Quality Assurance

ECSA Project: Data Analysis

Reneilwe Mphahlele

10/21/22

[Course title]

Table of contents

Introduction

For this assignment, sales data was provided, with an intention to analyse it and come up with conclusions relating to the data at hand. For analysis, A computer programme, RStudio, is utilised for analysing the given data.

The data given consists of features such as age, class, price, year, month, day, delivery time, and reasons for a sale.

In this report, data wrangling will be conducted, in order to make adequate analysis of the data. Furthermore, the data given will be analysed using descriptive statistics and process capability indices of data under the technology class data will be calculated. The report then utilises statistical process control (SPC) charts for all classes of data to determine whether specifications of each class are being met or not. This is then followed by hypothesis tests (multivariate analysis of variables). Lastly, exercises from the prescribed module textbook are done.

1. Data wrangling

Before the commencement of the data analysis of the data given, the given data must be wrangled. Data wrangling may be regarded as a process of cleaning the given data in order to perform data analytics in a correct manner that is of great value and data integrity.

For this assignment, it was determined that the raw data consists of missing pricing values. As such, the data had to then be divided in to two parts: data with proper numerical pricing values and data with missing pricing values. The two parts of the data are then stored in excel files. For the purpose of this assignment, we will use the data with proper numerical pricing values.

Below is a picture of the code that was used to wrangle the data:

```
function(x){

  salesTableNA <- data.frame(filter(x,is.na(x$Price)))

  write.csv(salesTableNA,"NAPriceTable.csv", row.names=TRUE)
}

function(y){

  validsalesTable <- data.frame(filter(y,!is.na(y$Price)))

  write.csv(validsalesTable,"ValidSalesTable.csv", row.names=TRUE)
}

#From now on we will use data from the Valid Sales Table for part 2

ValidSales <- read.csv("ValidSalesTable.csv")
```

2. Descriptive statistics

When one works with data, data analysis of the data needs to be conducted. In the data analysis, each feature of the data needs to be described using measures of central tendency (mean, mode and median) and standard measures of variation (standard deviation, percentiles etc). For this purpose, two data quality reports (one for categorical features and another for continuous features) that contain all required descriptive measures have been formulated below.

Data quality report for the continuous features of the valid sales data:

Feature	Count	Cardinality	Min	1st Qrt	Mean	Median3	3rd Qrt	Max	Std. dev
Age	69591	91	18	39	55	54	72	108	20.601
Price	69591	48895	30.95	410.61	9790	1635	7623.82	102 239.05	17 864.98
Delivery.time	69591	138	1	6	27.08	17	38	141	27.66

Data quality report for the categorical features of the valid sales data:

Feature	Count	Cardinality	Mode	Mode freq	Mode %	2nd Mode	2nd Mode freq	2nd Mode %
Class	69591	7	Gifts	15463	22.22	Technology	13722	19.72
Year	69591	9	2018	12872	18.50	2026	8616	12.38
Month	69591	12	3	5893	8.47	12	5832	8.38
Day	69591	30	5	2466	3.54	24	2444	3.51
Why.bought	69591	6	Recommended	37 298	53.60	Website	13097	18.82

In addition to the data quality reports, visualisations in the form of histograms and bar plots have been created to show what kind of distribution each feature of the given sales data has.

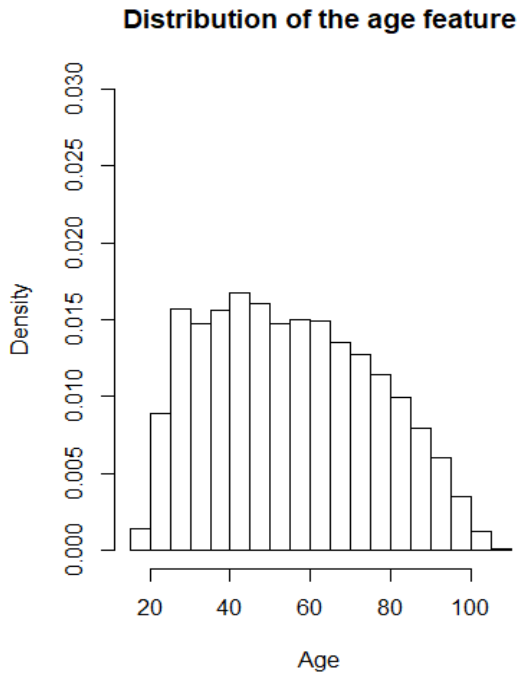


Figure 1: the age feature has a unimodal (skewed right) distribution of data

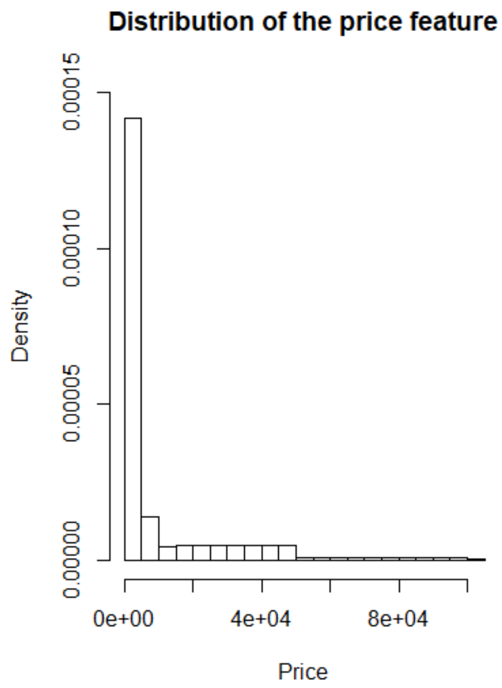


Figure 2: the price feature has an exponential distribution of data

Figure 3: the delivery time feature has a multimodal distribution of data

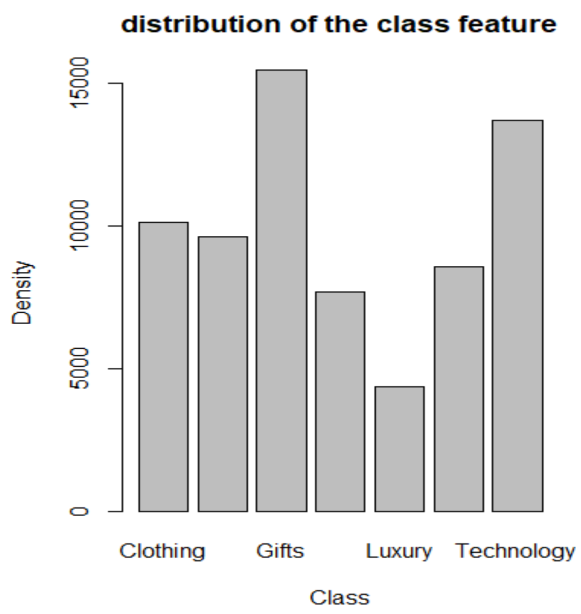


Figure 4: the class feature had a multimodal distribution of data

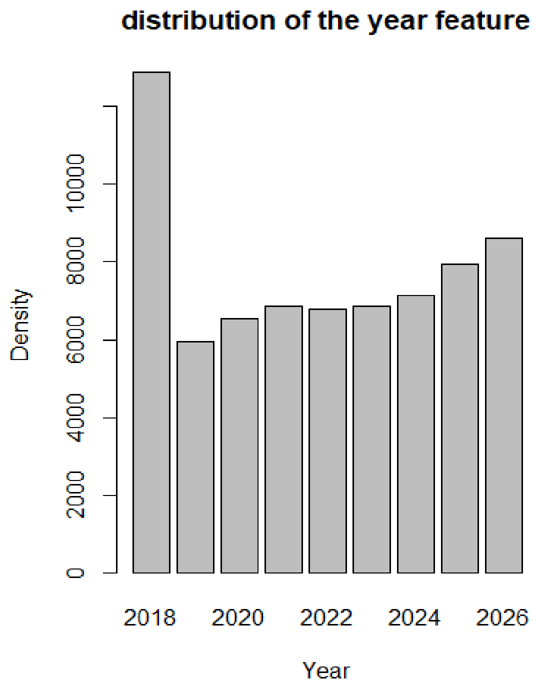


Figure 5: the year feature has a uniform distribution of data

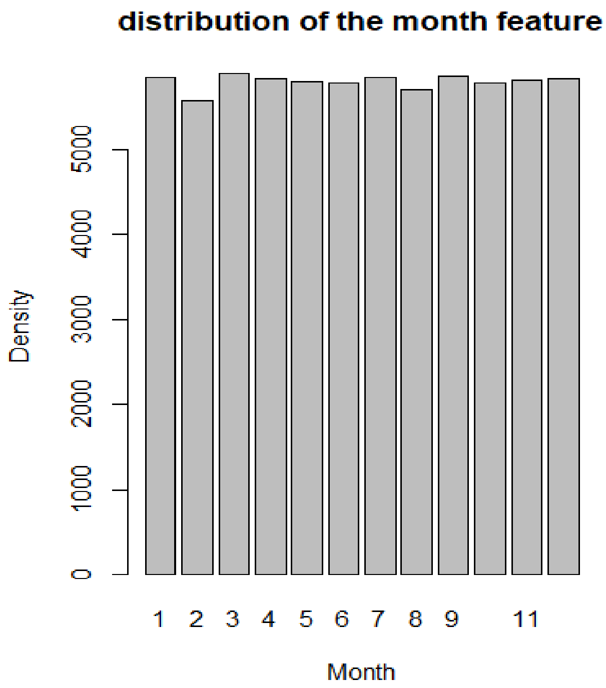


Figure 6: the month feature has a uniform distribution of data

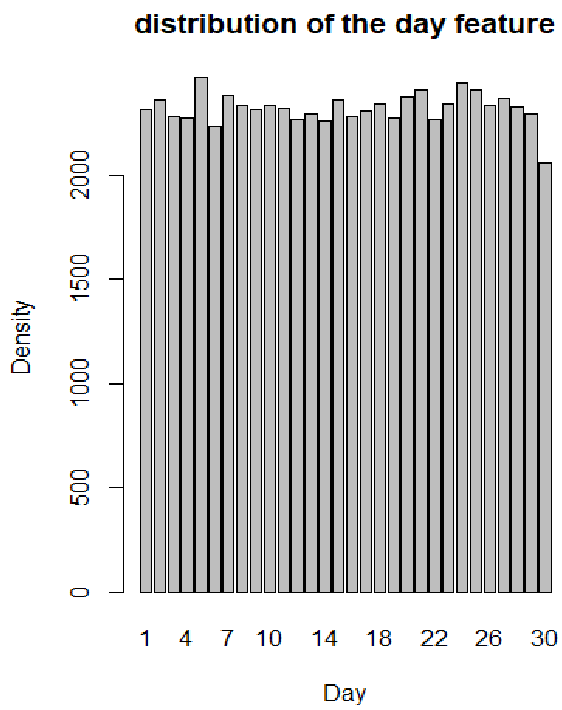


Figure 7: the day feature has a uniform distribution of data

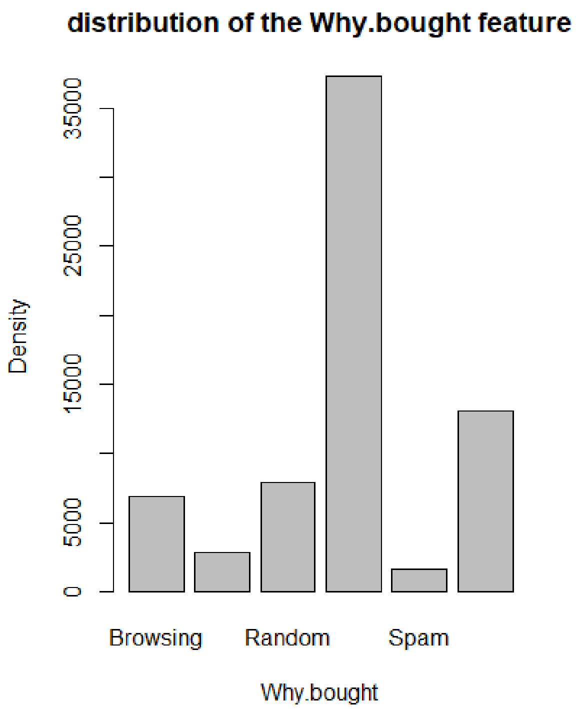


Figure 8: the why,bought feature has a multimodal distribution of data

Process capability indices

For this section of the report, the process capability indices of process delivery times of data that fell under the technology class were calculated. The C_p, C_{pu}, C_{pl}, and C_{pk} were calculated. The aim of these calculations is to analyse and determine the relationship of the specifications of the delivery process times and variation that the data of this feature has. In order to conduct the said indices, an upper specification limit (USL) of 48 days and a lower specification limit (LSL) of 0 days. A LSL of 0 days is logical on the basis that items cannot be delivered in less than 0 days. Items can however be delivered on the same day as they are ordered.

