2022

Quality Assurance 344

ECSA PROJECT REPORT

SHANE-LEE KUHN (20002262)

21 OCTOBER

Abstract

Industrial engineering students at Stellenbosch University are required to meet various requirements which are set by the Engineering Council of South Africa (ECSA) before they are awarded their degrees. In accordance with these requirements, the module Quality Assurance 344 specifically assesses what is known as "Graduate Attribute 4" which relates to the students ability to partake in investigations, experiments and data analysis. This report aims to demonstrate that the student has met the requirements that have been outlined by ECSA.

Table of Contents

List of Figures

List of Tables

Introduction

This report aims to demonstrate that students are able to fulfil the expectations of ECSA by developing an R script that can pre-process a large dataset that has been provided, perform exploratory visual analysis on the dataset apply the relevant methods of Statistical Process Control to the dataset and draw appropriate conclusions. The dataset specifically deals with client data of an online business. The data is then analysed to improve the overall functionality of the business and the experience of the clients.

Part 1: Data Wrangling

A large sales dataset that initially consisted of 180,000 instances and 10 features was provided. The dataset was manipulated in R and separated into valid data and invalid data. The valid data contained no missing values and no negative prices. The invalid data consisted of instances containing at least one missing value. A primary key was added to both valid and invalid datasets which kept track of the row number. Additionally, a secondary key was left in both datasets to allow one to trace which instance of the original dataset one was viewing. The valid dataset consists of 179,978 instances and 11 features while the invalid dataset consists of 17 instances and 11 features.

Part 2: Descriptive Statistics

The valid dataset was analysed in R in order to become acquainted with the data. Graphs of certain features are plotted in the figures which follow below and appropriate commentary is provided. It was decided to break the analysis of part 2 up into single feature analysis and multi-feature analysis.

Single Features

The plots and commentary that relate to the data analysis performed on single features follow below.

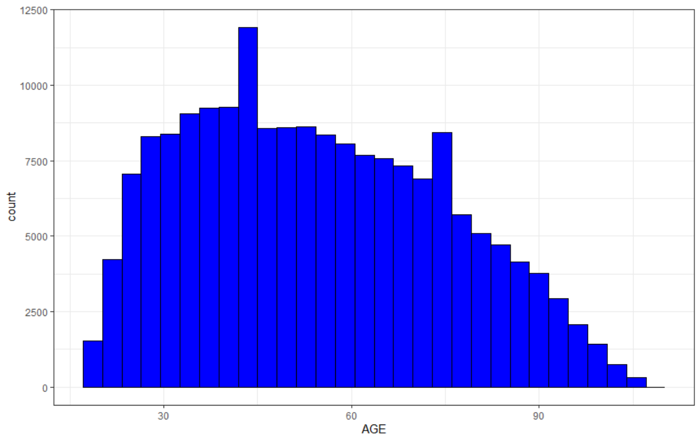


Figure 1: Age Distribution

A unimodal distribution of Age appears above in Figure 1. The distribution is skewed to the right. This means that the probability of customers buying from the client decreases as age increases once the peak of the distribution is passed at around the age of 40. This is expected of course, since many people in the population do not reach age of 90, for example.

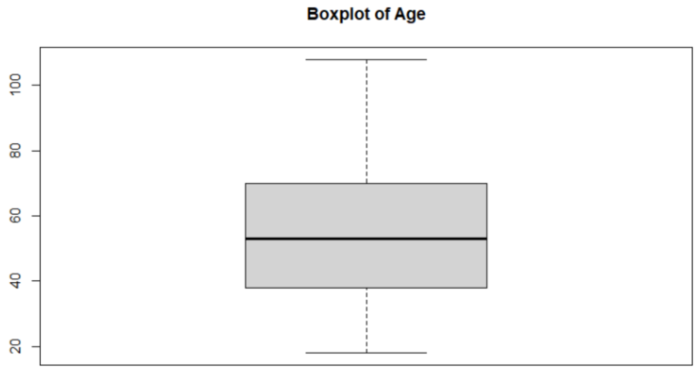


Figure 2: Boxplot of Age

From the boxplot above in Figure 2 it can be seen that the length of the top whisker is a bit longer than the length of the lower whisker. This corroborates the statement made previously that the distribution of age is skewed to the right.

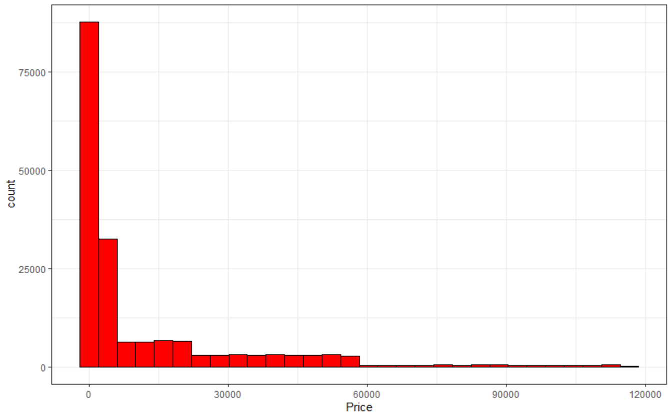


Figure 3: Distribution of Price

It can be seen above in Figure 3 that the Price feature exhibits an exponential distribution. This is not surprising since that means that most of the items that have been purchased by customers have been on the lower end of the price spectrum. In any business it is generally the case that most people do not buy the items that have higher prices since they might not be able to afford it or might feel like it is a waste of money to spend so much on a particular item.

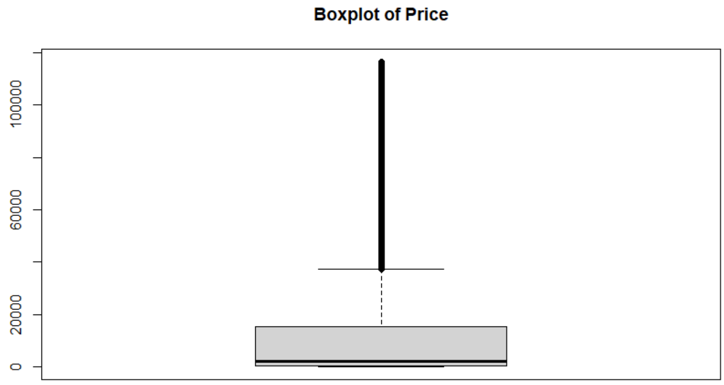


Figure 4: Boxplot of Price

The boxplot of Price in Figure 4 indicates that there are a number of outliers which is characteristic of an exponential distribution. This is most likely due to the high prices of items such as the Luxury items. It can also be seen that there is no lower whisker since there are no negative prices in the dataset.

The distribution of the delivery time feature can be seen below in Figure 5. This feature showed a multimodal distribution. Based on the project brief, it was assumed that the values included in the delivery time feature were given in hours. When considering the distribution below it can be seen that there are two classes of delivery items. There is one class of items that can be said to get delivered roughly within 1 day (24hours) and another class of items that generally take between 1 to 2.5days (24 to 60 hours) to be delivered.

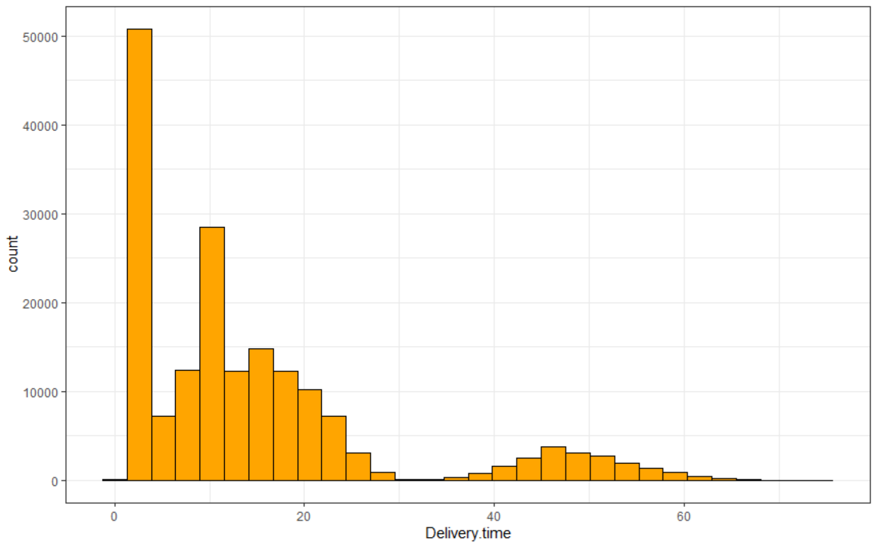


Figure 5: Distribution of Delivery Times

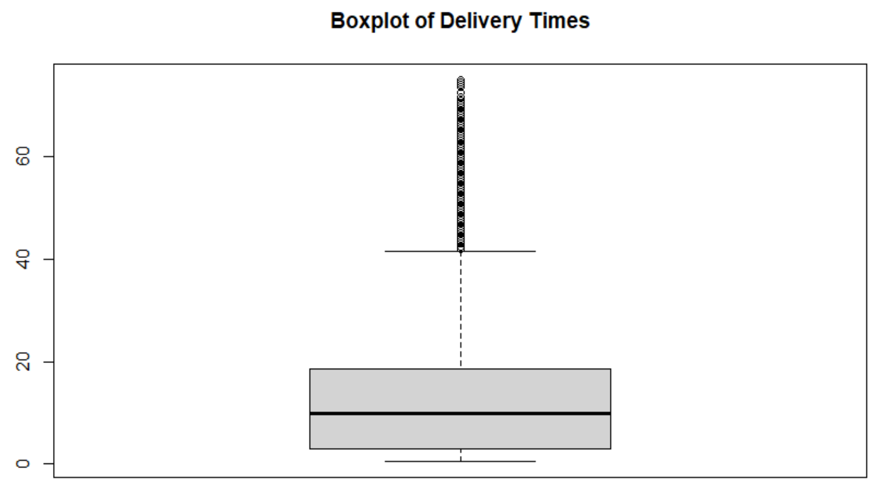


Figure 6: Boxplot of Delivery Times

The boxplot of the delivery times feature showed that there were numerous outliers in this feature, which can be seen in Figure 6. All these outliers fall into the second class of items mentioned above that take longer than 1 day to deliver.

