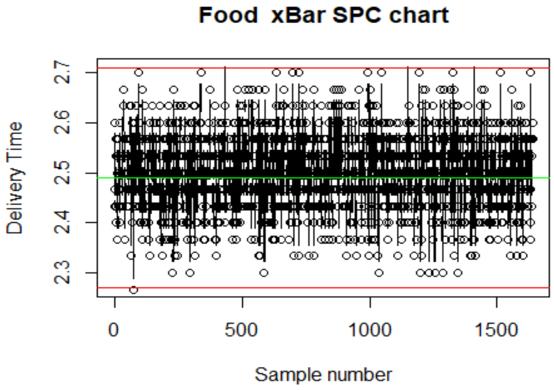
The x-Bar chart shows that the mean delivery times for clothing items remained relatively consistent over the years, while the s-chart shows outliers above the upper control limit from sample 1000 onwards that should be investigated.



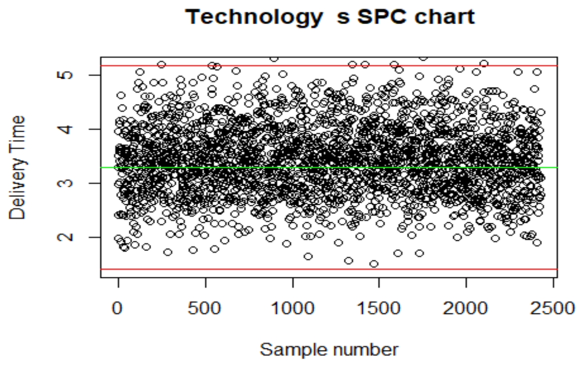


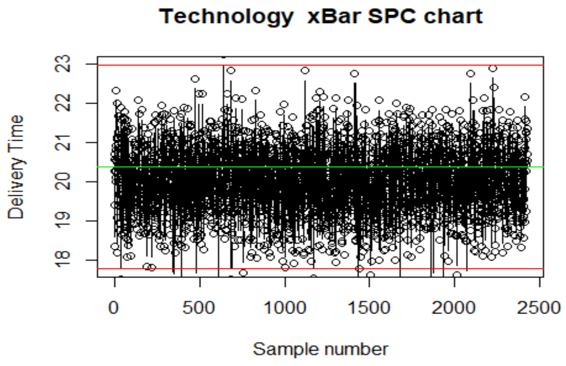
For the household class, the x-bar chart shows a significant rise in delivery time over the years. It would be expected that managers investigated the longer delivery times, but from the graph it is evident that is wat not detected or investigated.



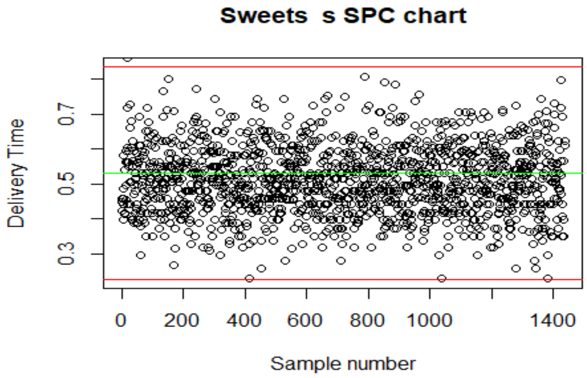


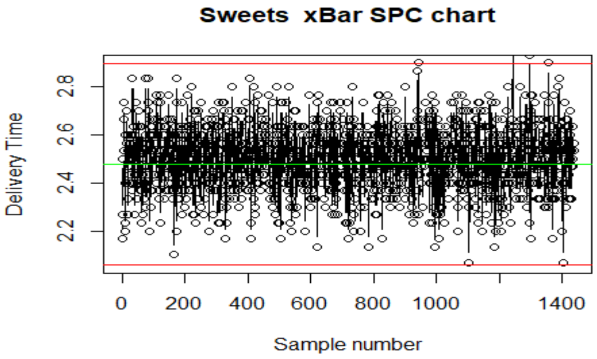
The x-chart for the food class shows that the process delivery times are in control, while some deviations above the upper control limit on the s-chart can be seen. This could be normal deviation in delivery times, since there are only a few outliers.