The calculations of the process capability indices are as follows:

$$C_p = \frac{USL-LSL}{6\sigma} = \frac{48-0}{6\times7.987665} = 1.001$$

$$C_{pu} = \frac{USL-\mu}{3\sigma} = \frac{48-41.94}{3\times7.987665} = 0.253$$

$$C_{pl} = \frac{\mu-LSL}{3\sigma} = \frac{41.94-0}{3\times7.987664} = 1.750$$

$$C_{pk} = \min(C_{pl}\ , C_{\boldsymbol{pu}}) = 0.253$$

$$k = \frac{2\times|\mu-T|}{USL-LSL}$$

$$= 0.748$$

Analysis of the process capability indices

The C_p of the process delivery times is 1.001. According to Evans and Lindsay (2017), this means that the process is capable of meeting their specification due to the process variation being smaller than the specification range.

Regarding process centering, we examine the values of the C_{pu} and C_{pl} values, which are the upper and lower one-sided index respectively. The C_{pu} value is smaller than the C_{pl} value. Evans and Lindsay (2017) interpret this as the process mean being closer to the upper specification limit.

Using the k value that is calculated, we are able to determine how well is the process centred.. The value is 0.748, meaning that the process is tending towards being centered at or beyond the specification limits (Evans and Lindsay, 2017).

3. Control charts

For this section of the work, X-bar and s control charts are analysed as a means of determining the mean and variation of the process under the various classes the data is under using the first 25 samples of 15 sales each.

Analysis of the X-bar and s control charts of data under the Technology class

This sub-section focuses on analysing the x-bar and s control charts of data under the technology class.

X-bar control chart of the technology class

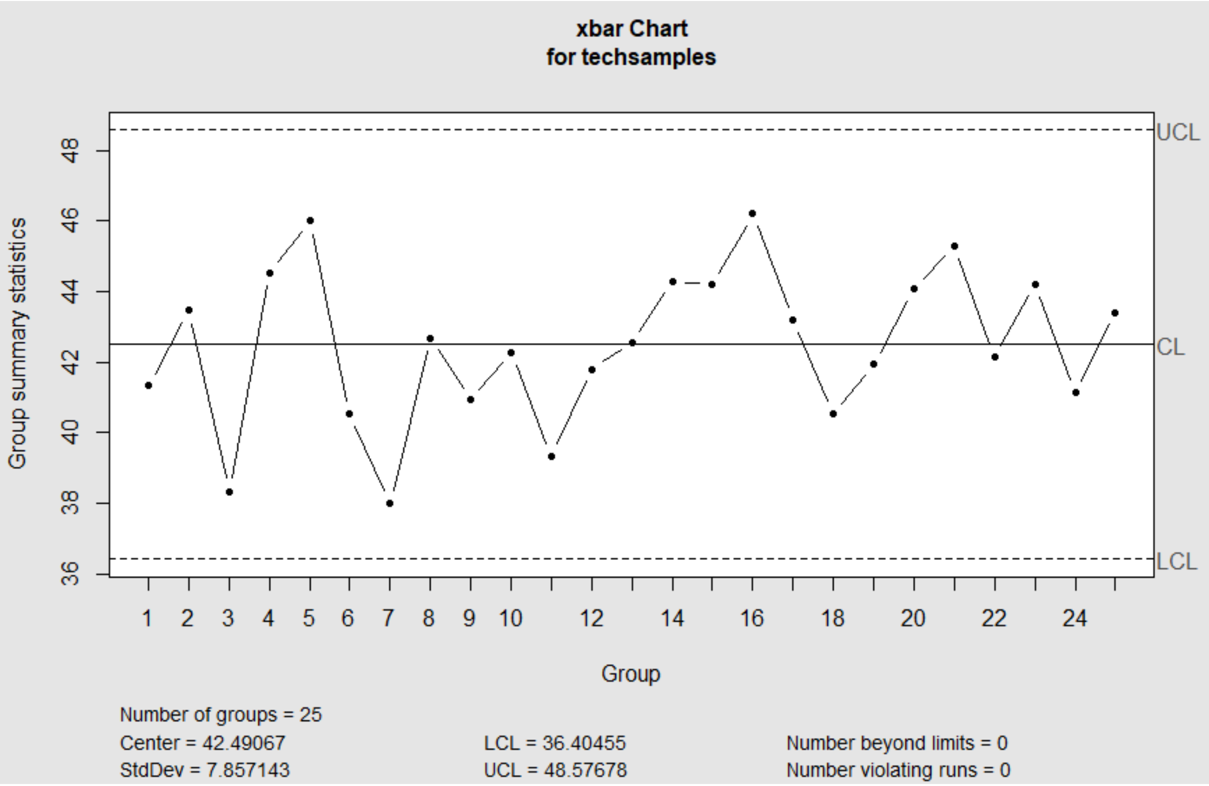


Figure 9: The X-bar control chart of the technology class

The graph shows no noticeable trend. Furthermore, the number of samples are evenly spread between the two sides of the centre line with most being close samples being close to the line mentioned. With this stated, it can be concluded that the delivery times of the technology class data are in statistical control.

S control chart

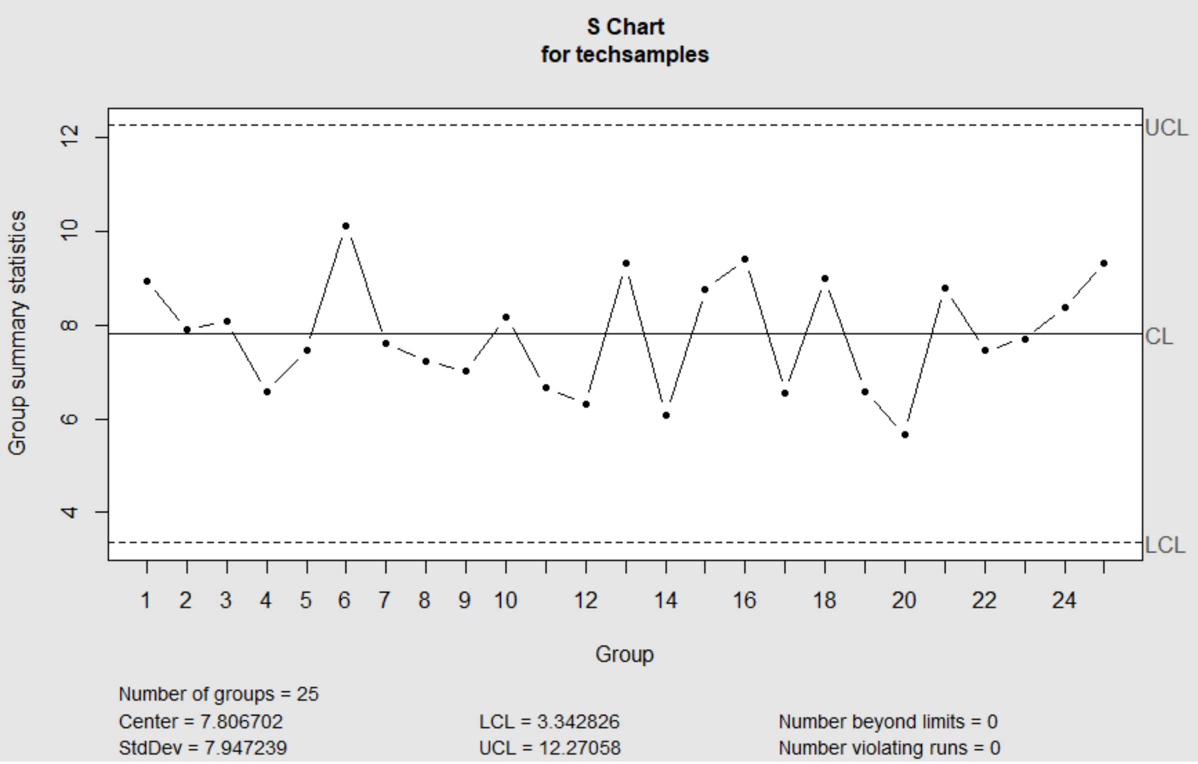


Figure 10: the s control chart of the technology class

The graph shows no noticeable trend. Furthermore, the number of samples are evenly spread between the two sides of the centre line with most being close samples being close to the line mentioned. With this stated, it can be concluded that the delivery times of the technology class data are in statistical control.

Analysis of the X-bar and s control charts of data under the Clothing class

This sub-section focuses on analysing the x-bar and s control charts of data under the clothing class.

x-bar control chart

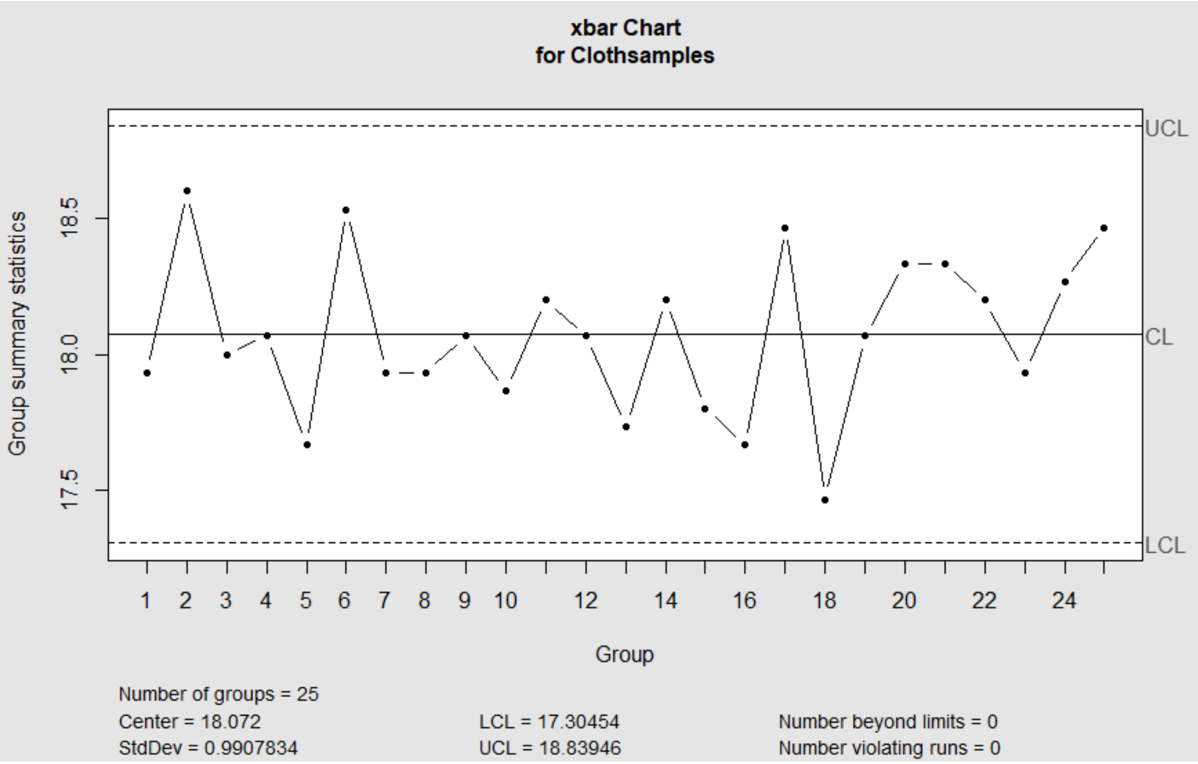


Figure 11: the X-bar control chart of the clothing data

Like the X-bar chart, the graph shows no noticeable trend. Furthermore, the number of samples are evenly spread between the two sides of the centre line with most being close samples being close to the line mentioned. In addition, three data points are directly on the center line. The 18th data point is close to the lower control limit. With this stated, it can be concluded that the delivery times of the technology class data are in statistical control.