Multiple features

The plots and commentary that relate to the data analysis performed by comparing multiple features follow below.

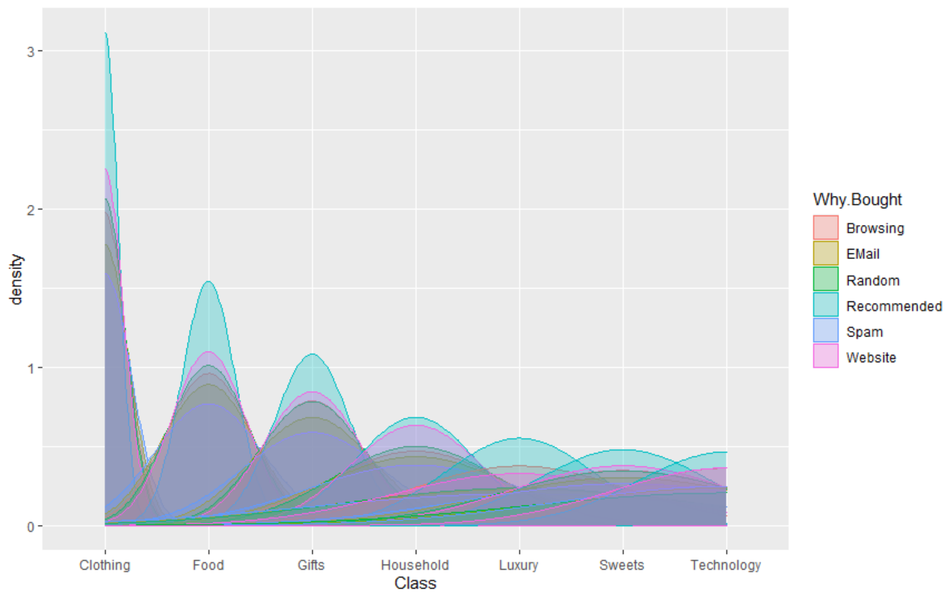


Figure 7: Density Distribution of Class

A density plot of the class feature appears above in Figure 7. This plot depicts the reason why customers buy all the different classes of items. The plot indicates the main reason for people buying from the client is due to recommendations. The reason that contributes to the second most sales of items, apart from the Luxury items, is attributed to Website.

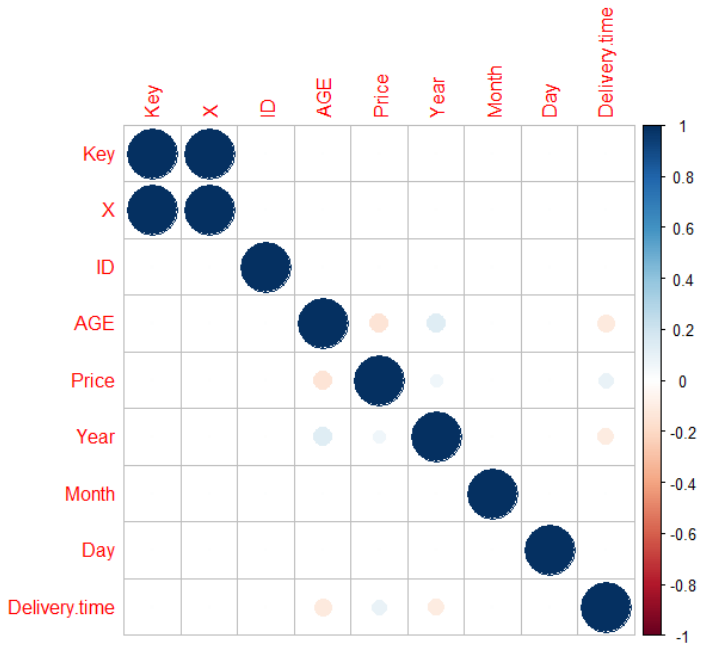


Figure 8: Correlation Matrix of numerical features

The correlation matrix above in Figure 8 shows the relationship between all the numerical features. As can be seen on the y-axis on the right-hand side, darker blue indicates a strong relationship. Based on the correlation matrix there are no features that are strongly correlated other than “Key” and “X” but this, of course, is expected given the way “Key” was created. There seem to be no strong correlations that might be useful.

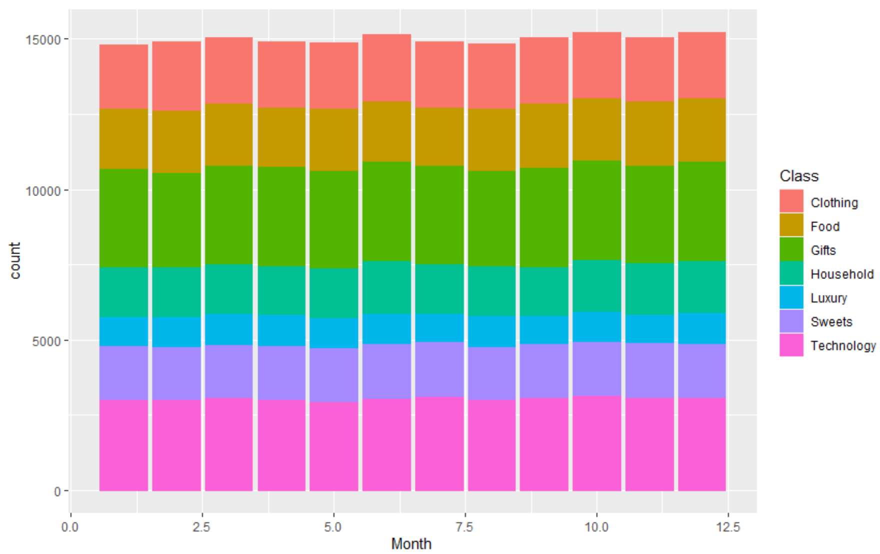


Figure 9: Plot of the variation of Class items across month

The plot above in Figure 9 shows how much of each class of item was sold across all 12 months over the years provided in the dataset. There seems to be little variation across months since the amount of each item sold seems to remain relatively constant. The plot seems to indicate that gifts are the items that generally sold the most and the Technology Items sold the second most. Management should note that these two items are the most popular and try to capitalize on this by maybe doing promotions of these items from time to time to attract even more buyers.

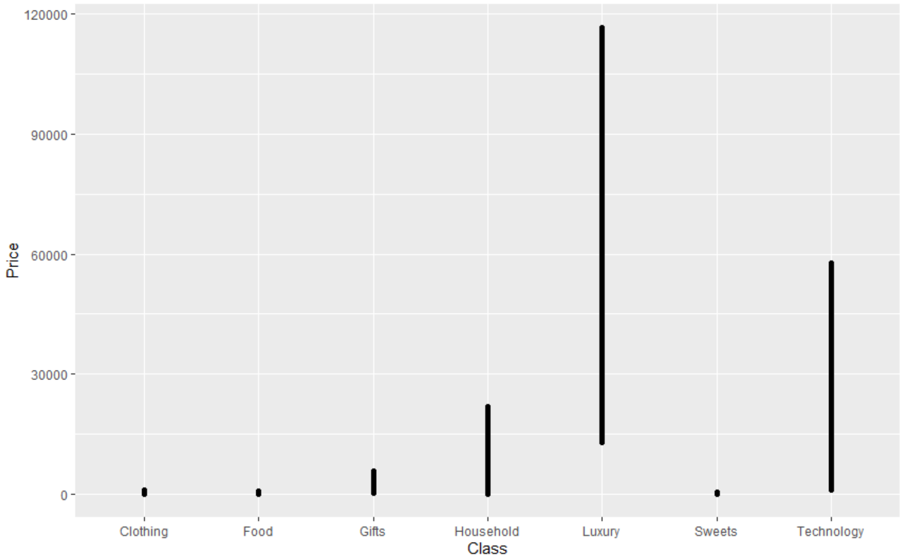


Figure 10: Plot of Class vs Price

The plot above in Figure 10 shows the relationship between Price and Class. It can be seen that the Luxury items are generally the most expensive items. This confirms the hypothesis that was made under Figure 4 previously, where it was hypothesized that the Luxury items were contributing to the outliers in the boxplot of the Price feature. It can also be seen that the Technology items are the second most expensive items.