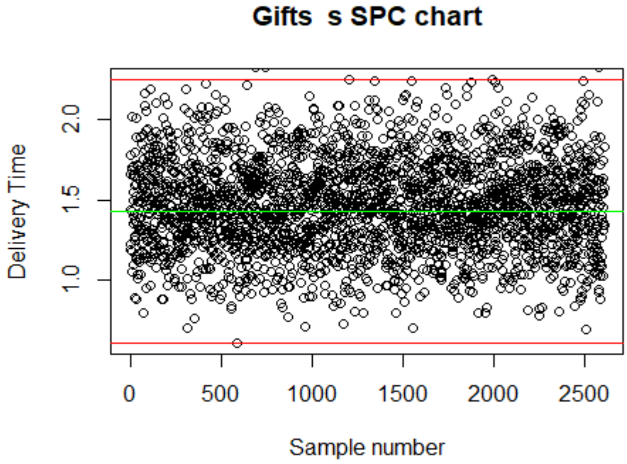


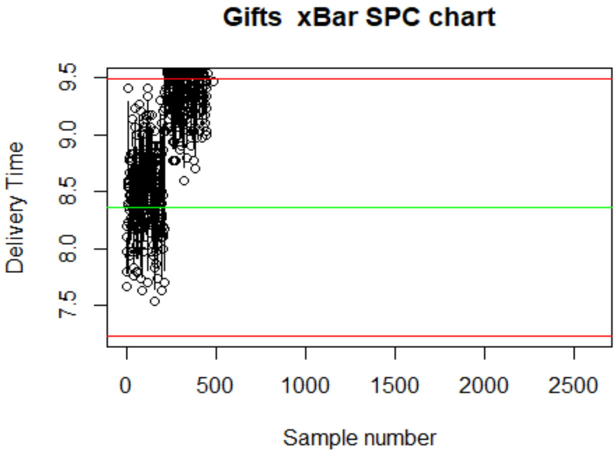
The x- and s-chart for technology shows the process is in control with some normal outliers above the upper control limit on the s-chart.



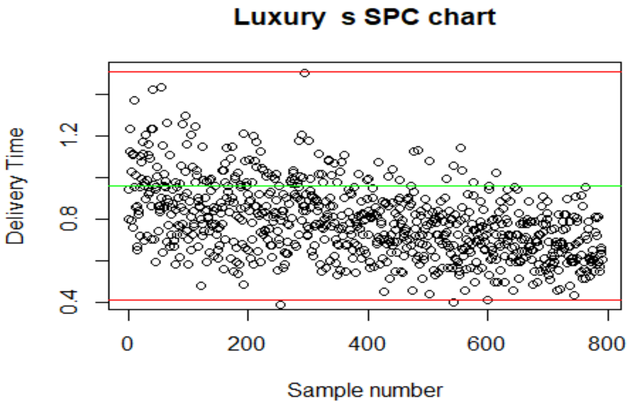


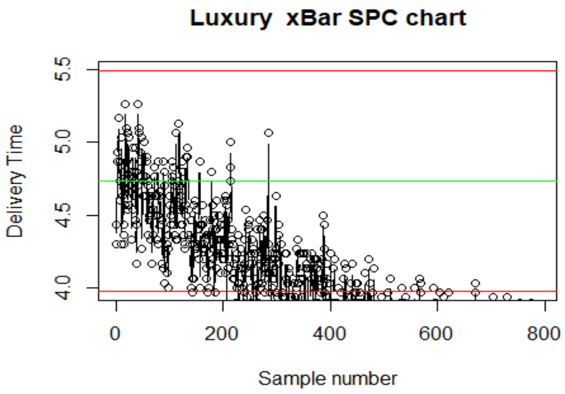
For sweets the process in in control. There are so few samples lying outside of the limits.