s control chart

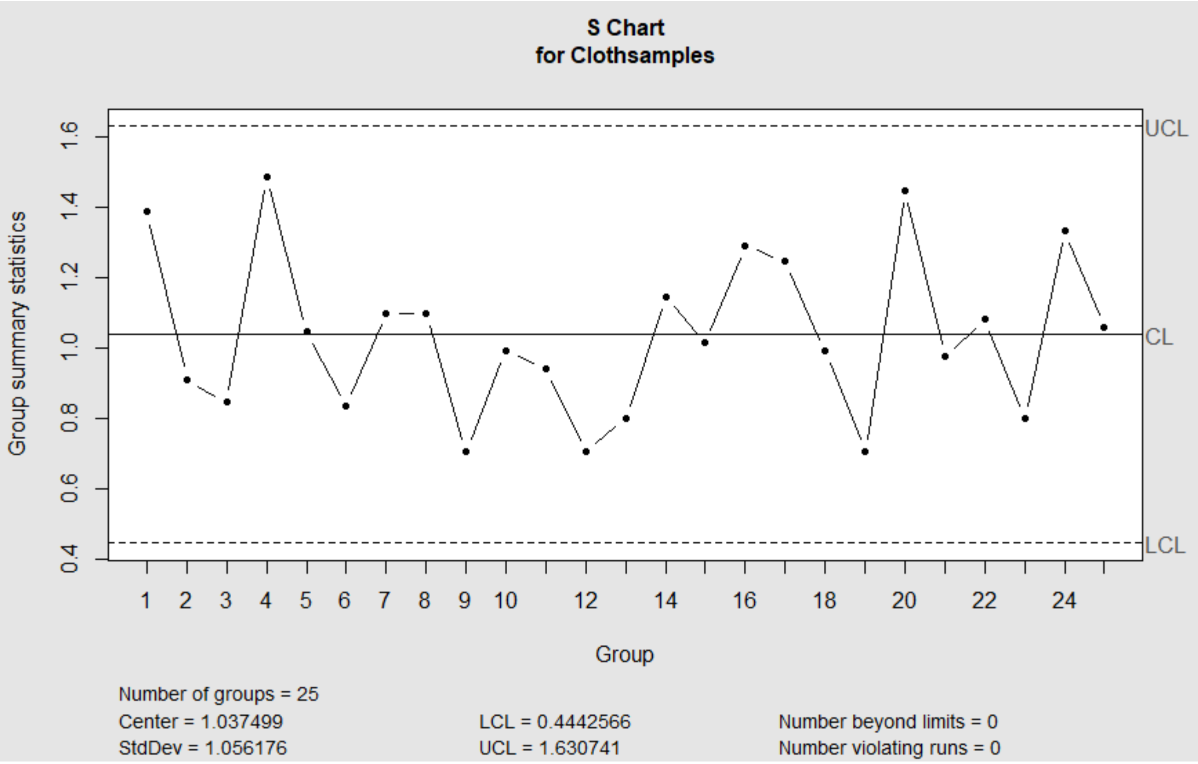


Figure 12: the s control chart for the clothing data

The graph shows a little trend between datapoints 10-12. Possible cause of this trend include the driver getting lost whilst delivering clothing goods or the driver being given the incorrect delivery address by office admin. Furthermore, the number of samples are not evenly spread between the two sides of the centre line with most being under the line mentioned. With this stated, it can be concluded that the delivery times of the technology class data are not in statistical control.

Analysis of the X-bar and s control charts of data under the Household class

This sub-section focuses on analysing the x-bar and s control charts of data under the housing class.

X-bar control chart

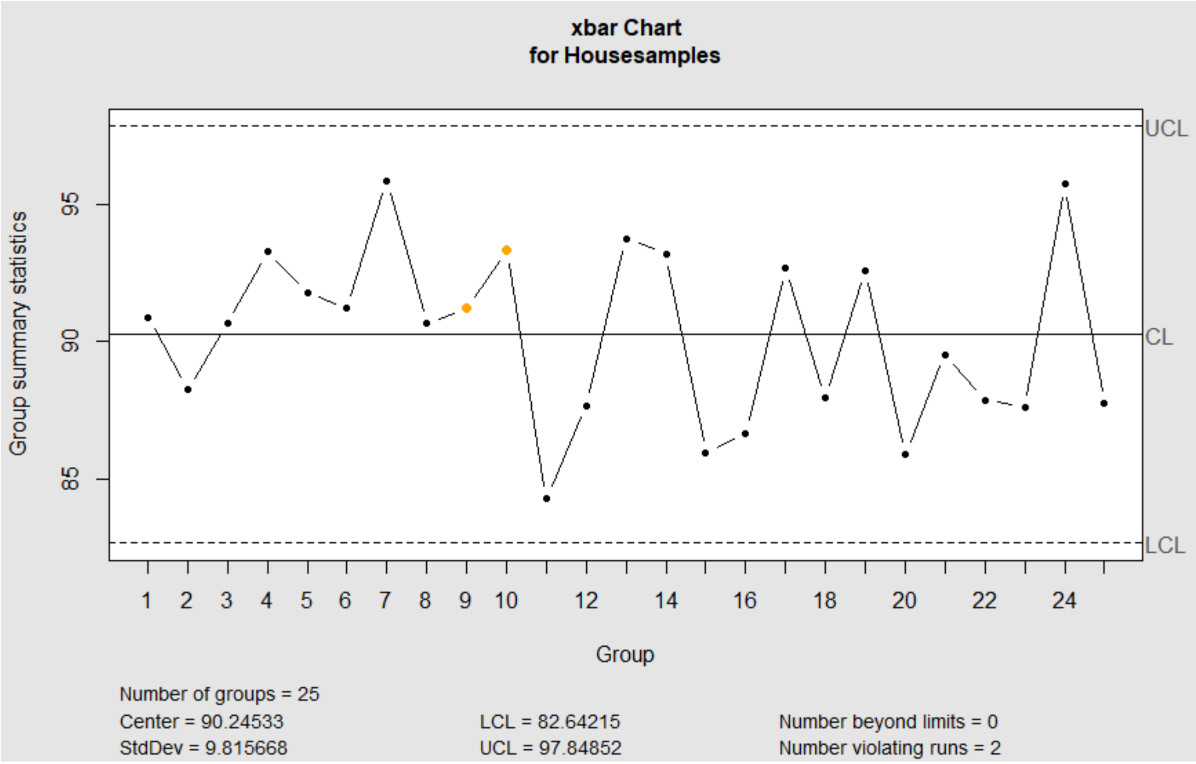


Figure 13: The X-bar control chart of the housing data

The graph shows no noticeable trend. Furthermore, the number of samples are not evenly spread between the two sides of the centre line with most being in above the line mentioned. In addition, there are two points in the upper half of the graph that have errors (violating runs). Possible causes of this could be an error in the calculations done, a broken tool, missing values etc. With this stated, it can be concluded that the delivery times of the technology class data are not in statistical control.

S control chart

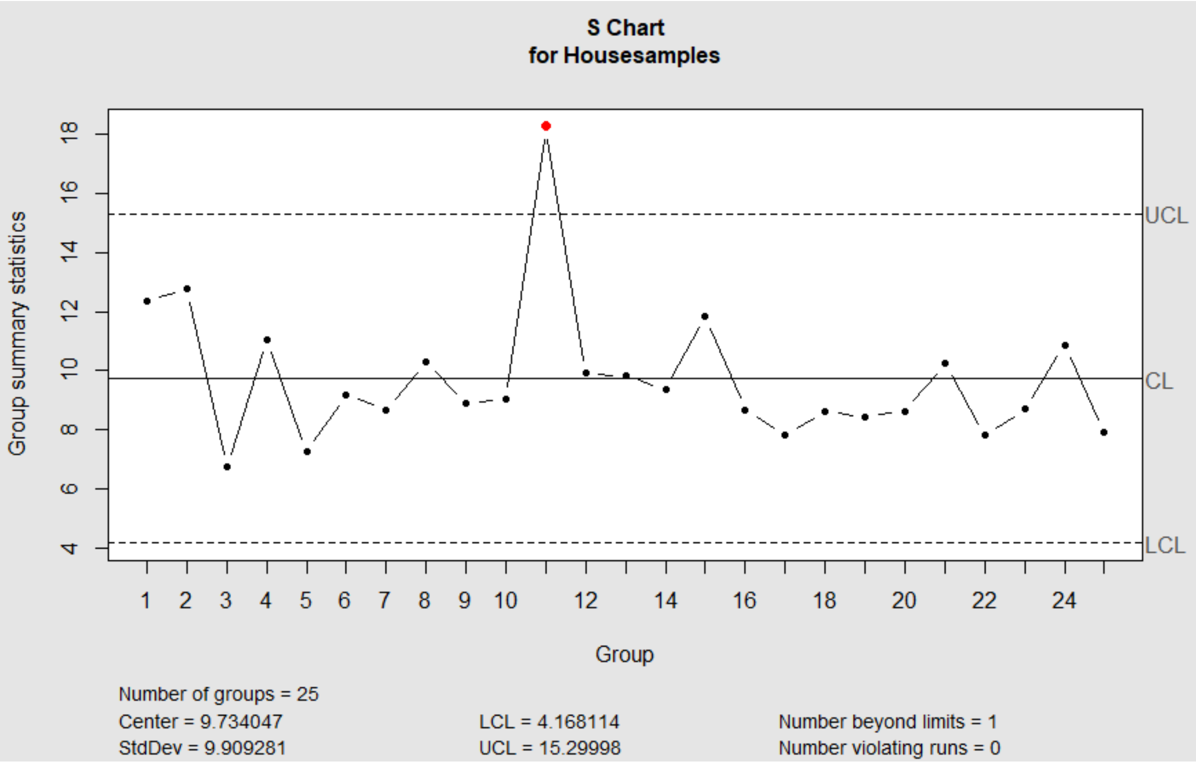


Figure 14: the s control chart of the housing data

The graph shows two noticeable trends. The first trend (between data points 12 and 14) has a decreasing pattern and the second one (between data points 17 and 21) has a gradual increasing pattern. Possible reasons for this would be different drivers being used for delivering goods, delivery vehicles needing repairs. There is also one data point (data point 11) that is plotted beyond the upper control limit. The possible cause for this would be an error in which the upper control limit was calculated, a cancelled delivery, or a delivery vehicle crashing. Furthermore, the number of samples are not evenly spread between the two sides of the centre line with most being plotted on the lower half of the graph. With this stated, it can be concluded that the delivery times of the technology class data are not in statistical control.

Analysis of the X-bar and s control charts of data under the Luxury class

This sub-section focuses on analysing the x-bar and s control charts of data under the luxury class.

X-bar control chart

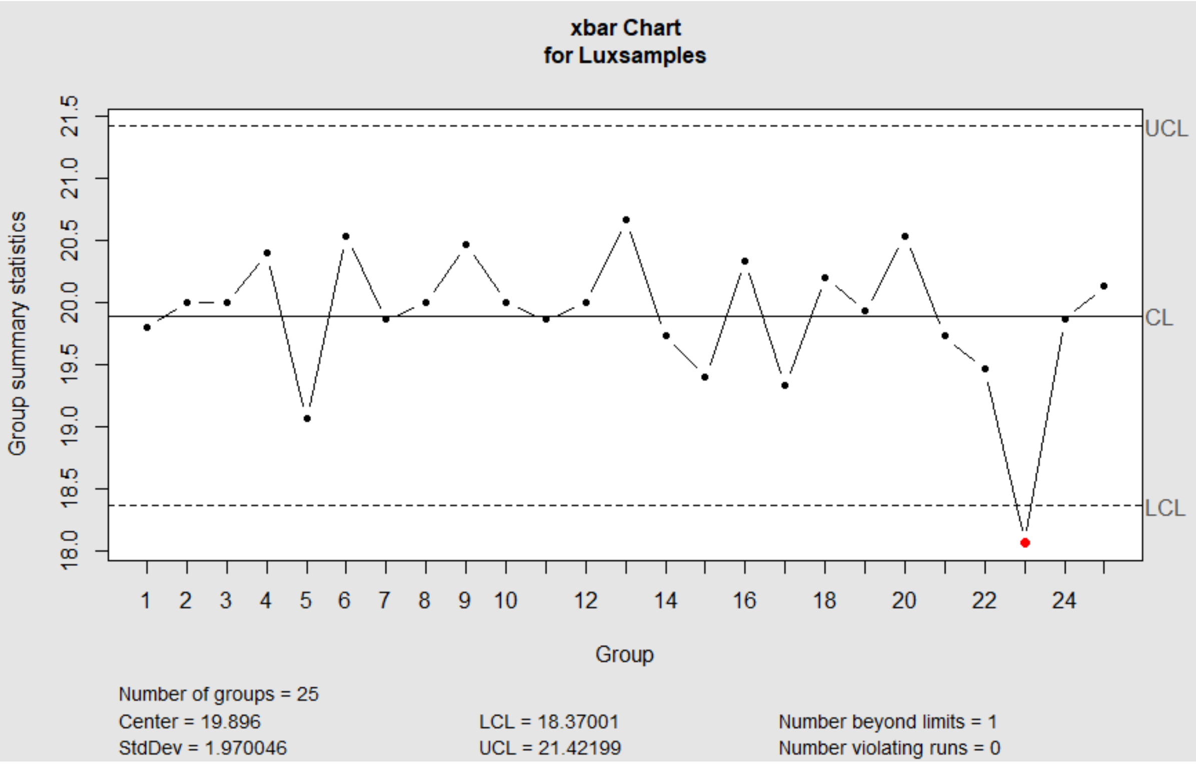


Figure 15: The X-bar control chart of the luxury data

The graph shows no noticeable trend. Furthermore, the number of samples are not evenly spread between the two sides of the centre line with most (14 data points out of 25) being located on the upper half of the line mentioned. In addition, one data point is beyond the lower control limit. Possible causes for this is a calculation error in the calculation of the lower limit, a delivery vehicle broke down or the driver had an accident. With this stated, it can be concluded that the delivery times of the luxury class data are not in statistical control.