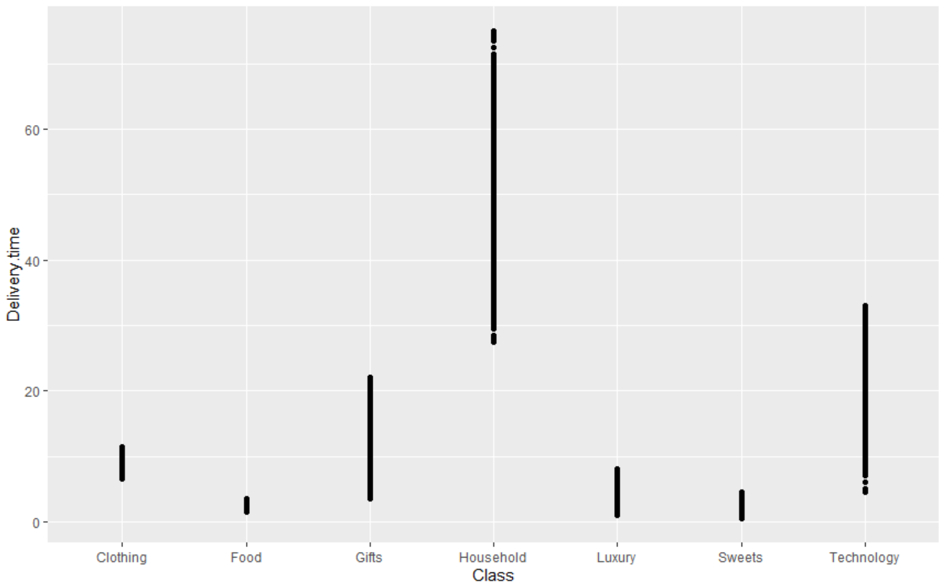


Figure 11: Plot of Class vs Delivery times

The plot above in Figure 11 shows the relationship between Class and Delivery time. It can be seen that household items generally take the longest to get delivered. The items that have the second longest delivery times are the technology items. In Part 4 of this report, it will be shown how the average delivery time for technology items can be improved.

Process Capability for the process delivery times of technology class items

It was assumed that USL = 24 hours and LSL = 0. It should be mentioned that LSL is logical since it is the lowest level that a measurement can reach and be acceptable to the client. The mean and standard deviation were calculated and the values obtained were $\mu = 20.011$ and $\sigma = 3.502$. The Process Capability Indices were then calculated using the relevant values and tabled below in Table 1.

Process Capability Index	Value
C_p	1.142
C_{pu}	0.380
C_{pk}	0.380
C_{pl}	1.905

Table 1: Process Capability Indices

It can be seen from Table 1 above that the Capability Performance has a low value of 0.380. Generally, a value of 1 means that the process is barley capable. Seeing that this value is well below 1, management should investigate what is causing this.

Part 3: Statistical Process Control

The valid data was arranged by year, month and day and ordered from the oldest to the latest. 30 samples of 15 instances were made use of to calculate various control limits. The Control charts that resulted from this analysis appears below in Figure 12-15.

X-bar chart for the delivery process times of all classes. [Samples 1-30]

	myClassNames	UCLXT	U2SIGMAT	U1SIGMAT	CLXT	L1SIGMAT	L2SIGMAT	LCLXT
1	Clothing	9.40493352386633	9.25995568257756	9.11497784128878	8.97	8.82502215871122	8.68004431742245	8.53506647613367
2	Household	50.2483278659662	49.0196259847182	47.7909241034702	46.5622222222222	45.3335203409742	44.1048184597263	42.8761165784783
3	Food	2.70580826490218	2.63414037583517	2.56247248676816	2.49080459770115	2.41913670863414	2.34746881956713	2.27580093050012
4	Technology	22.9746158797126	22.1078920679566	21.2411682562005	20.3744444444444	19.5077206326884	18.6409968209323	17.7742730091763
5	Sweets	2.89310560598847	2.75632327755553	2.61954094912259	2.48275862068966	2.34597629225672	2.20919396382378	2.07241163539084
6	Gifts	9.48856467334077	9.11274681926422	8.73692896518766	8.36111111111111	7.98529325703456	7.60947540295801	7.23365754888145
7	Luxury	5.49396512637278	5.24116193610037	4.98835874582796	4.73555555555556	4.48275236528315	4.22994917501074	3.97714598473833

Figure 12: X- chart values for samples 1-30

S-Chart for the delivery process times of all classes. [Samples 1-30]

	myClassNames	UCLS	U2SIGMAS	U1SIGMAS	CLS	L1SIGMAS	L2SIGMAS	LCLS
1	Clothing	0.866559568463719	0.761455227250562	0.656350886037405	0.551246544824249	0.446142203611092	0.341037862397935	0.235933521184778
2	Household	7.34418006586244	6.45341013420991	5.56264020255739	4.67187027090486	3.78110033925233	2.89033040759981	1.99956047594728
3	Food	0.428372325526015	0.376415376798093	0.324458428070171	0.272501479342249	0.220544530614327	0.168587581886404	0.116630633158482
4	Technology	5.18056970372824	4.55222240293678	3.92387510214531	3.29552780135385	2.66718050056238	2.03883319977091	1.41048589897945
5	Sweets	0.817573461203726	0.71841060713067	0.619247753057614	0.520084898984558	0.420922044911502	0.321759190838447	0.222596336765391
6	Gifts	2.24633333311156	1.9738772969496	1.70142126078763	1.42896522462567	1.15650918846371	0.884053152301749	0.611597116139788
7	Luxury	1.51105176847233	1.32777746576534	1.14450316305835	0.961228860351357	0.777954557644365	0.594680254937373	0.411405952230381

Figure 13: S-chart values for samples 1-30

X-bar chart for the remaining delivery process times of all classes. [Samples 31-end]

	myClassNames	UCLS	U2SIGMAS	U1SIGMAS	CLS	L1SIGMAS	L2SIGMAS	LCLS
1	Clothing	0.955583584833818	0.839681589387477	0.723779593941136	0.607877598494795	0.491975603048454	0.376073607602113	0.260171612155772
2	Household	8.31397709212241	7.30558122768347	6.29718536324454	5.2887894988056	4.28039363436667	3.27199776992773	2.2636019054888
3	Food	0.438386897396513	0.385215288976071	0.332043680555629	0.278872072135187	0.225700463714744	0.172528855294302	0.11935724687386
4	Technology	5.41575511914017	4.75888236083903	4.10200960253789	3.44513684423675	2.78826408593561	2.13139132763447	1.47451856933333
5	Sweets	0.805107617784373	0.707456736238007	0.609805854691641	0.512154973145275	0.41450409159891	0.316853210052544	0.219202328506178
6	Gifts	2.31461677296082	2.03387869108347	1.75314060920612	1.47240252732877	1.19166444545142	0.910926363574064	0.630188281696713
7	Luxury	1.22567582091096	1.07701454661896	0.928353272326961	0.779691998034961	0.631030723742962	0.482369449450963	0.333708175158964

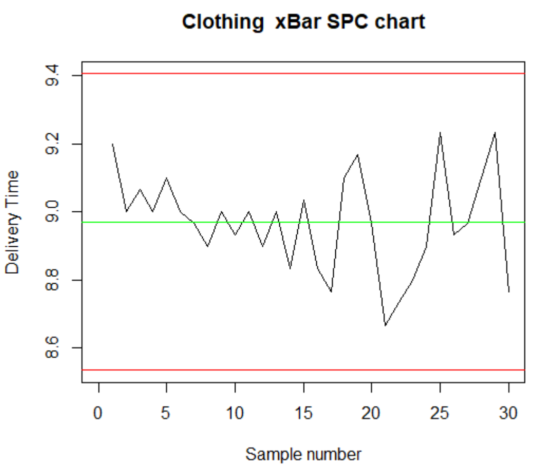
Figure 14: X-chart values for samples 31-end

S-Chart for the remaining delivery process times of all classes. [Samples 31-end]

	myClassNames	UCLS	U2SIGMAS	U1SIGMAS	CLS	L1SIGMAS	L2SIGMAS	LCLS
1	Clothing	0.955583584833818	0.839681589387477	0.723779593941136	0.607877598494795	0.491975603048454	0.376073607602113	0.260171612155772
2	Household	8.31397709212241	7.30558122768347	6.29718536324454	5.2887894988056	4.28039363436667	3.27199776992773	2.2636019054888
3	Food	0.438386897396513	0.385215288976071	0.332043680555629	0.278872072135187	0.225700463714744	0.172528855294302	0.11935724687386
4	Technology	5.41575511914017	4.75888236083903	4.10200960253789	3.44513684423675	2.78826408593561	2.13139132763447	1.47451856933333
5	Sweets	0.805107617784373	0.707456736238007	0.609805854691641	0.512154973145275	0.41450409159891	0.316853210052544	0.219202328506178
6	Gifts	2.31461677296082	2.03387869108347	1.75314060920612	1.47240252732877	1.19166444545142	0.910926363574064	0.630188281696713
7	Luxury	1.22567582091096	1.07701454661896	0.928353272326961	0.779691998034961	0.631030723742962	0.482369449450963	0.333708175158964

Figure 15: S-chart values for samples 31-end

An example of the graphical representations of the X-bar chart(1:30) and S-chart(1:End) for Clothing is seen in figure below. It was decided to include one graph for 30 samples and one graph for one to the end to give the reader a sense of what both look like. The X-charts(1:30) and S-charts(1:End) for the rest of the classes can be found in Appendix A.