For the gifts class the process is completely out of control. The delivery time means increased drastically and were never investigated or picked up by management. The fact that the standard deviation is constant means that the delivery time means just kept going up every year.



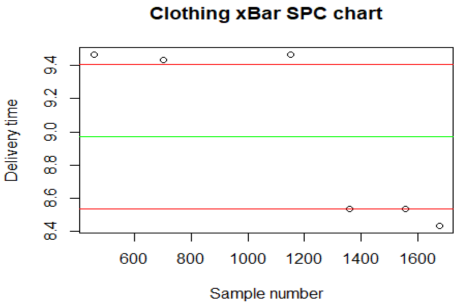


The mean delivery time of luxury items decreased significantly. This could be an admin problem, meaning that delivery times were incorrectly logged. This should have been detected and investigated by management, but the s-chart shows consistent standard deviation meaning that the mean delivery times just kept decreasing every year.

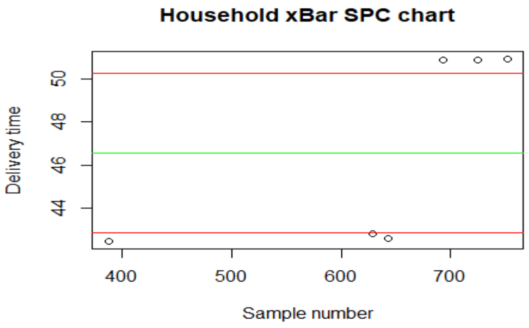
Part 4: Optimising the delivery process

Samples that give indications of out-of-control processes

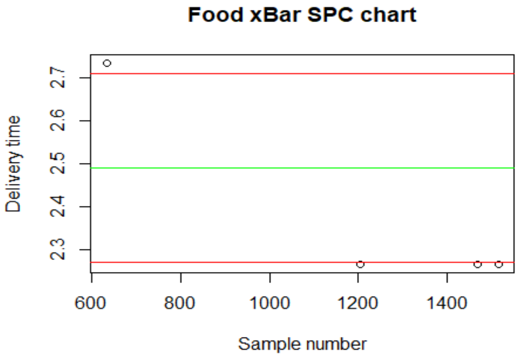
We are looking at every class to see where to out of control samples of the delivery times lie on the x-Bar SPC chart. In some cases, there were more than 6 out of control samples. The plots below only show the first 3 samples above the UCL and last 3 samples below the LCL.



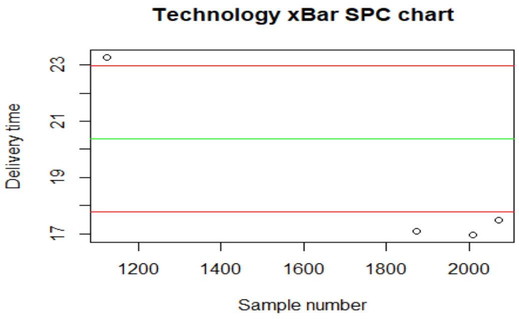
The clothing class has samples that are above the upper control limit and samples that are below the lower control limit.



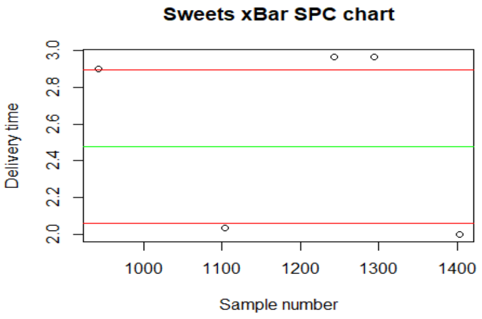
The household class has samples that are all above the upper control limit and lower control limit. There are more samples in frequent years that are out of control. This samples should be inspected.



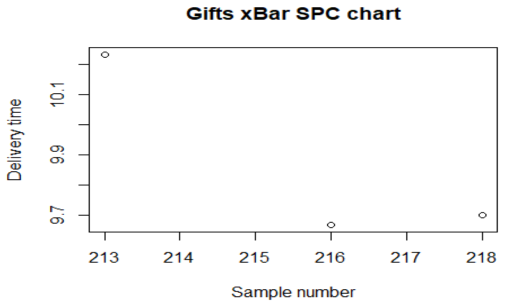
There are a few out of control samples in the food class that are just below the LCL.