S control chart

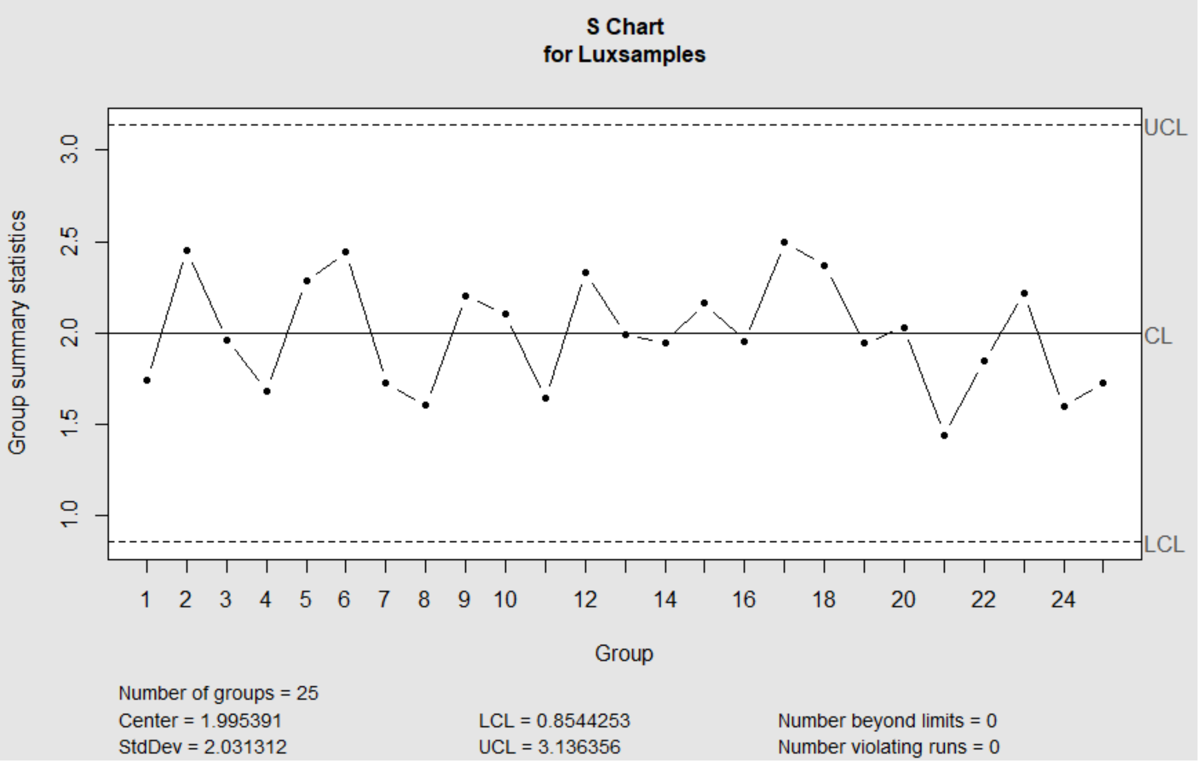


Figure 16: The s control chart for the luxury data

The graph shows no noticeable trend. The data points are plotted randomly and are within the control limits. Furthermore, the number of samples are not evenly spread between the two sides of the centre line with most being located in the lower half of the line mentioned. With this stated, it can be concluded that the delivery times of the luxury class data are not in statistical control.

Analysis of the X-bar and s control charts of data under the Food class

This sub-section focuses on analysing the x-bar and s control charts of data under the food class.

X-bar control chart

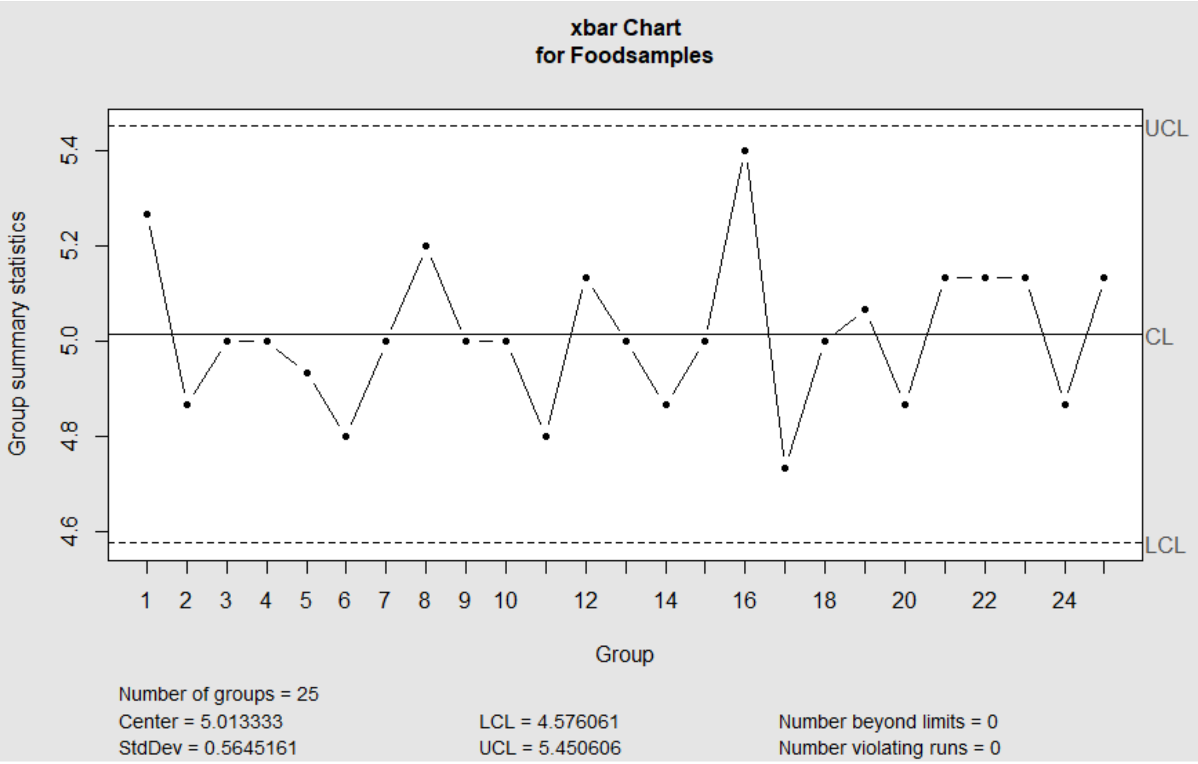


Figure 17: The X-bar chart for the food data

The graph shows no noticeable trend with no points located outside the control limits. Furthermore, the number of samples are not evenly spread between the two sides of the centre line with most (16 data points out of the 25) being below the line mentioned. Most points are located close to the center line. With this stated, it can be concluded that the delivery times of the technology class data are in statistical control.

S control chart

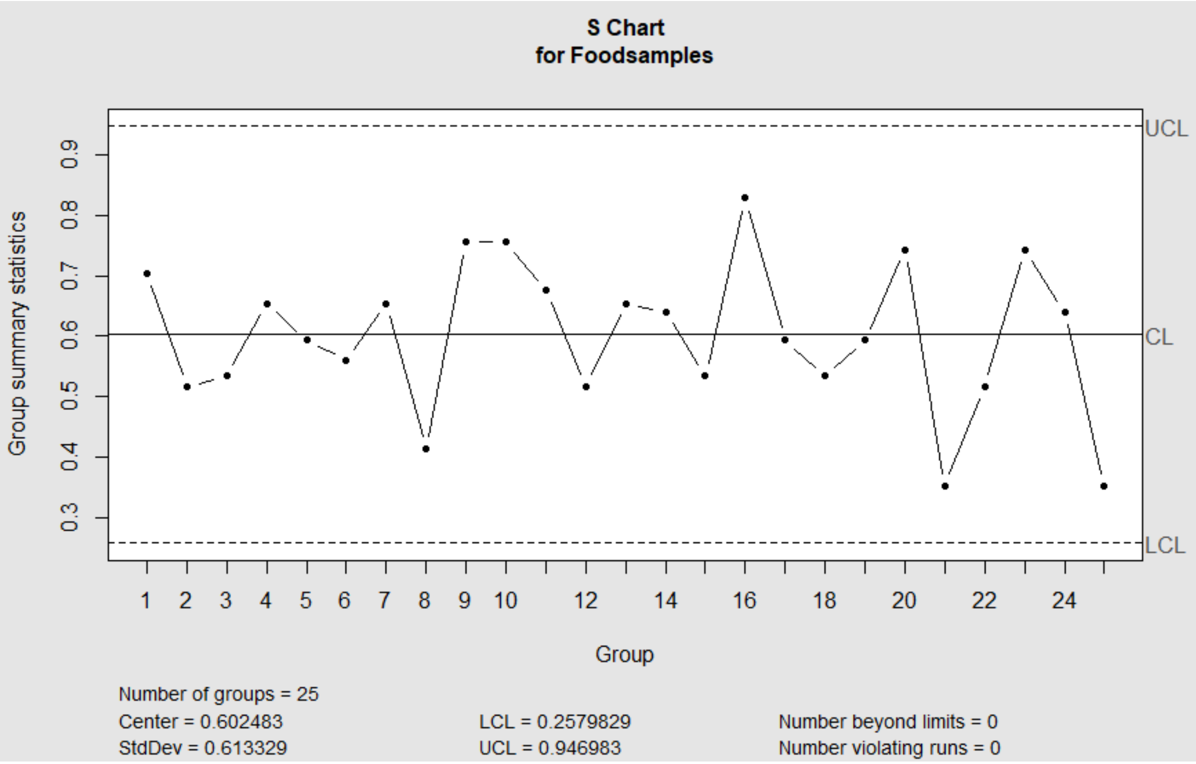


Figure 18: The s control chart of the food data

The graph shows one noticeable trend. From data point 20, the points are plotted much further from the centreline, indicating that the delivery process of delivering food was disturbed. Possible causes include prolonged delivery times due to traffic, unplanned stops on the road and changing of drivers. Furthermore, the number of samples are evenly spread between the two sides of the centre line with most being close samples being close to the line mentioned. With this stated, it can be concluded that the delivery times of the technology class data are in statistical control.

Analysis of the X-bar and s control charts of data under the Gifts class

This sub-section focuses on analysing the x-bar and s control charts of data under the gifts class.