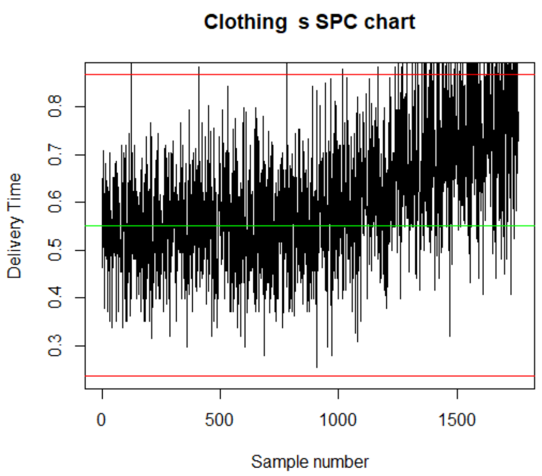


Figure 16: X-chart(1:30) and S-chart(1:End) for Clothing

Part 4: Optimising Delivery Processes

The delivery times of all the samples in the dataset were analysed in order to make appropriate deductions about the delivery process.

In the figures below, X-bar charts are plotted for all classes of items and tabled values of the total samples of each class appear, as well as the first 3 samples and the last 3 samples where applicable.

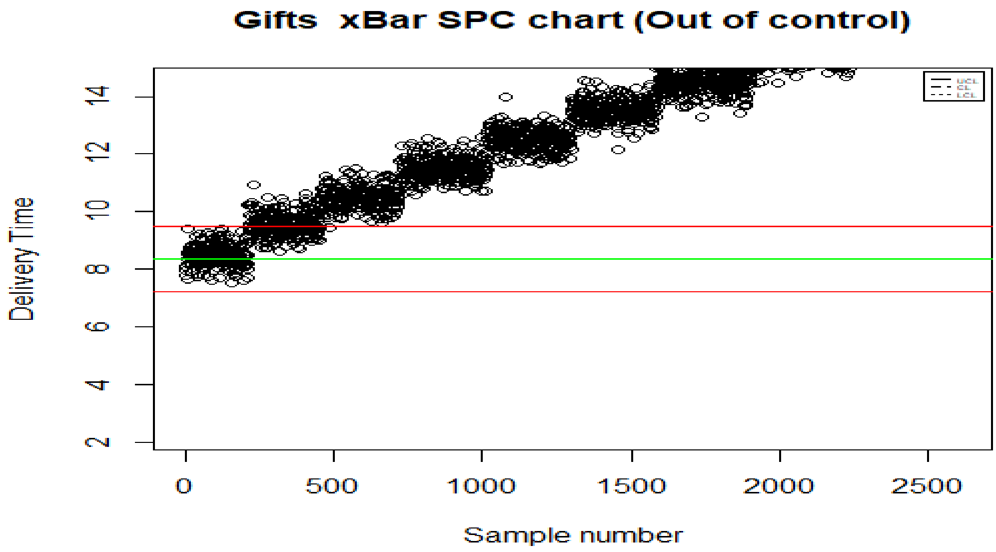


Figure 17: X-bar chart for Gifts

Total Samples	First 3 Samples	Last 3 Samples
2287	213, 216, 218	2607, 2608, 2609

Table 2: Out of control samples for Gifts

In the gifts class there was a concerning 2287 samples that were outside the outer control limits. All of the samples outside the outer control limits are above the UCL. This should be investigated by management with immediate effect.

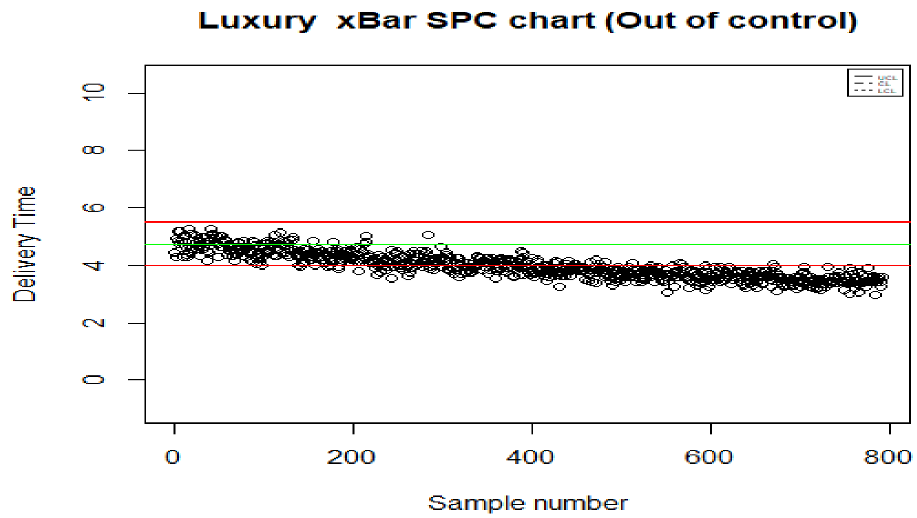


Figure 18: X-bar chart for Luxury

Total Samples	First 3 Samples	Last 3 Samples
440	142, 171, 184	789, 790, 791

For the Luxury class a large number of samples were outside of the outer control limits. All the samples that were outside the outer control limits were below the LCL. Management should look into why so many samples are not within Statistical Process Control.

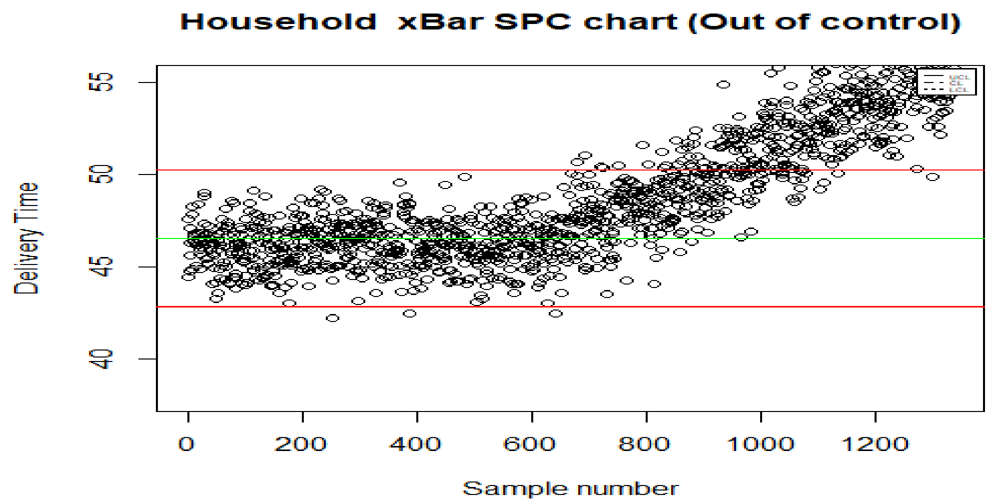


Figure 19: X-bar chart for Household

Total Samples	First 3 Samples	Last 3 Samples
395	252, 387, 643	1335, 1336, 1337

Table 3: Out of control samples for Household

For the Household class a large number of samples were outside the outer control limits as well. Most of these samples were above the UCL. This should be a concern for management and the process should be investigated.

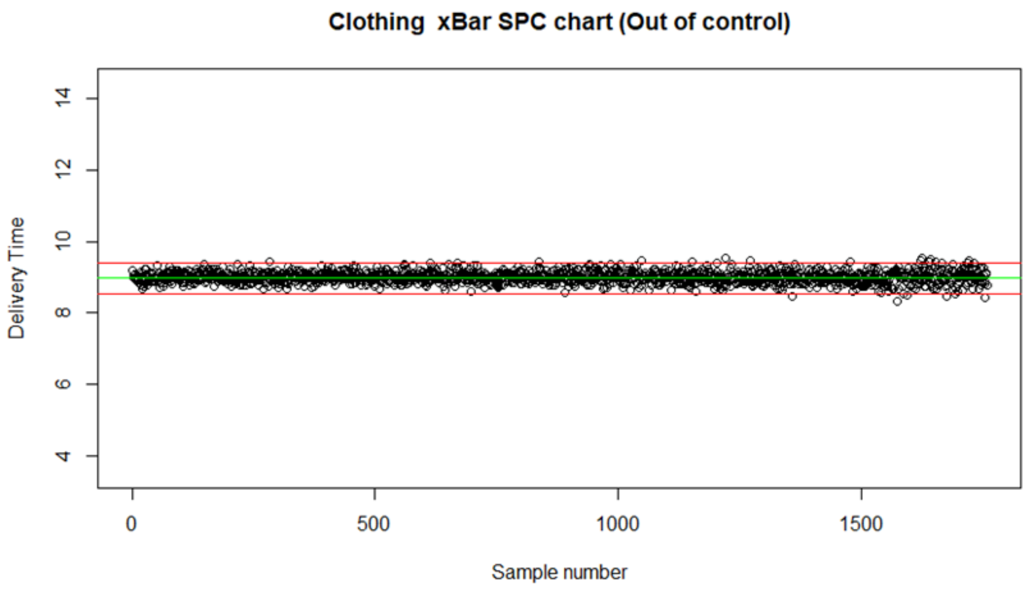


Figure 20: X-bar chart for Clothing

Total Samples	First 3 Samples	Last 3 Samples
20	282, 837, 1048	1695, 1723, 1756

Table 4: Out of control samples for Clothing

The clothing class showed 20 samples that were outside the outer control limits. Most of the samples that were outside the outer control limits were above the UCL. This process should probably be investigated by management.