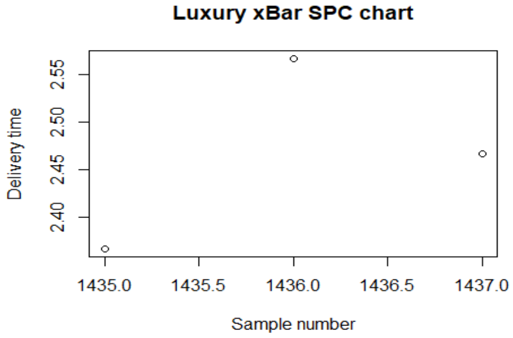
The technology class has samples above and below the upper and lower control limits. The samples below the LCL should be inspected as they are far under the LCL.



The out-of-control samples for sweets are above the UCL and below the LCL.



The out-of-control samples for gifts are all below the lower control limit of 9.488. This could indicate that the delivery times were incorrectly processed or that some workers are doing their job wrong.



The luxury class only has samples that are below the LCL of 3.977.

Most consecutive sample standard deviations

The probability that the delivery times are between -0.3 and 0.4 is slim. This could indicate and out of control process. We look at the most consecutive sample delivery times of each class to see if there is a problem.

Class	Most consecutive samples	Index of last sample
Clothing	27	451
Household	14	195
Food	26	176
Technology	17	2108
Sweets	24	1047
Gifts	27	37
Luxury	0	0

For the clothing and gifts classes 27 is the most consecutive samples, but the index is quite low, meaning the problem was resolved and that there is no issue anymore. Technology and sweets have 17 and 24 consecutive samples, respectively, but their indexes are high, meaning this variability was quite recent and should be inspected.

Likelihood of making a type I error

The probability of making a manufacturers (type 1) error is the probability of rejecting Ho when Ho is not true. This means thinking the process is out of control when it is in fact in control.

For A it would be the probability that the means fall outside of the control limits when they are in fact inside the control limits.

For B it would be that there are many consecutive delivery times between -0.3 and 0.4, when they are not within those bounds.

The probability for both is always the same and that probability is 0.0026998.

How to centre the process for the best profit of technology class

The delivery process can be improved to be faster or adapted to be slower. Both options come with certain costs. Here we evaluate how to centre the technology delivery process for the best profit. A brute force method was used, and every option evaluated.

Movement	Cost
0 hours	R 849 549
-1 hours	R 461 000
-2 hours	R 349 704.5
-3 hours	R 340 892.5
-4 hours	R 387 517

From the table it is evident that the process should be centred around -3 hours for the minimum cost.

Probability of making a type II error

The probability of a type 2 error is failing to reject Ho when Ho is false. This means thinking the process is in control, when, in fact it is not.

For A it would be the probability that the sample means are within the outer control limits, when they are in fact not.

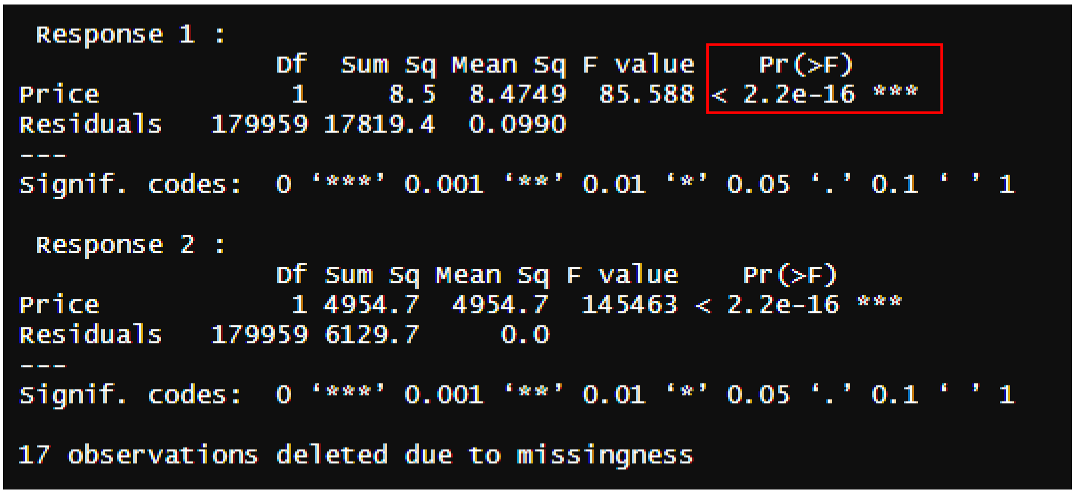
For B it would be the probability that there are no large consecutive delivery times within the -0.3 and 0.4 bounds, when in fact there are large consecutive delivery times within that bound.