X-bar control chart

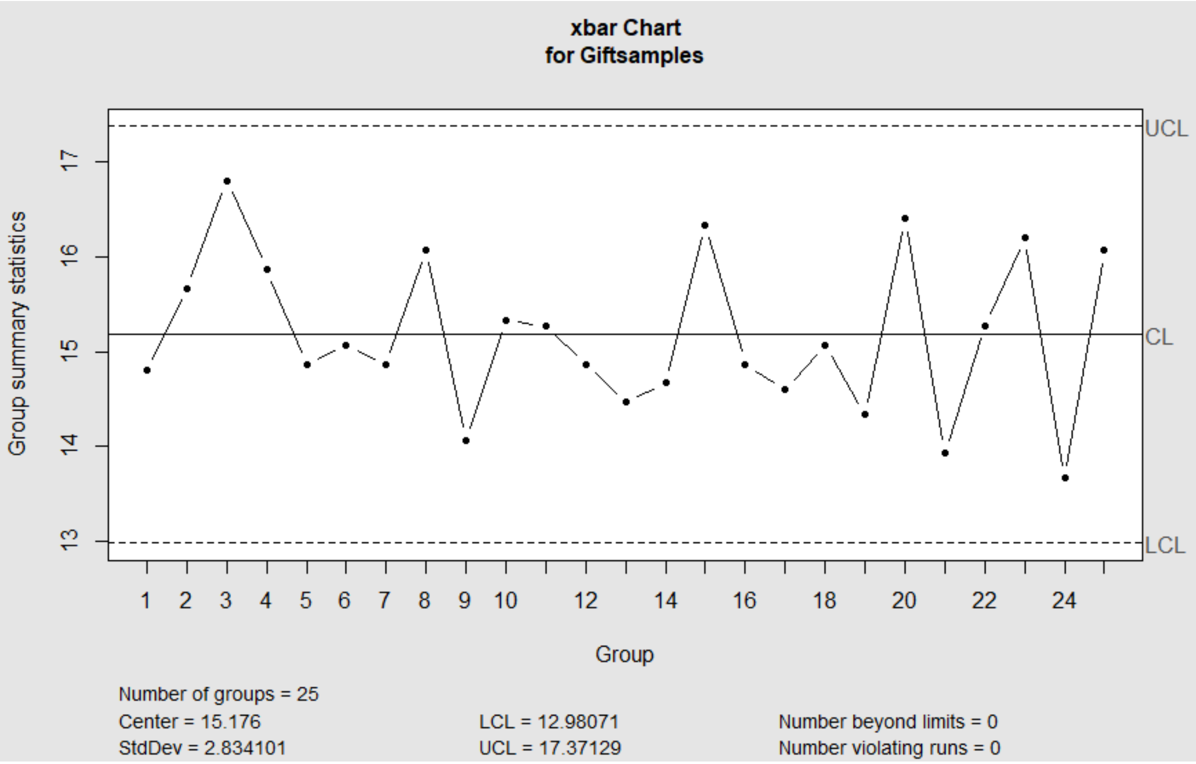


Figure 19: The X-bar chart of the gift data

The graph shows one noticeable trend. Similar to the s control chart of the food data, the points (from data point 19) are plotted much further from the centreline, indicating that the delivery process of delivering food was disturbed. Possible causes include prolonged delivery times due to traffic, unplanned stops on the road and changing of drivers. Furthermore, the number of samples are not evenly spread between the two sides of the centre line with most being plotted close to the line mentioned. With this stated, it can be concluded that the delivery times of the technology class data are not in statistical control.

S control chart

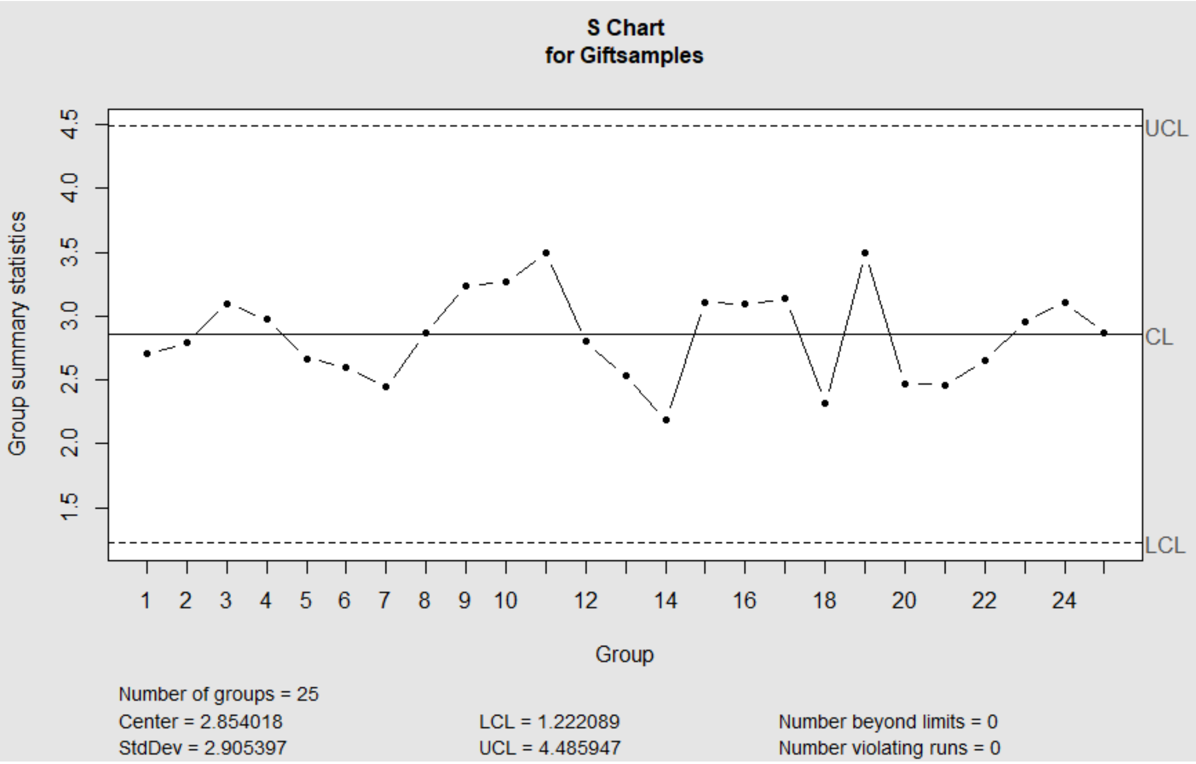


Figure 20: The s control chart of the gift data

The graph shows no noticeable trend. The datapoints are further located close to the center line. There are also less spikes on this data compared to the other s control charts Furthermore, the number of samples are evenly spread between the two sides of the centre line with most being close samples being close to the line mentioned. With this stated, it can be concluded that the delivery times of the technology class data are in statistical control.

Analysis of the X-bar and s control charts of data under the Sweets class

This sub-section focuses on analysing the x-bar and s control charts of data under the sweets class.

X-bar control chart

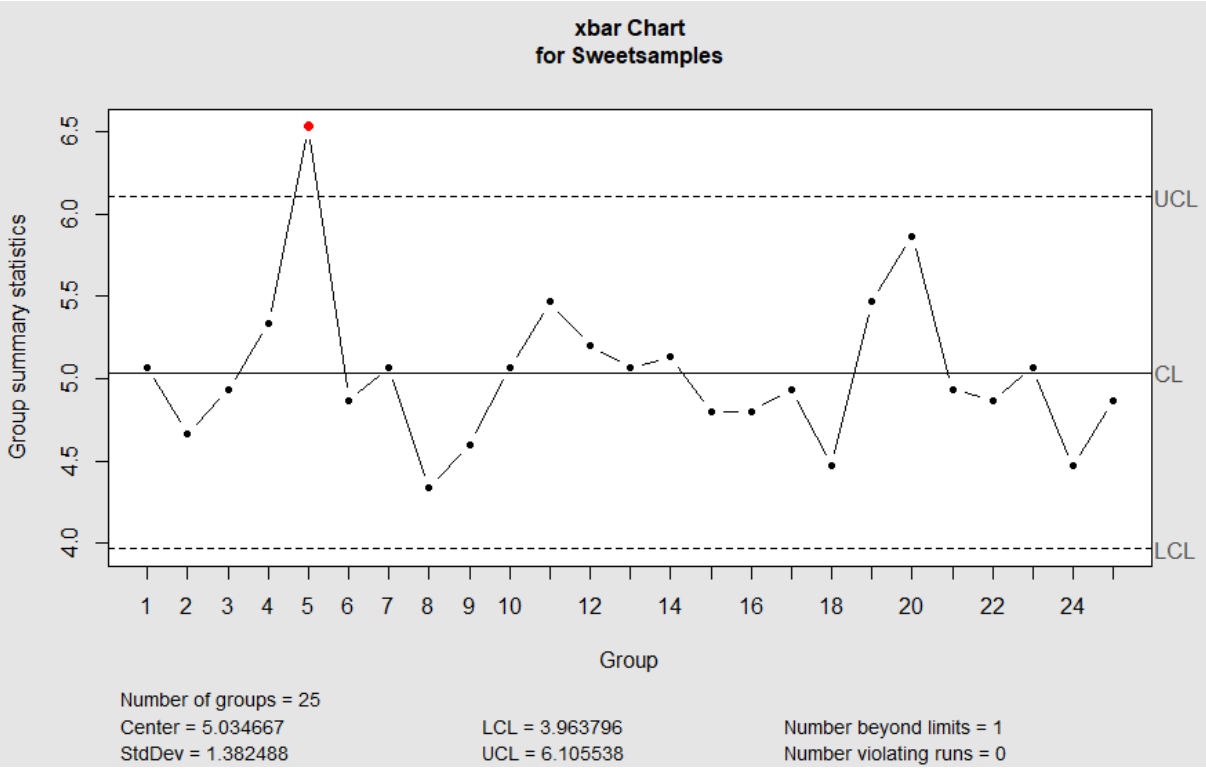


Figure 21: The X-bar control chart for the sweets data

The graph shows one noticeable trend. From data point 11 to data point 18, there is a decreasing pattern of the points. In addition, one data point is beyond the lower control limit. Possible causes for this is a calculation error in the calculation of the upper control limit, a delivery vehicle broke down or the driver had an accident. Furthermore, the number of samples are evenly spread between the two sides of the centre line with most being close samples being close to the line mentioned. With this stated, it can be concluded that the delivery times of the sweet class data are not in statistical control.

S control chart

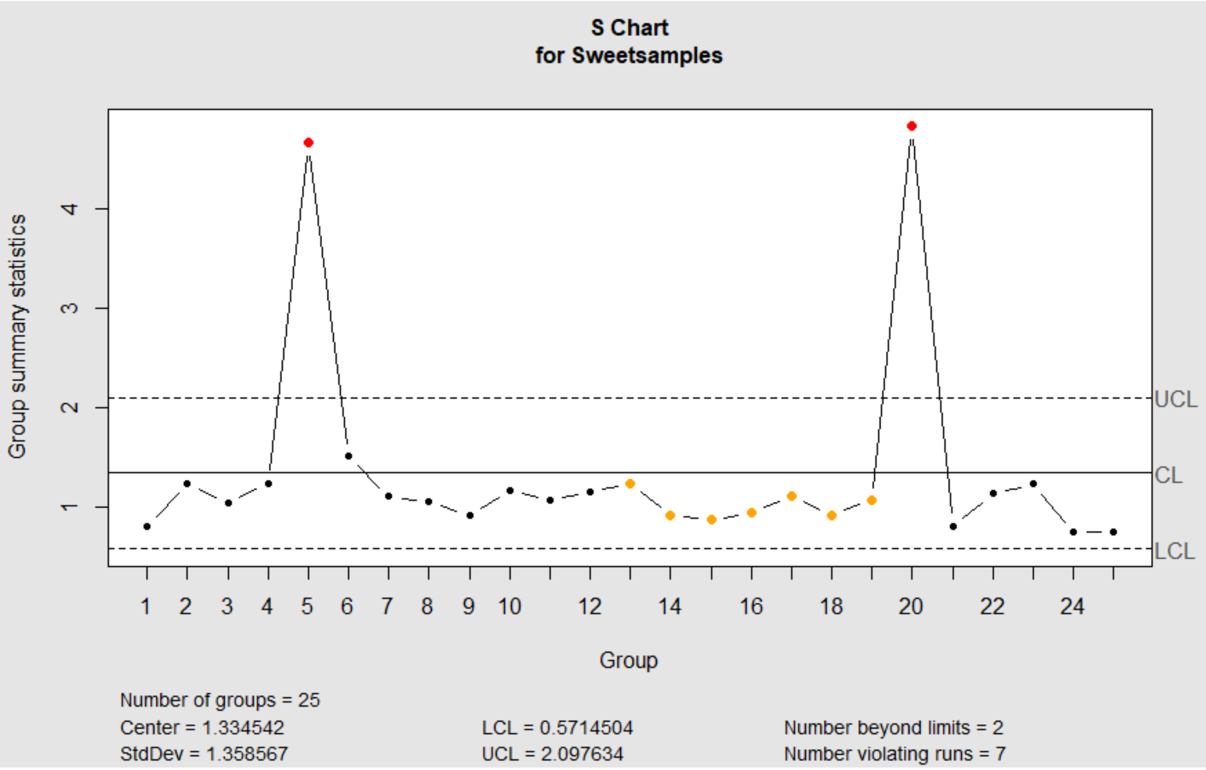


Figure 22: The s control chart of the sweet data

The graph shows one noticeable trend. The data points below the center line have a wave-like (cyclic) pattern. Furthermore, the number of samples are not evenly spread between the two sides of the centre line with most data points (22 data points out of 25) being close to the line mentioned. Two of the three points above the center line are beyond the upper control limit, possibly due to incorrect calculation of the control limits. With this stated, it can be concluded that the delivery times of the sweet class data are not in statistical control.