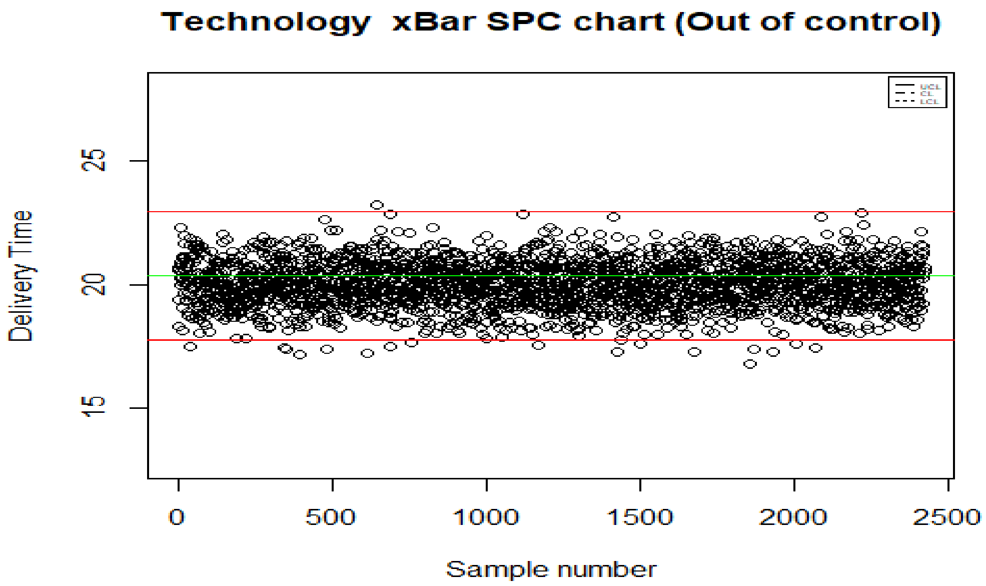


Figure 21: X-bar chart for Technology

Total Samples	First 3 Samples	Last 3 Samples
19	37, 345, 353	1933, 2009, 2071

Table 5: Out of control samples for Technology

The Technology class showed 19 samples that were outside the outer control limits. The vast majority of the samples that lied outside the outer control limits, except for one, were below the LCL. Management should probably investigate the process.

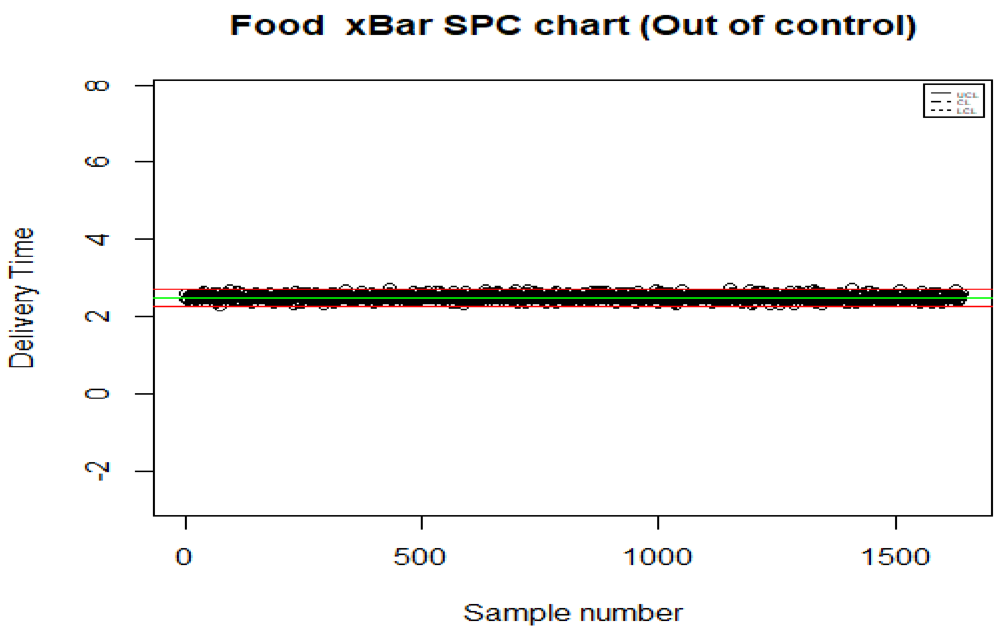


Figure 22: X-bar chart for Food

Total Samples	First 3 Samples	Last Sample
4	75, 432, 1149	1408

Table 6: Out of control samples for Food

The Food class showed that only 4 samples were outside of the control limits. Management should not investigate this process.

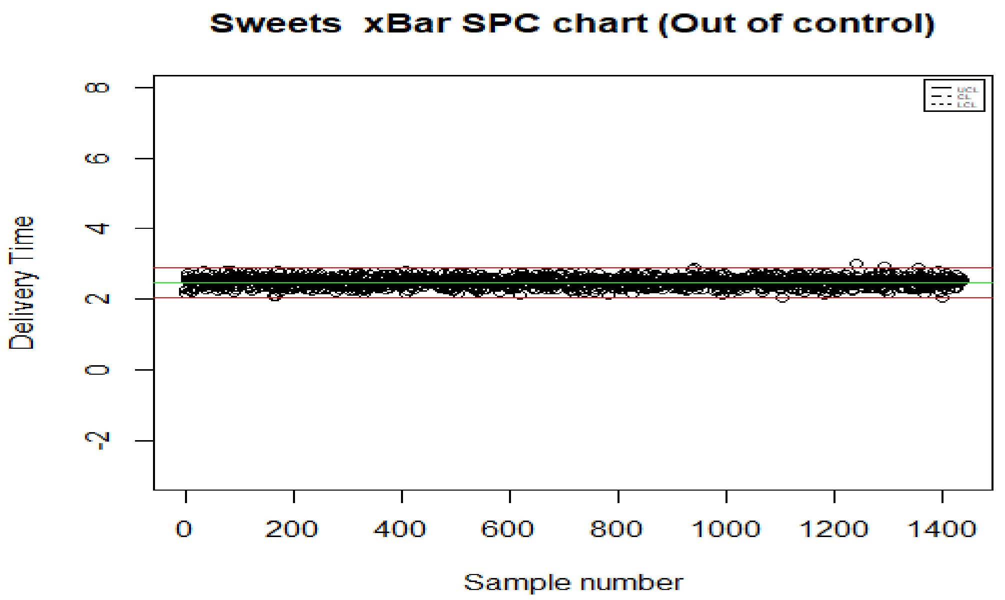


Figure 23: X-bar chart for Sweets

Total Samples	First 3 Samples	Last Sample
4	942, 1243, 1294	1358

Table 7: Out of control samples for Sweets

The sweets class also showed that only 4 samples were outside of the control limits. Management should not investigate this process.

<u>4.1 B: Most consecutive samples</u>

The most consecutive samples of "s-bar or sample standard deviations" between -0.3 and 0.4 sigma control limits and the ending sample numbers were found. It should be noted that where there were multiple occurrences of the most consecutive samples, all the ending sample numbers were provided. This can be seen below in Table 8.

Class	Most Consecutive Samples	Ending Sample Number
Technology	6	1776
Clothing	4	223, 1121
Household	3	45, 588, 647, 766, 843
Luxury	4	63
Food	4	85, 223, 640, 879
Gifts	7	2477
Sweets	5	316

Table 8: Most consecutive samples between -0.3 and 0.4 sigma control limits

4.2: Type 1 Error

The null hypothesis for a type 1 error is that the process is in control. If the samples are converted to a standard normal distribution with a mean of 0 and a standard deviation of 1, then LCL = -3 and the UCL = 3. The type 1 error was then calculated in R using pnorm(-3)*2, which yielded a value of 0.0027 or 0.27%. Therefore, the probability of management investigating a process that is actually stable is 0.27%.

4.3: Hours for best profit

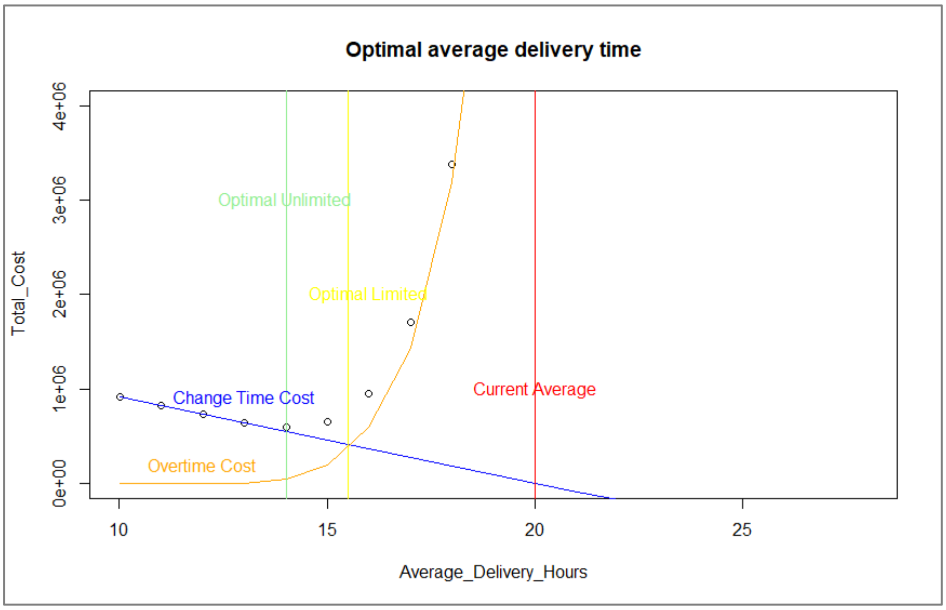


Figure 24: Optimal Delivery

As can be seen in Figure 24 above, the optimal average delivery time was calculated by finding the point where the cost to reduce the average time intersected with the cost to deliver technology items slower than 26 hours. It can be seen that the average delivery time was 20 hours for the technology class. It was then found that the optimal average delivery time, that would minimise the total costs incurred, was around 16 hours. This means that the business would need to reduce the current average delivery time by 4 hours. A total cost of R589306 will be incurred by the business as a result of this.