can only be made when the process is not in control. not at centre, or variation has become too large. This probability is 0.4883177.

Manova tests

Test 1

After evaluating all the results in the report so far, a manova test will be done to evaluate the price difference between the household and luxury classes.

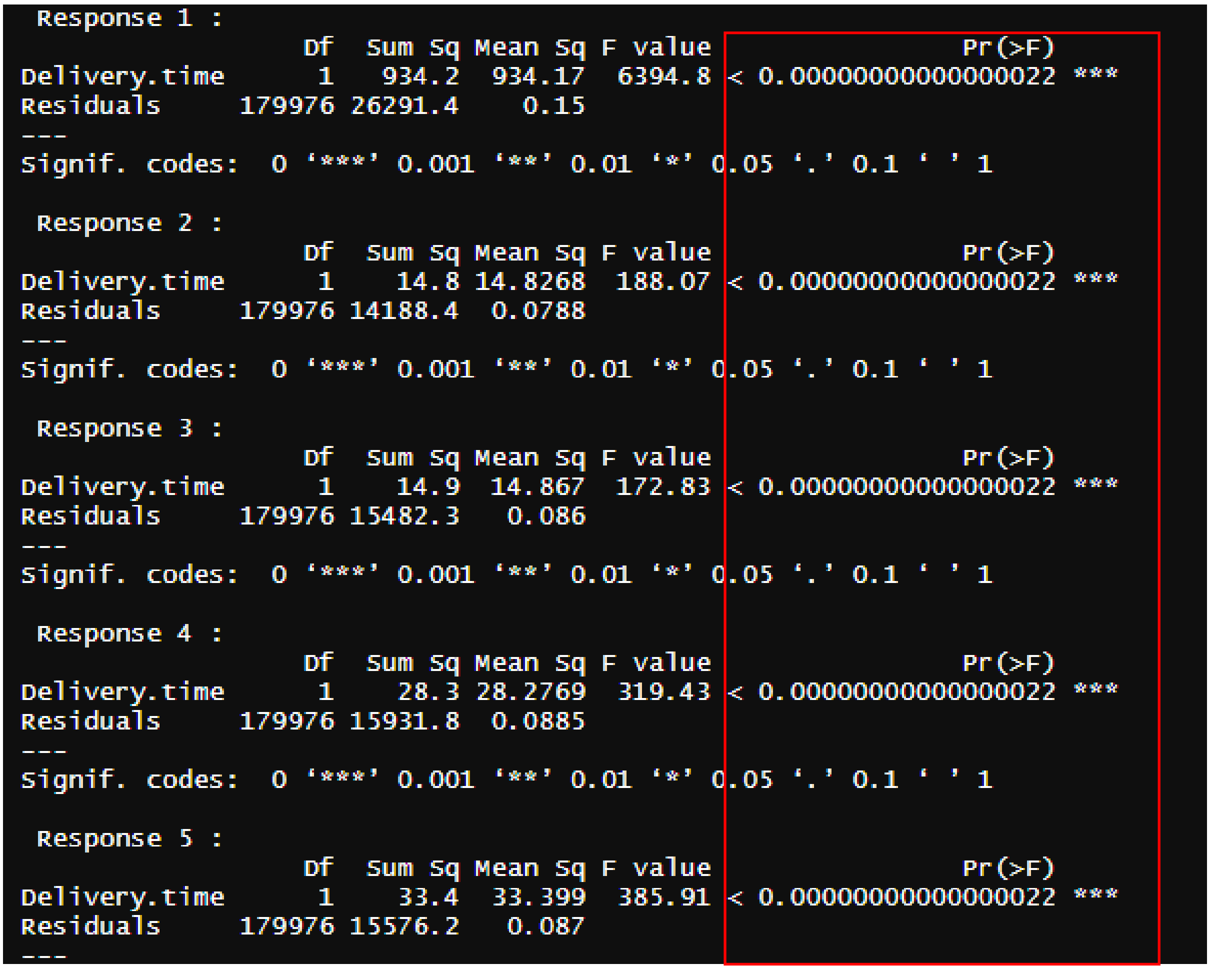


```
 Response 1 :
                 Df  Sum Sq Mean Sq F value    Pr(>F)    
Price             1     8.5  8.4749  85.588 < 2.2e-16 ***
Residuals    179959 17819.4  0.0990                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 2 :
                 Df Sum Sq Mean Sq F value    Pr(>F)    
Price             1 4954.7  4954.7  145463 < 2.2e-16 ***
Residuals    179959 6129.7     0.0                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

17 observations deleted due to missingness
```