Class	UCL	U2Sigma	U1sigma	CL	L1Sigma	L2Sigma	LCL
Technology	48.577	48.838	44.664	42.491	40.317	38.143	36.405
Clothing	18.839	18.654	18.363	18.072	17.781	17.489	17.305
Household	97.849	96.567	93.403	90.245	87.087	83.929	82.642
Luxury	21.422	20.994	20.445	19.896	19.347	18.798	18.370
Food	5.451	5.326	5.169	5.013	4.857	4.701	4.576
Gifts	17.371	16.823	15.999	15.176	14.353	13.529	12.981
Sweets	6.106	5.959	5.497	5.035	4.572	4.110	3.964

4. Optimising the delivery processes

In this part of the report, an in-depth analysis of the data that we are given is analysed by means statistical process controlling. In addition, all data was used except data between the 1st and 25th data set

Using R programming as a tool of analysis and calculation…

Optimising the delivery process for the Technology class data

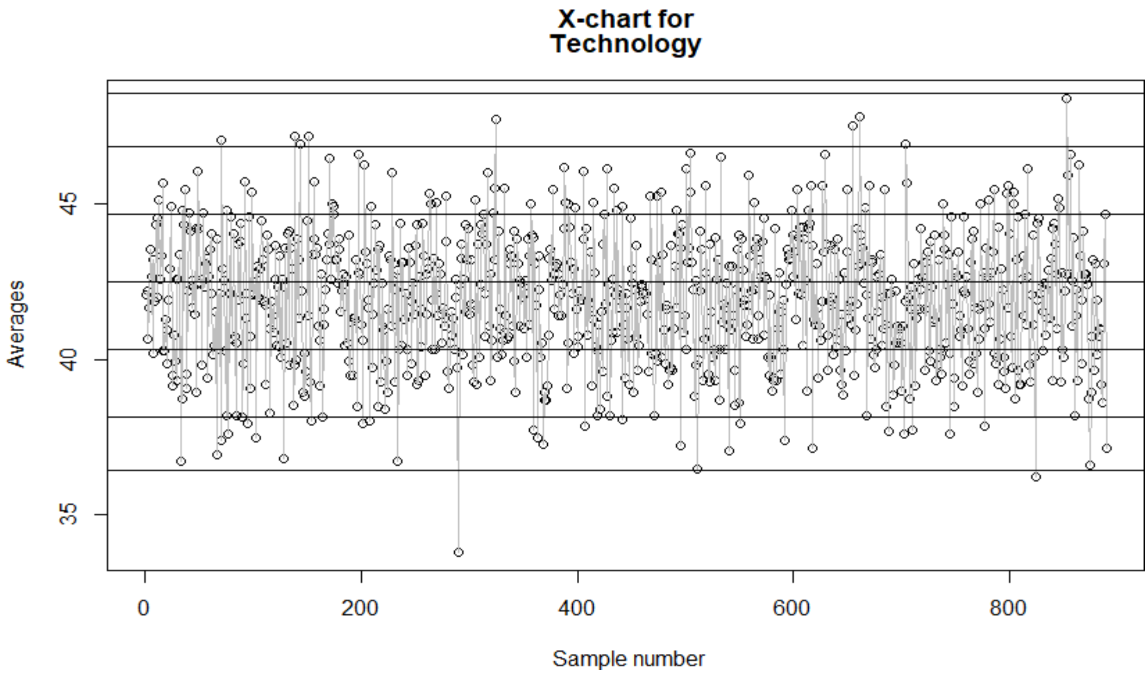


Figure 23: The X-bar chart for the rest of the technology data

27 samples are out of control.

Data points 47 to 51 showed a decreasing pattern on the bottom half of the graph.

Data points 165 to 172 have the most consecutive samples above the center line.

Data points 3 to 20 are the most consecutive sample that are between -1 and 1 sigma control limits.

Optimising the delivery process for the Clothing class data

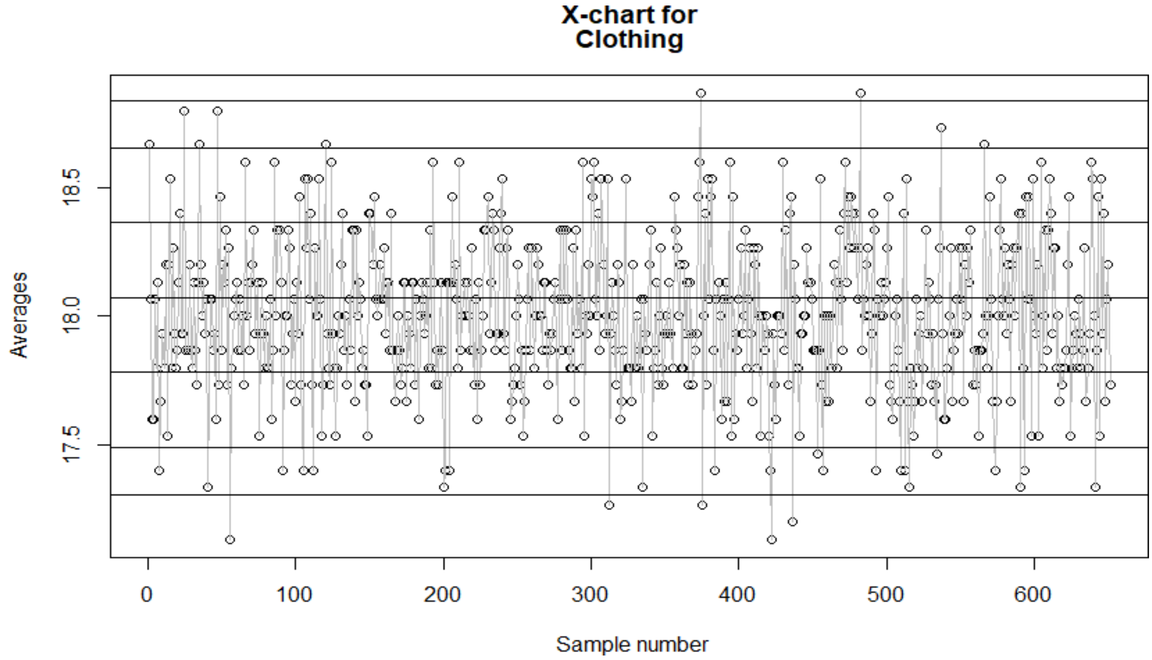


Figure 24: The X-bar chart for the rest of the clothing data

5 data points are out of control and they are points 240, 290, 312, 315 and 332.

Data points 52 to 54 showed a decreasing pattern on the bottom half of the graph.

Data points 71 to 78 have the most consecutive samples above the center line.

Data points 41 to 53 are the most consecutive sample that are between -1 and 1 sigma control limits.

Optimising the delivery process for the Housing class data

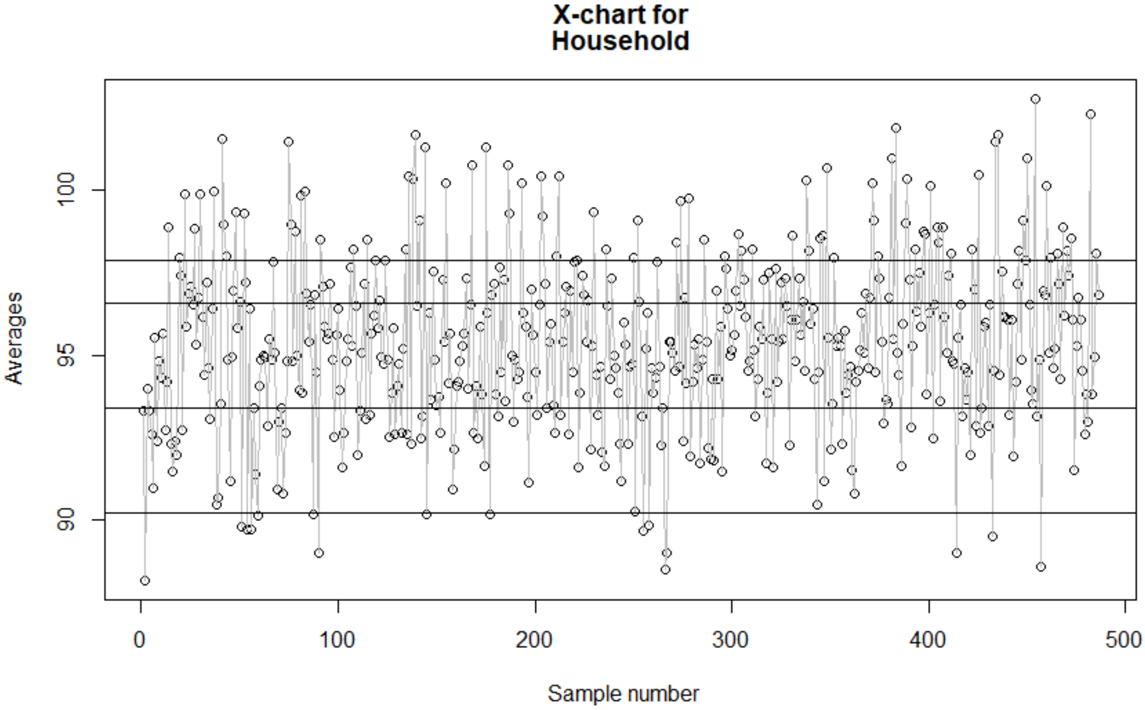


Figure 25: The X-bar chart for the rest of the household data

87 samples are out of control, with the first three being data points 161, 166 and 169 and the last three being data points 266, 267 and 268.

Data points 47 and 51 showed a decreasing pattern on the bottom half of the mean.

Data points 167 to 268 have the most consecutive samples above the center line.

Data points 44 to 59 are the most consecutive sample that are between -1 and 1 sigma control limits.

Optimising the delivery process for the Luxury class data

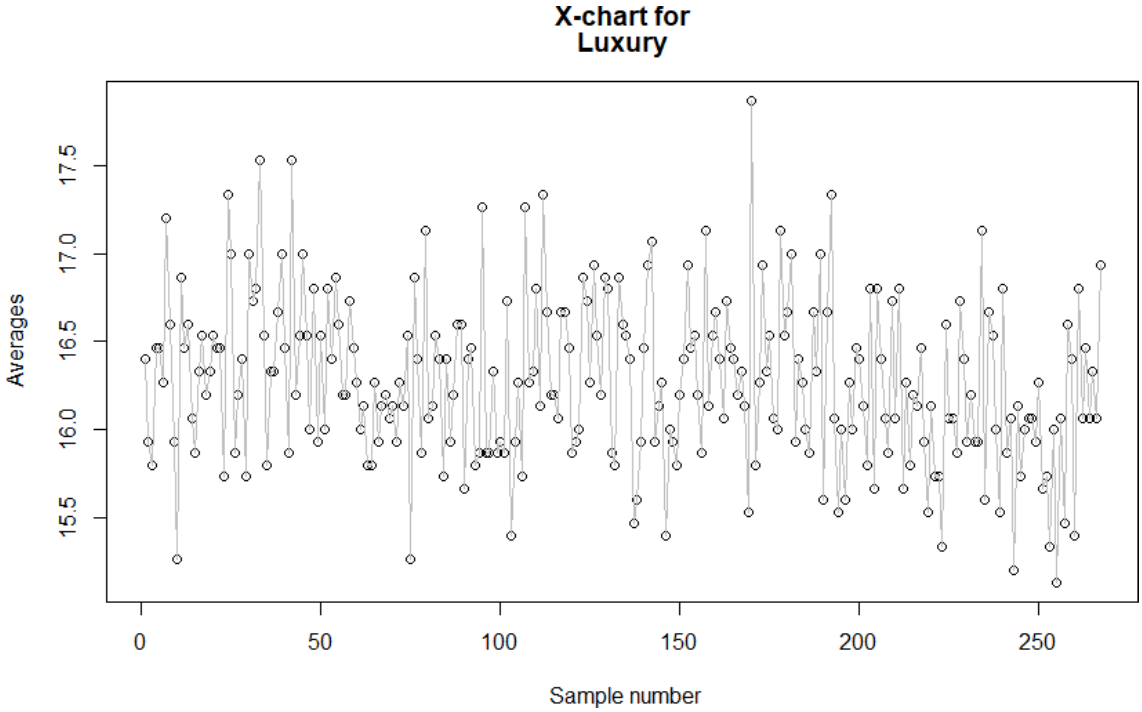


Figure 26: The X-bar chart for the rest of the luxury data

130 samples are out of control, with the first three being data points 3, 8 and 13 and the last three being data points 143, 144 and 145.