4.4 Type 2 Error

The type 2 error was calculated as seen in the figure below.

```
UCL_Tech<-XBarbarTotal[[4]]+0.789*SbarTotal[[4]] #22.974
LCL_Tech<-XBarbarTotal[[4]]-0.789*SbarTotal[[4]] #17.774
SD_Tech<-((TechUCL-TechLCL)/6) #0.8667
T2Error<-(pnorm(TechUCL,mean=23,sd=Standarddev)-pnorm(TechLCL,mean=23,sd=Standarddev))
## the probabilty of making a type 2 error is 0.48831767
```

Figure 25: R code extraction for type 2 error

Part 5: DOE and MANOVA

The MANOVA was performed on "Why.Bought" which is an indicator of the reasons why customers purchase items. The MANOVA was done to determine whether the factors of Age, Price and Delivery time influence the reasons that customers purchase items significantly.

Null hypothesis: The factors of Age, Price and Delivery time do not significantly affect why customers purchase certain classes of items.

Alternative hypothesis: At least one factor has a significant influence on the reason why customers purchase certain classes of items.

```
Descriptive:
   Why.Bought      n    AGE     Price  Delivery.time
1    Browsing  18990 53.845 16133.660         14.740
2       EMail   7224 55.757  6661.580         14.422
3      Random  13121 56.963  4288.261         14.179
4 Recommended 106975 54.482 13441.292         13.226
5        Spam   4208 54.659  9360.900         15.235
6     Website  29443 53.965 11017.647         19.033

Wald-Type Statistic (WTS):
           Test statistic df   p-value
Why.Bought "12950.513"    "15" "<0.001"

modified ANOVA-Type Statistic (MATS):
           Test statistic
Why.Bought       12412.12

p-values resampling:
           paramBS (WTS) paramBS (MATS)
Why.Bought "<0.001"      "<0.001"
```

Figure 26: MANOVA results of R code

In Figure 26 above the results of the MANOVA that was done R can be seen. Alpha was chosen to be 0.05 which means that a p-value of less than 0.05 would indicate that the test is significant. It can be seen that a p value of less than 0.001 was attained for the test which means that at least one factor does affect the reason why people buy certain classes of items. Boxplots of the factors of Price, Age and Delivery time against Why.Bought can be seen in the figures which follow below.

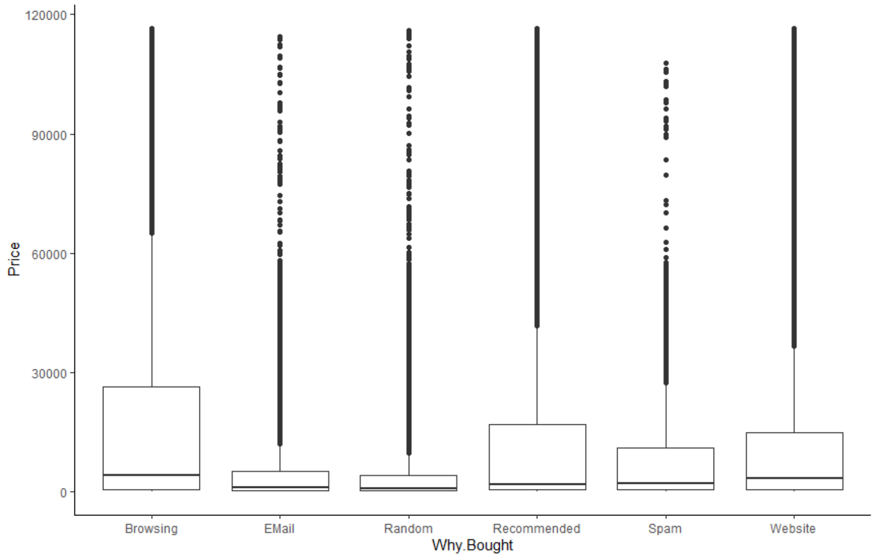


Figure 27: Price vs Why.Bought

In Figure 27 above it can be seen that Price has an impact on the mean value of the reasons why customers purchase items.

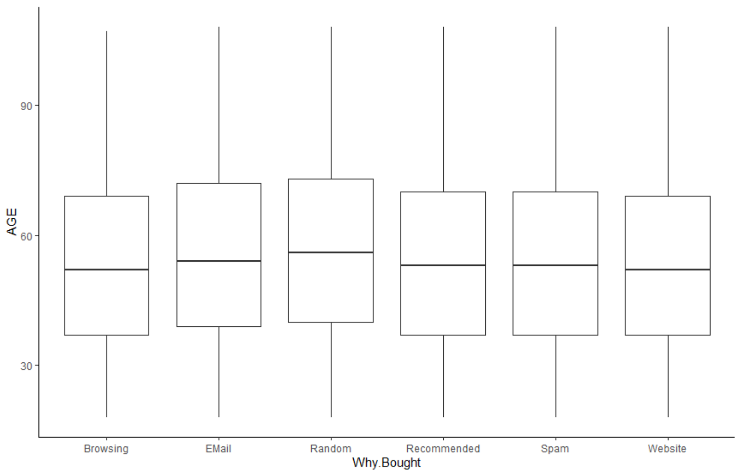


Figure 28: Age vs Why.Bought

It can be seen in Figure 28 above that Age has no significant effect on the reasons why people purchase certain items. The mean ages for all customers remain relatively constant across the reasons that influenced customers to purchase.

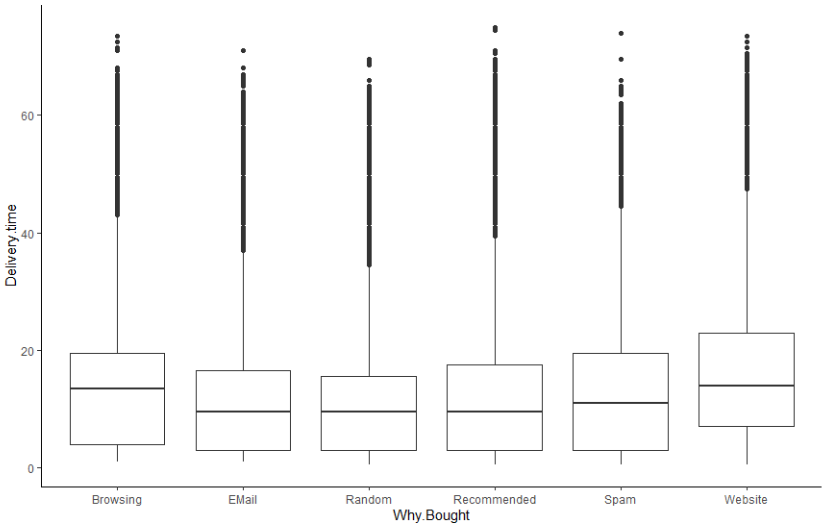


Figure 29: Delivery.time vs Why.Bought

In Figure 29 above it can be seen that Delivery time has a slight impact on the mean value of the reasons why customers purchase items.

Part 6: Reliability of Service and Products

<u>6.1</u>

<u>Problem 6 of Chapter 7</u>

It is given that the thickness of a refrigerator part at Cool Food Inc. is 0.06 ± 0.04 cm. It is also given that it costs $45 dollars to scrap a part that is not within specifications. The Taguchi loss function is as follows: $L = k(y\text{-}m)^2$

After substituting into the equation it is found that $L = 28125(y\text{-}0.06)^2$

<u>Problem 7 of Chapter 7</u>

a) If the cost to scrap a part reduces to 35$ the Taguchi loss function then becomes $L= 21875(y – 0.06)^2$.
b) If the process deviation from the target reduces to 0.027cm, the Taguchi loss becomes $L = 21875(0.027)^2 = 15.95$

<u>6.2</u>

<u>Problem 27 of Chapter 7</u>

Given the following information about the reliability of machines at Stages A,B, and C.

Machine	Reliability
A	0.85
B	0.92
C	0.90

Table 9: Machine reliability values

a) Assuming only one machine at each stage,

$R_A*R_B*R_C = 0.85*0.92*0.9 = 0.7038$

b) If both machines are functioning at all 3 stages,
Reliability = $(1 – (1 – 0.85)^2) * (1 – (1 – 0.92)^2) *(1 – (1 – 0.9)^2) = 0.9615$ or 96.15%

It is clear that by having 2 machines that function at each stage the reliability improves by 26%.

<u>6.3</u>

The dbinom() function was used in R which lead to the following probabilities being obtained:

The probability of a reliable number of vehicles = 0.92782

The probability of a reliable number of drivers = 0.94713

The number of reliable delivery days per year was then calculated by multiplying the weighted average of the reliable number of drivers and vehicles. The number of reliable days also increased by one if the number of vehicles increased by one, as can be seen in Table 10 below.