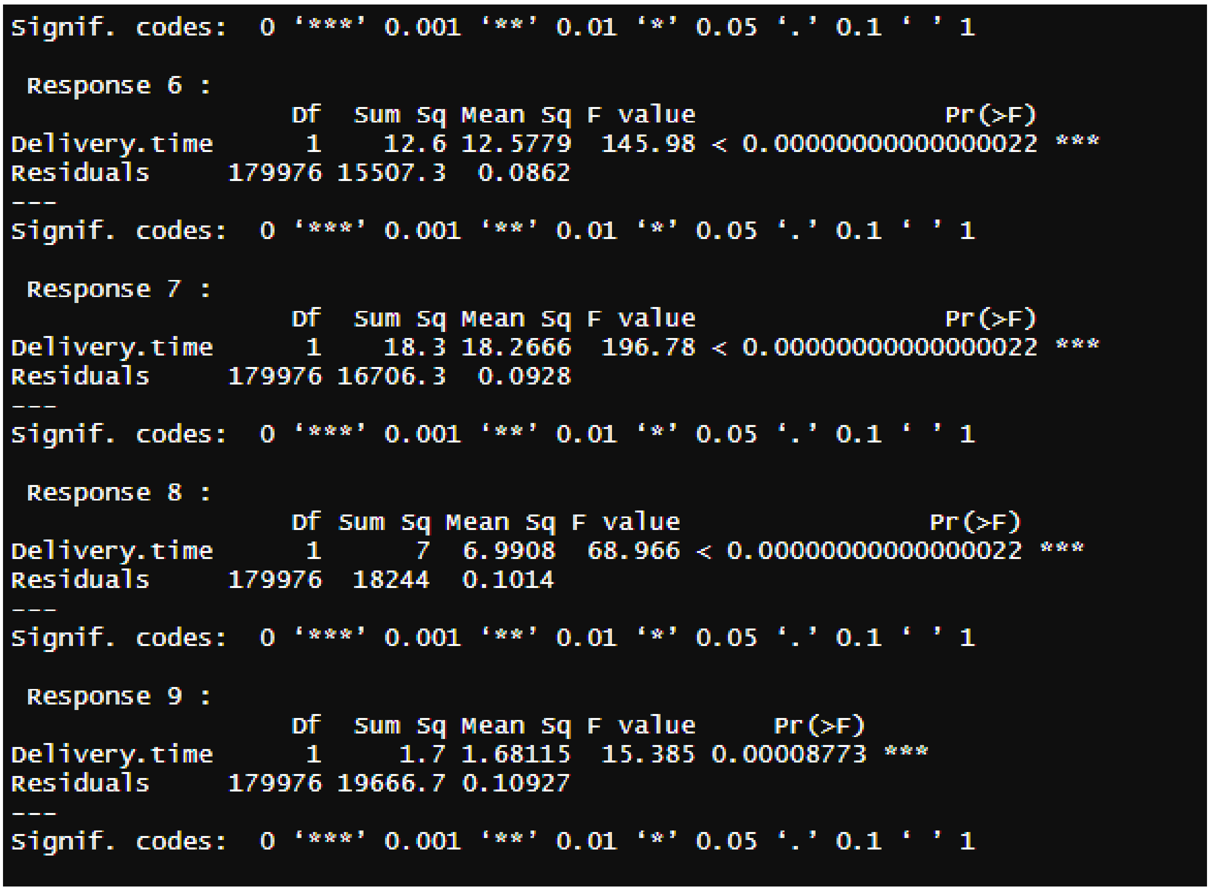
The P value is very small meaning that the null hypothesis is rejected. In this case the null hypothesis is that the process is in control.