Data points 111 to 115 showed a decreasing pattern on the bottom half of the graph.

There are no most consecutive sample above the center line.

Only data point 1 has the most consecutive sample that are between -1 and 1 sigma control limits.

Optimising the delivery process for the Food class data

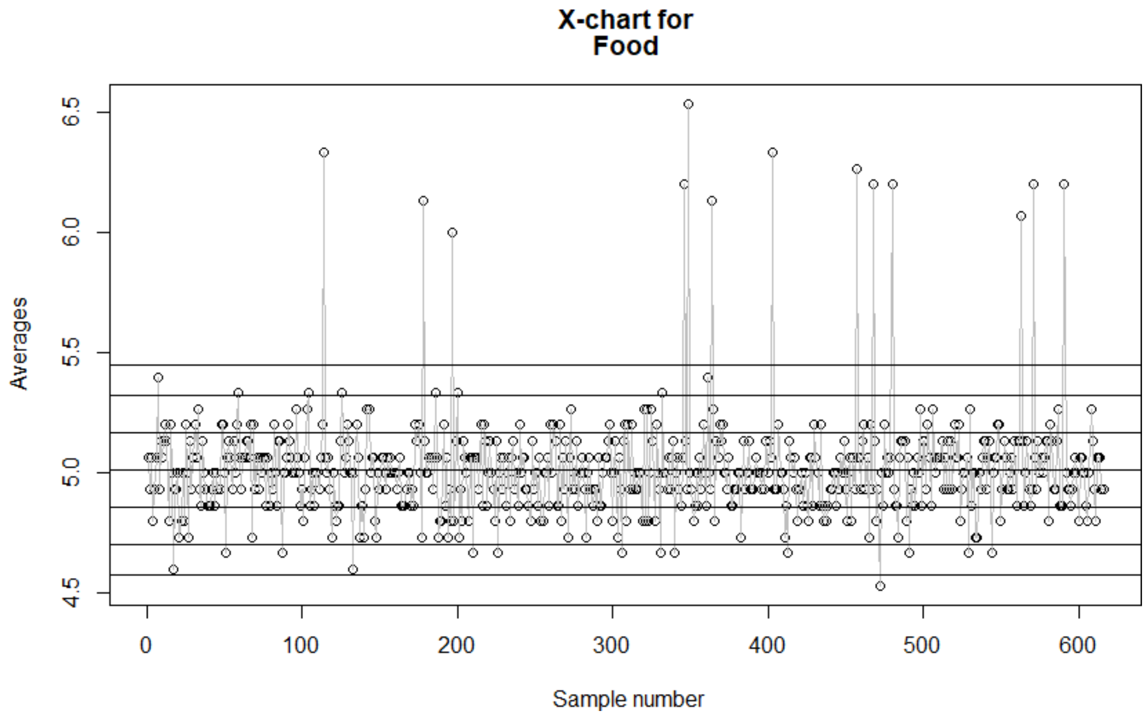


Figure 26: The X-bar chart for the rest of the food data

15 samples are out of control, with the first three being data points 1, 70 and 83 and the last three being data points 305, 311 and 326.

Data points 78 to 80 showed a decreasing pattern on the bottom half of the graph.

Data points 170 to 179 have the most consecutive samples above the center line.

Data points 181 to 190 are the most consecutive sample that are between -1 and 1 sigma control limits.

Optimising the delivery process for the Gifts class data

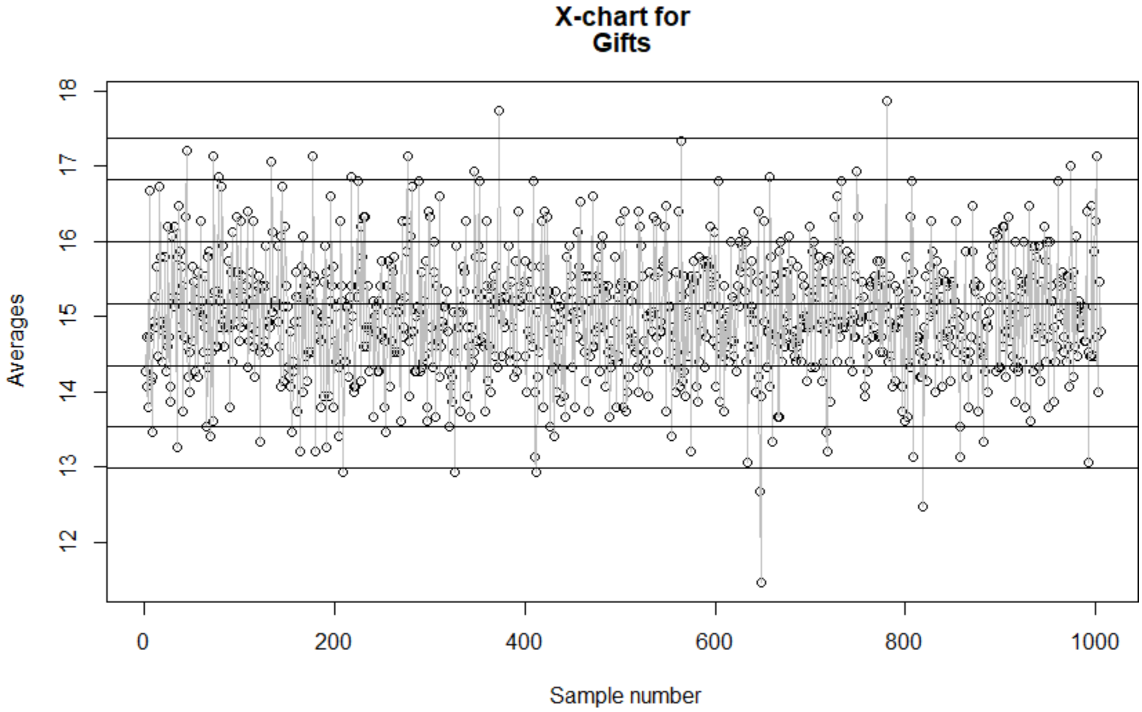


Figure 27: The X-bar chart for the rest of the gifts data

409 samples are out of control, with the first three being data points 1, 70 and 83 and the last three being data points 305, 311 and 326.

Data points 537 to 540 showed a decreasing pattern on the bottom half of the graph.

Data points 165 to 172 have the most consecutive samples above the center line.

Data points 169 to 184 are the most consecutive sample that are between -1 and 1 sigma control limits.

Optimising the delivery process for the Sweets class data

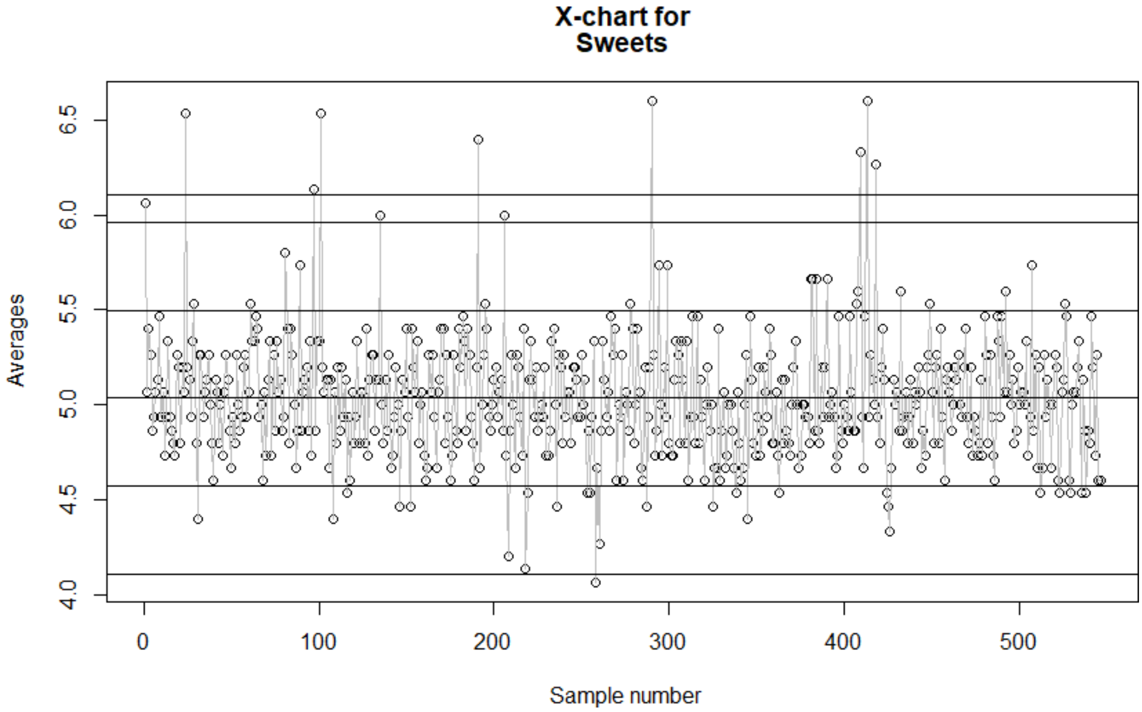


Figure 28: The X-bar chart for the rest of the sweets data

Data points 25, 46,125, 163 and 240 were out of control.

Data points 21 to 23 showed a decreasing pattern on the bottom half of the graph.

Data points 210 to 217 have the most consecutive samples above the center line.

Data points 218 to 231 are the most consecutive sample that are between -1 and 1 sigma control limits.

Likelihood of making a Type I error

Whilst doing the various calculations of analysis for the section above, the type 1 error for A and B was calculated. From the calculations conducted, it it determined that the probability of making a Type I error is the same for all clusters. For A, the Type 1 probability is 0.0027 and 0.683 for B.

Central processing for best profit

When analysing the data of the technology class, we see that in many cases, there are delivery times that are greater than the 48 days. From the problem statement, it can be deduced that delivery times greater than 48 days result in a loss of sales. In order to counteract such a result, the reduction of the average delivery time needs to occur. The reduction will cost money, but will be much less than the sales that can be lost if the delivery times aren't corrected to be below 48 days.

In this section, the days in which to delivery process can be centred is determined.

Using R programming, the following graph gives us the number of days in which the delivery process times can be centred.