Number of Vehicles	**Number of drivers**	**Reliable delivery days per year**
21	21	343
22	21	344

Table 10: Reliable delivery days per year

Conclusion

The analysis done indicates that gifts and technology items are the items that are most popular amongst clients. Management should take not of this and do everything that they can to retain the customers who purchase these two items. The delivery times of the Gifts, Luxury and Household classes had many samples that were outside the outer control limits and management needs to investigate what the cause this is. It was also seen that the cost of delivering technology items can be reduced to a minimum by paying to reduce the average delivery time by 4 days. It is also clear that the business should have 2 machines which will improve reliability by 26%.

References

Frost, J., n.d. *Statistics by Jim.* [Online]
Available at: https://statisticsbyjim.com/basics/skewed-distribution/
[Accessed 9 October 2022].

Muniz, H., 2017. *PrepScholar.* [Online]
Available at: https://blog.prepscholar.com/skewed-right
[Accessed 9 October 2022].

Schalkwyk, T. v., n.d. *SUNLEARN.* [Online]
Available at:
https://learn.sun.ac.za/pluginfile.php/3514418/mod_resource/content/1/QA344%20Statistics.pdf
[Accessed October 2022].

Appendix A

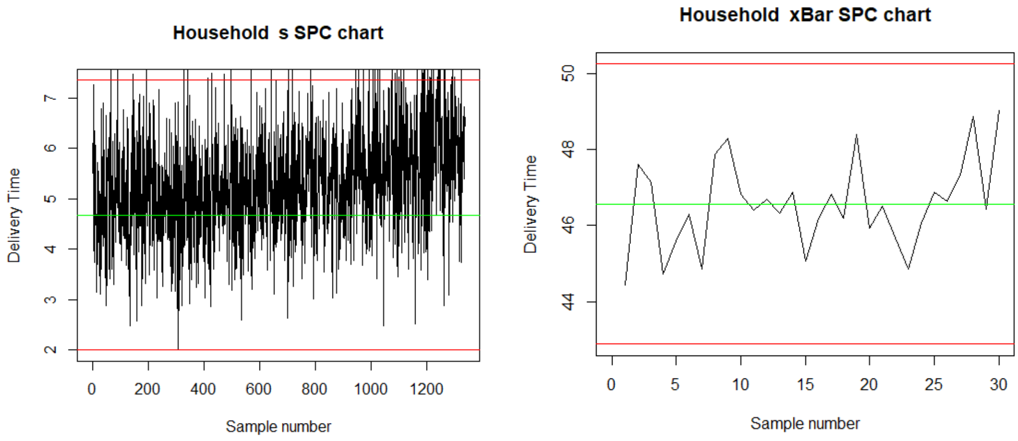