Test 2

The delivery times in the different years (2021-2029) are evaluated.



```
 Response 1 :
                  Df  Sum Sq Mean Sq F value                Pr(>F)    
Delivery.time      1   934.2  934.17  6394.8 < 0.00000000000000022 ***
Residuals     179976 26291.4    0.15                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 2 :
                  Df  Sum Sq Mean Sq F value                Pr(>F)    
Delivery.time      1    14.8 14.8268  188.07 < 0.00000000000000022 ***
Residuals     179976 14188.4  0.0788                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 3 :
                  Df  Sum Sq Mean Sq F value                Pr(>F)    
Delivery.time      1    14.9  14.867  172.83 < 0.00000000000000022 ***
Residuals     179976 15482.3   0.086                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 4 :
                  Df  Sum Sq Mean Sq F value                Pr(>F)    
Delivery.time      1    28.3 28.2769  319.43 < 0.00000000000000022 ***
Residuals     179976 15931.8  0.0885                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 5 :
                  Df  Sum Sq Mean Sq F value                Pr(>F)    
Delivery.time      1    33.4  33.399  385.91 < 0.00000000000000022 ***
Residuals     179976 15576.2   0.087                                  
---
```



```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 6 :
                  Df  Sum Sq Mean Sq F value                Pr(>F)    
Delivery.time      1    12.6 12.5779  145.98 < 0.00000000000000022 ***
Residuals     179976 15507.3  0.0862                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 7 :
                  Df  Sum Sq Mean Sq F value                Pr(>F)    
Delivery.time      1    18.3 18.2666  196.78 < 0.00000000000000022 ***
Residuals     179976 16706.3  0.0928                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 8 :
                  Df Sum Sq Mean Sq F value                Pr(>F)    
Delivery.time      1      7  6.9908  68.966 < 0.00000000000000022 ***
Residuals     179976  18244  0.1014                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 9 :
                  Df  Sum Sq Mean Sq F value     Pr(>F)    
Delivery.time      1     1.7 1.68115  15.385 0.00008773 ***
Residuals     179976 19666.7 0.10927                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For this test, we also look at the P- value. Again, the p-value is very small, indicating that the null hypothesis should be rejected. The process is not in control.

Part 6: Reliability of the service and products

6.1 Reliability of service and products

Problem 6 page 359

Cost to scrap: \$45

Specification: $0.06 \pm 0.04\ cm$

Taguchi loss function: $L(x) = k(x - T)^2$

$= 28\,125(x - T)^2$

Problem 7 page 359

(a) Cost to scrap: \$35
Specification: $0.06 \pm 0.04\ cm$
Taguchi loss function: $L(x) = k(x - T)^2$
$= 21\,875(x - T)^2$

(b) Process deviation reduced to $0.027\ cm$
$L(x) = 21875(0.027 - 0.04)^2$
$= \$\ 3.70$

6.2 System reliability

(a) if only one machine works at each stage, the machines are in series.

$R_s = 0.85\ \times 0.92\ \times 0.9$

$= 70.38\%\ Reliability$

(b) If 2 machines are at each stage, The machines are in parallel:

$R_p = (1-(1-0.85)) \times (1-(1-0.92)) \times (1-(1-0.9))$

$= 99.88\%\ Reliability$

Reliability is drastically improved from series to parallel.

6.3 Binomial properties