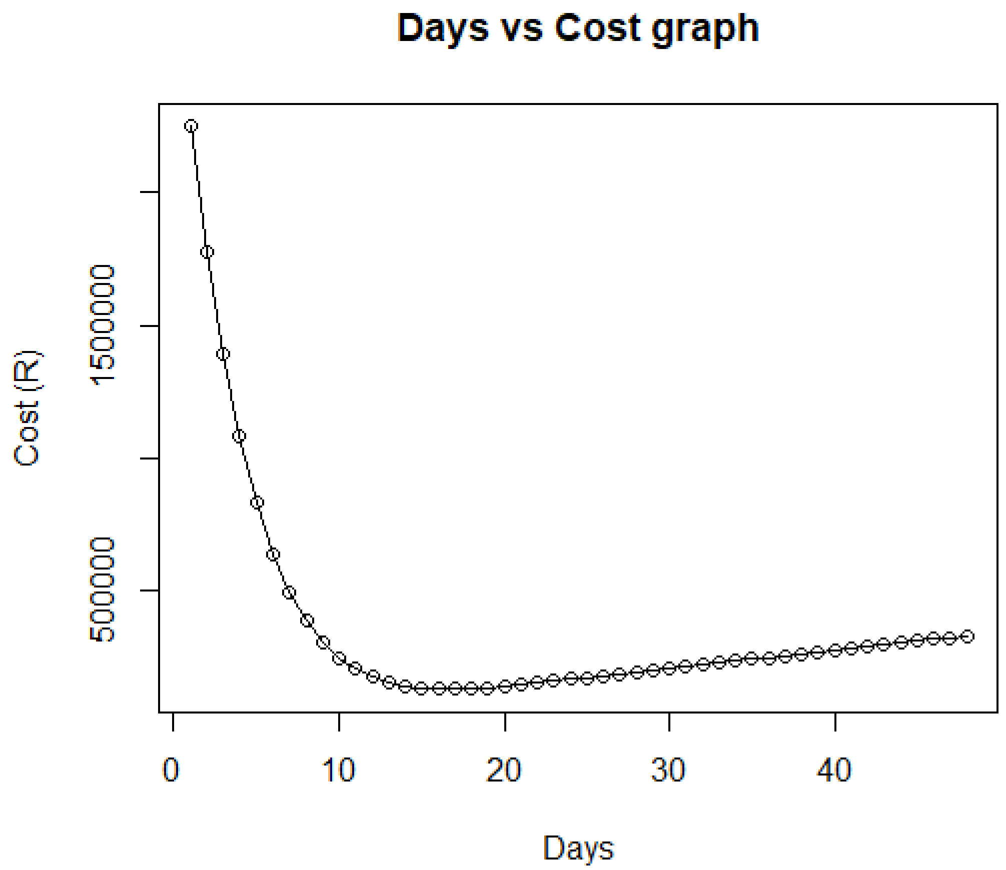


Figure 29: The days vs cost graph

Upon analysis of the graph, it can be deduced that the delivery process times for the technology class should be centred at 17 days in order to achieve maximum profit.

Likelihood of making a Type II error

In this section, the Type II error was calculated. Iw was determined that the probability of a type II error being made for A in Technology is 0.66. This means that the probability of assuming a process is running well when it isn't doing well is quite high. As such, from the data give, failures in the process are not easily identifiable.

5. DOE and MANOVA

In this section of the assignment, a hypothesis test using Multivariate Analysis of Variance (MANOVA) was conducted using data that consists of the technology class.

For the testing, the question that was used to establish the testing is as follows: Does the age of customers and pricing of the products sold affect the reason our customers buy a certain technological product? As such, our null hypothesis (H_O) is "the age of customers and pricing of technological products affect the reason why a certain technological product is bought." Our first alternate hypothesis (H_1) is then "the age of customers does not affect the reason why a certain technological product is bought." Our second alternate hypothesis is "the pricing of the technological products does not affect the reason why a certain technological product is bought." Lastly, our third alternate hypothesis (H_3) is "the age of customers does not affect the reason why a certain technological product is bought."

For analysis and discussion, we shall use a significance level of 0.05. As such, $\alpha = 0.05$.

Results and discussion of the MANOVA results

Using R programming, the whole testing was conducted and the results are as follows:

In response to Age:					
	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Why.Bought	5	1783	356.67	1.1867	0.3128
Residuals	13716	4122212	300.54		

In response to Price:					
	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Why.Bought	5	576810000	115362171	0.5674	0.7251
Residuals	13716	2.7889E+12	203334294		

For analysis, the significant level is compared to the Pr values in both tables. If the two Pr values are lower than the significant level (α), then H_3 cannot be rejected. If one of the Pr values is smaller than the significant level, then either H_1 or H_2 cannot be rejected. If the two Pr values are higher than the significant level, then the null hypothesis cannot be rejected.

From the results in the tables above, it can be deduced that the Pr value linked to age and the Pr value linked to price are larger than the significant value, meaning that the that the null hypothesis cannot be rejected. In conclusion, the age of customers and pricing of the products doesn't affect the reason our customers buy a certain technological product.

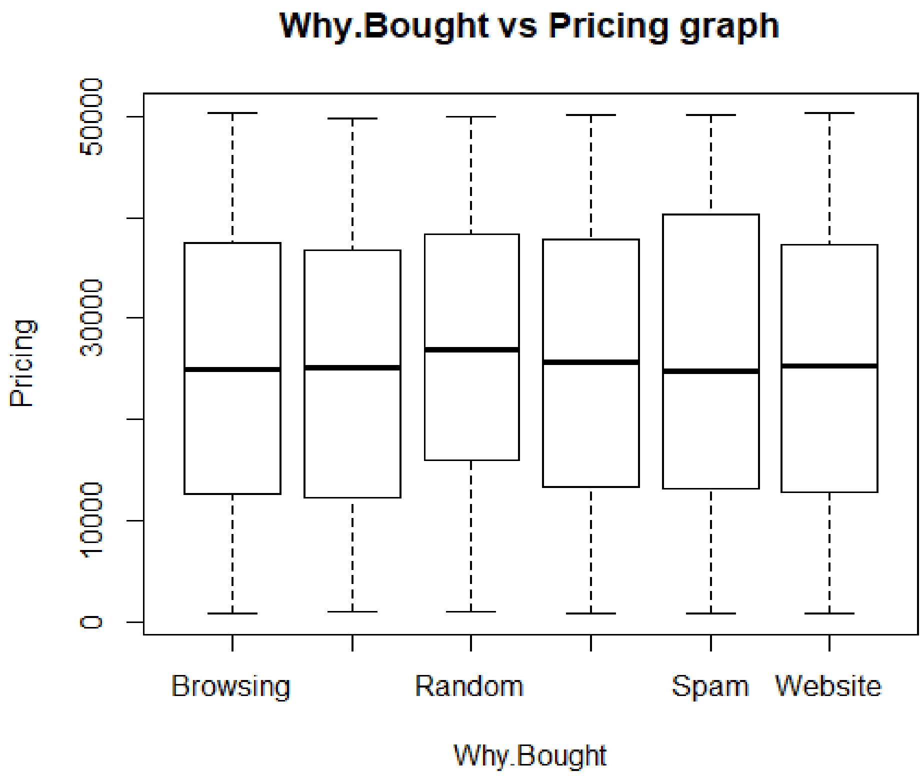


Figure 23: The Why.Bought vs Pricing graph

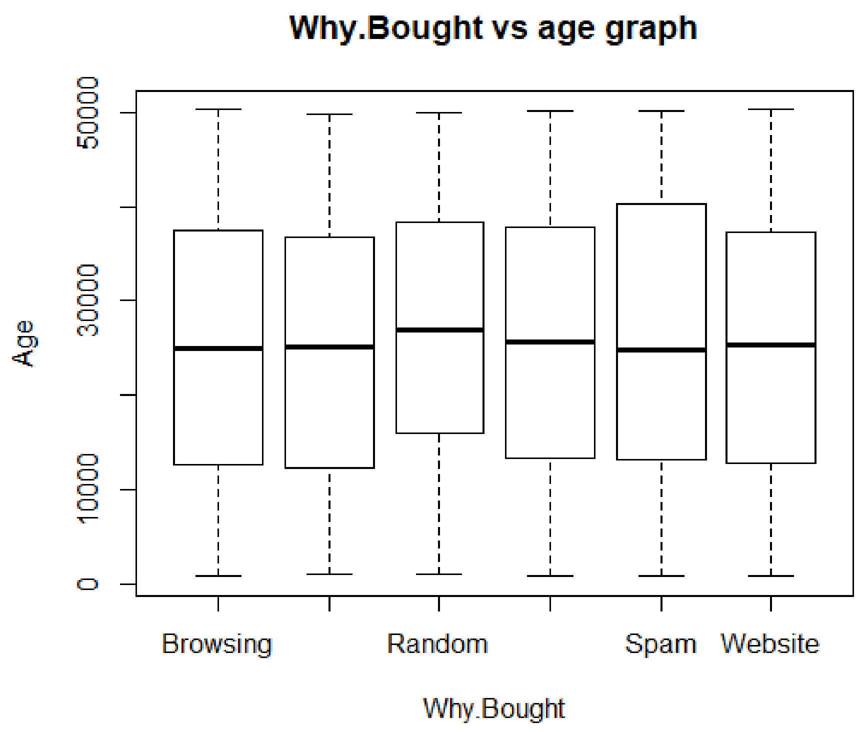


Figure 3: the Why.Bought vs Age graph

Upon analysing these two graphs, we see that there isn't any change or relationship that can be deduced.

6. Reliability of the service and products

In this section, exercises from the textbook are done.

The Taguchi Loss function

Problem 6

For the blueprint specification, we are given a tolerance of 0.035 cm and a scrap cost of $30. Using these two values, we can get the k constant for the Taguchi loss function.

Therefore: $$k = \frac{30}{(0.035)^2}$$

$$k = \$24489.8$$

The Taguchi Loss function is as follows:

$$L(x) = 24489.8\,(x - T)^2$$

Problem 7

a. Similar to problem six, we need to get the k constant. Therefore:

$$k = \frac{25}{(0.035)^2}$$

$$k = \$20\,408.16$$

The Taguchi function can be written as follows:

$$L(x) = \$20\,408.16(x - T)^2$$

b. The problem states that the process deviation from target can be reduced to 0.027 cm. the Taguchi loss can then be computed as follows:

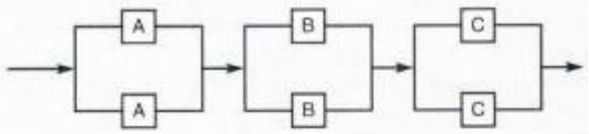
$$L(x) = \$20\,408.16(0.027 - 0.040)^2$$

$$= \$20\,408.16 \times 0.000169$$

$$= \$3.45$$

Reliability

For reliability, we are given problem 27 in chapter 7 to solve. A picture of problem 27 is given below.



Production System

The reliabilities of the machines are as follows:

Machine	Reliability
A	0.85
B	0.92
C	0.90

a. We are asked to analyse the system reliability, assuming that there is only machine operating at each operation section. The system reliability is calculated as follows:

$$System\ reliability(R_S) = R_A \times R_B \times R_C$$
$$= 0.85 \times 0.92 \times 0.90$$
$$= 0.704$$

b. We need to calculate the reliability of the system when both machines at each of the three stages are working. For this, the method of calculating reliability using series calculations is used in conjunction with the parallel systems method of calculating reliability.

$$System\ reliability(R_S) = R_A \times R_B \times R_C$$
$$= (1 - (1 - 0.85)^2) \times (1 - (1 - 0.92)^2) \times (1 - (1 - 0.90)^2)$$
$$= 0.9775 \times 0.9936 \times 0.99$$
$$= 0.961$$

From the two results given above, it can be deduced that the reliability of the system given greatly improves when both machines at each stages works.

Binomial probabilities

In this section of the work, the probability of how many days per year can reliable delivery times be expected is calculated and analysed.

In the problem statement, the following binomial formula needs to be used:

$$f(x) = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

The issue that arises is that the p-value is not given. Using R programming, it was determined that the value of the p-value is 0.02.

Using the formula above and the p-value using R programming, it is deduced that out of 365 days in a year, we can expect 194 days to consist of reliable delivery times. The increment of the number of vehicles to 16 is not a big one and as such one can assume that our answer will change. Therefore, if our vehicle fleet increases to 16, then the number of days in which reliable delivery times occurs will still be 194 days.

Conclusion

The data given had many missing values and as such data wrangling had to be conducted, in order to ensure that the analysis that deduced is of good quality and had integrity. After the wrangling, a data quality report was compiled, explaining the measures of central tendencies (mean, mode, range etc) that each feature of the dataset contained. Furthermore, a visual representation of what type of distribution that each feature has is given. Process capability indices for the data under technology is given and it was determined that the process is capable of meeting their specification due to the process variation being smaller than the specification range.

In the third and fourth section of the assignment, X-bar charts and s-control charts were analysed. It was determined that although the business is generally capable of meeting the specifications, in some cases specifications are not met. This is assumed to be due to change of drivers, breakdown of vehicles, etc. Trends were also analysed. Multivariate analysis of variables was also conducted. It was determined that Type I errors and Type II errors with Type II errors having a greater probability of occurring. With this result, it was determined that failures and issues weren't easily identifiable in the process.

Lastly, exercises from the prescribed textbook were done.

References

Evans, J.R. & Lindsay W.M. Managing for Quality and Performance Excellence. 10th Edition. United States: Conveo