Figure 30: S-chart(1:End) and X-bar chart(1:30) of Household

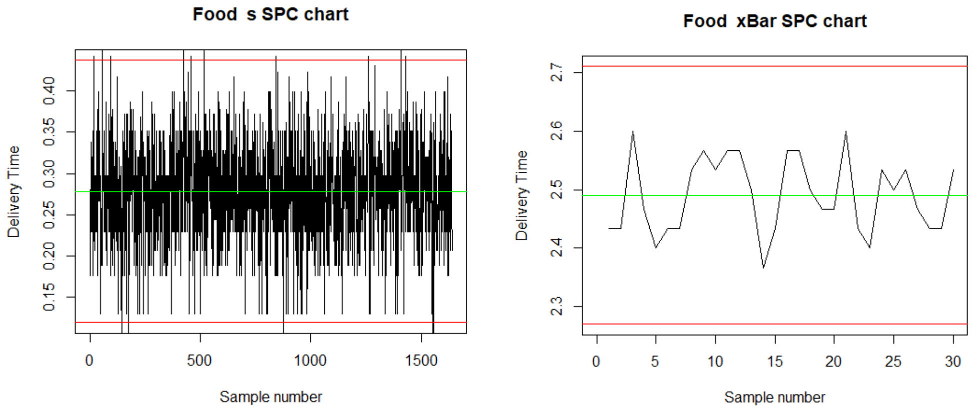


Figure 31: S-chart(1:End) and X-bar chart(1:30) of Food

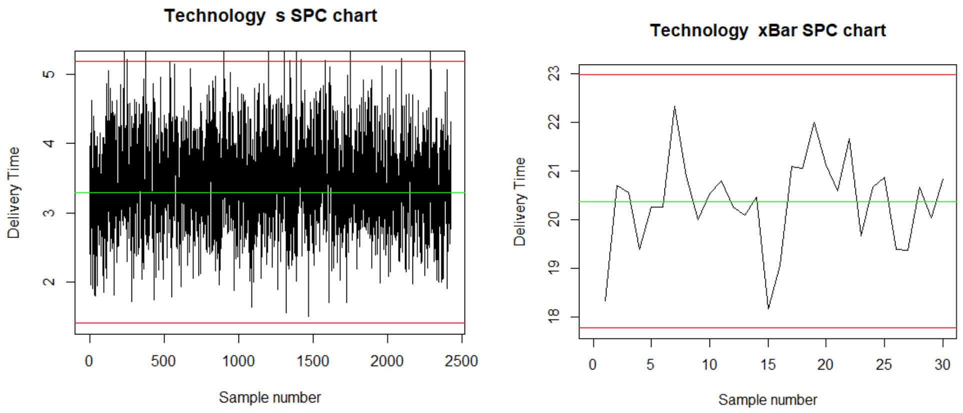


Figure 32: S-chart(1:End) and X-bar chart(1:30) of Technology

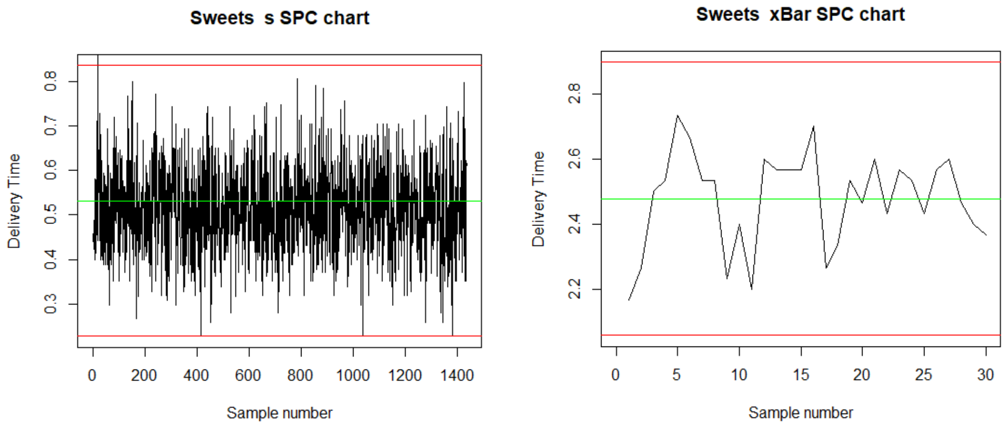


Figure 33: S-chart(1:End) and X-bar chart of Sweets

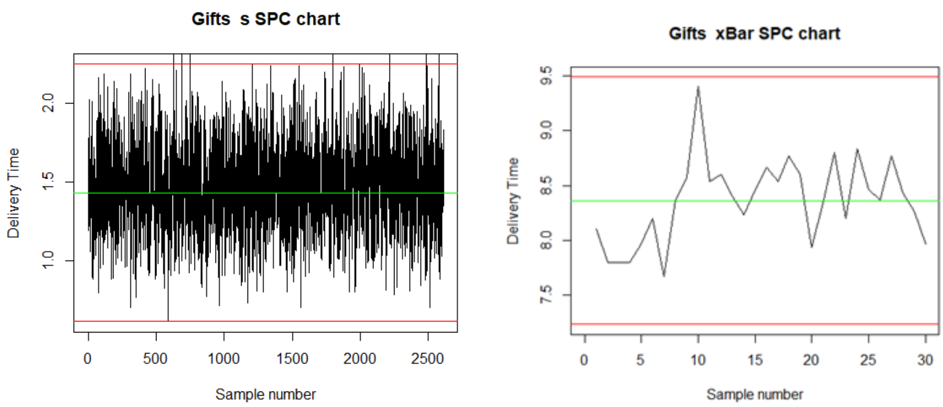


Figure 34: S-chart(1:End) and X-bar chart(1:30) of Sweets

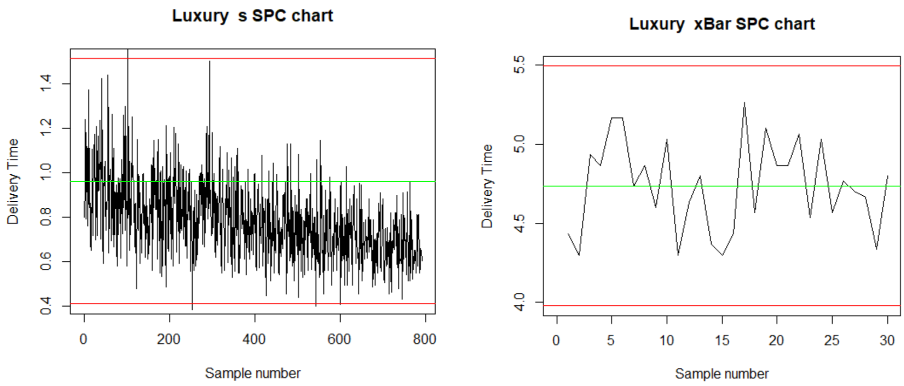


Figure 35: S-chart(1:End) and X-bar chart(1:30) for Luxury

Appendix B

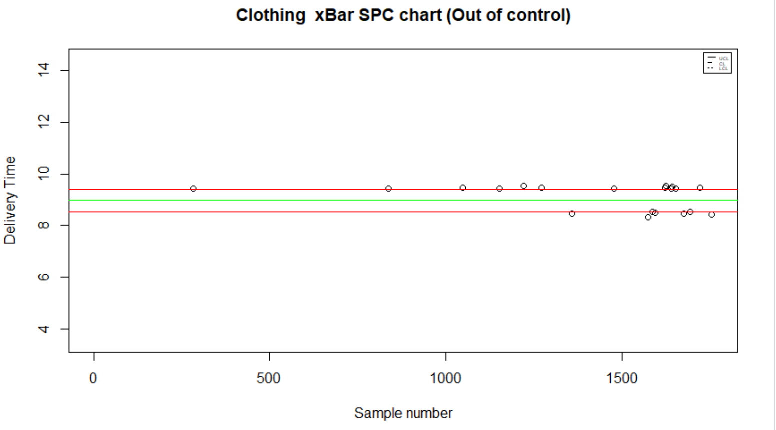


Figure 36: X-bar chart for out of control Clothing samples

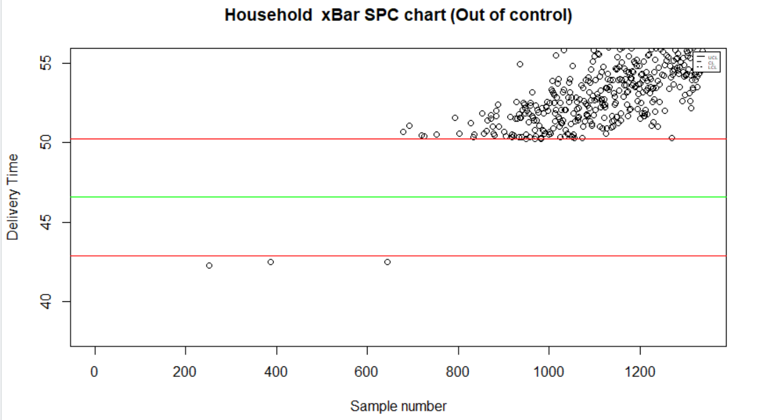


Figure 37: X-bar chart for out of control Household samples

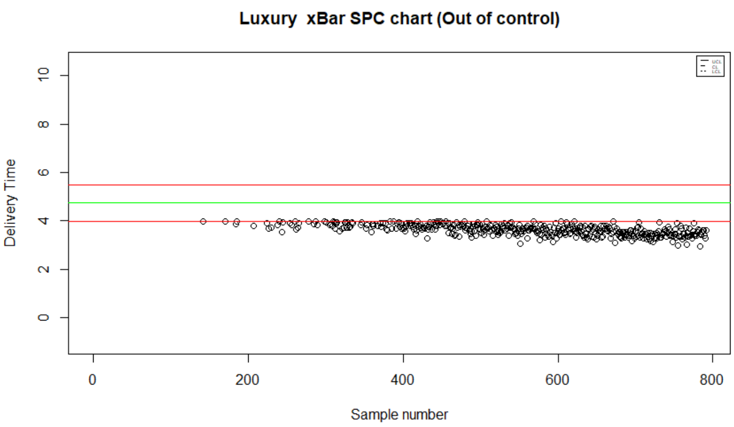


Figure 38: X-bar chart for out of control Luxury samples

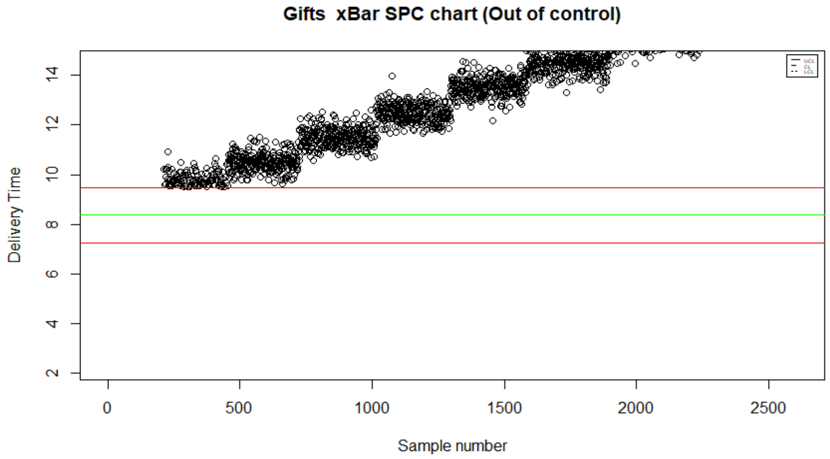


Figure 39: X-bar chart for out of control Gifts samples

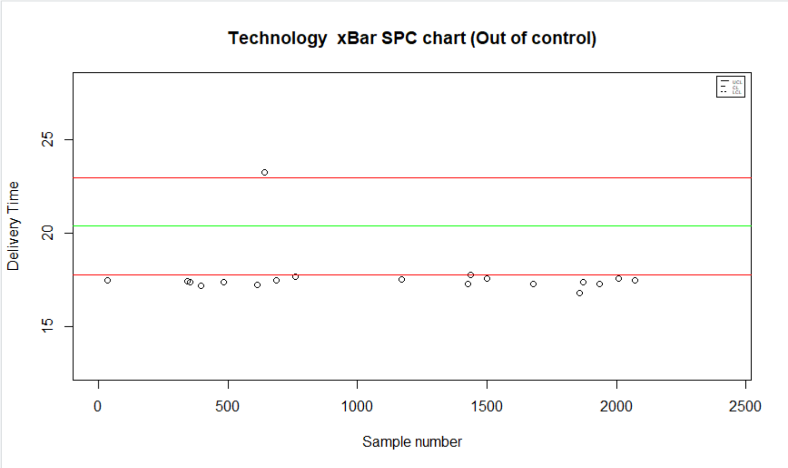


Figure 40: X-bar chart for out of control Technology samples

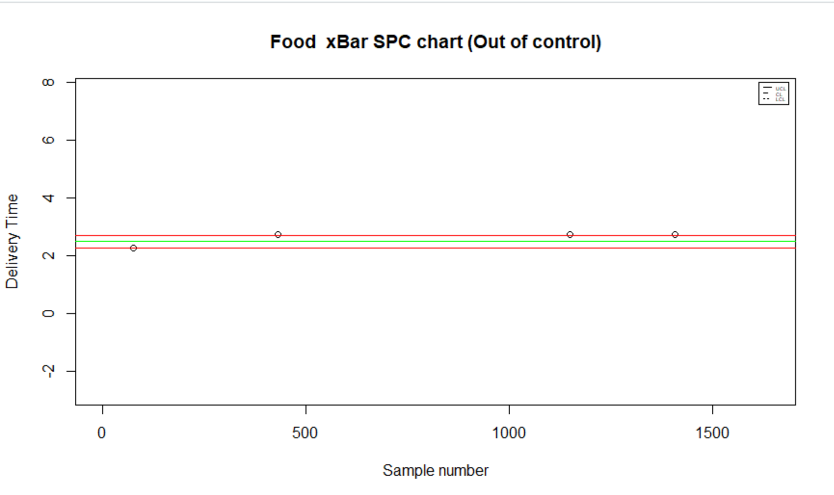


Figure 41: X-bar chart for out of control Food samples

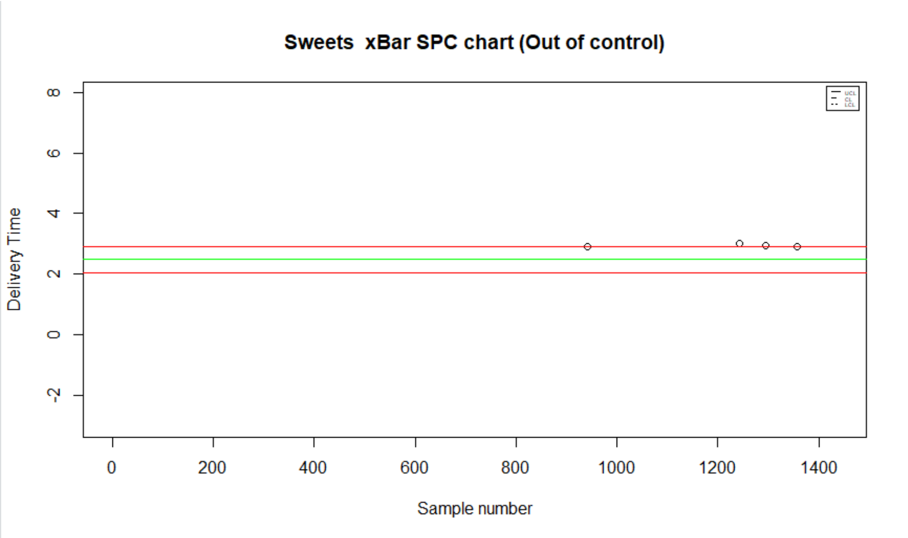


Figure 42: X-bar chart for out of control Sweets samples