We are trying to estimate how many days of the year we can expect reliable delivery times.

Probabilities that the number of vehicles are available:

Weighted p value: $= 0.007052725$

Probability that the number of vehicles fail in days:

$P(0\ fail) = 0.8619$

$Expected\ number\ of\ days: 1344.5$

$P(1\ fail) = 0.1285583$

$Expected\ number\ of\ days: 200.6$

$P(2\ fail) = 0.0091$

$Expected\ number\ of\ days: 14.24$

$P(3\ fail) = 0.0004$

$Expected\ number\ of\ days: 0.6408$

$P(4\ fail) = 0.0000131$

$Expected\ number\ of\ days: 0.0204815$

Expected percentage of reliable days: 99.96%

Expected number of days: 364.84

Probability that drivers are available:

Weighted P: 0.00322

$P(0\ fail) = 0.934$

$Expected\ number\ of\ days: 1457.79$

$P(1\ fail) = 0.0634$

$Expected\ number\ of\ days: 98.94$

$P(2\ fail) = 0.0020$

$Expected\ number\ of\ days: 3.198$

$P(3\ fail) = 0.000042$

$Expected\ number\ of\ days: 0.0654$

Expected percentage of reliable days: 99.99%

Expected number of days: 364.98

<u>Vehicles and drivers:</u>

$$P(Reliable) = P(Vehicles\ reliable) \times P(Drivers\ reliable)$$

$$= 0.9996 \times 0.9999$$

$$= 99.95\%\ Reliability$$

This equates to 364.83 reliable days.

When the question changes to 22 vehicles and 21 drivers:

<u>Reliability of vehicles:</u>

$P(0\ fail) = 0.8558$

$Expected\ number\ of\ days: 1335.06$

$P(1\ fail) = 0.1337$

$Expected\ number\ of\ days: 208.619$

$P(2\ fail) = 0.00997$

$Expected\ number\ of\ days: 15.5587$

$P(3\ fail) = 0.00047$

$Expected\ number\ of\ days: 0.7367$

<u>Reliability of drivers:</u>

The same as in the previous part.

<u>Vehicles and drivers:</u>

$P(Reliable) = P(Vehicles\ reliable) \times P(Drivers\ reliable)$

$$= 0.9998 \times 0.9999$$

$$= 99.99\%\ Reliability$$

This equates to 364.97 reliable days. Thus, if we increase the number of vehicles by one, the number of reliable days stay the same.

Conclusion:

This report evaluates client data of an online business. Data wrangling and data analysis, manova tests, statistical process control, reliability of system and products, optimisation, is done.

Data wrangling indicated missing features. These missing features are removed. Data analysis is done to inspect data and retrieve meaningful information. This information includes the discovery that the price distribution of the different classes varies greatly and that the delivery times drastically change from some years to the next. Manova tests are done on these discoveries. Statistical process control samples all the classes and evaluates delivery times. The luxury, gifts and household classes showed deviations of delivery samples that should be inspected. The optimisation process looks at exactly where the problem lies with delivery times. The clothing, gifts and food class shows the most inconsistencies with the process. The reliability of the system is high (99.98%).

Bibliography

ISIXSIGMA, 2022. *ISIXSIGMA.* [Online]
Available at: https://www.isixsigma.com/dictionary/lower-control-limit-lcl/
[Accessed 18 October 2022].

R Coder, 2022. *R Coder.* [Online]
Available at: https://r-coder.com/plot-r/#:~:text=PLOT%20in%20R%20%E2%AD%95%20%5Btype,%2C%20add%20text%2C%20label%20points%5D
[Accessed 15 October 2022].

R Documentation, 2022. *R Documentation.* [Online]
Available at: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/uniroot
[Accessed 13 October 2022].