

ECSA GA 4 REPORT

QUALITY ASSURANCE 344

Francis, WW, Mnr
21708126

Table of Contents

1. Introduction	1
2. Data Wrangling	2
3. Descriptive Statistics	2
3.1. <i>Analyzing the Valid Data Set</i>	2
3.2. <i>Analysis of Process Capability Indices</i>	6
4. Statistical Process Control	7
4.1. <i>Control Charts</i>	7
4.2. <i>Controlled Samples of Delivery Process Times</i>	10
5. Optimizing the delivery processes	13
5.1. <i>Sampled controlled data</i>	13
5.2. <i>Hypothesis testing: Type I</i>	14
5.3. <i>Costs involved</i>	14
5.4. <i>Hypothesis Testing: Type II</i>	15
6. MANOVA Results	15
7. Reliability of the service and products	17
7.1. <i>Taguchi Loss Function</i>	17
7.2. <i>Reliability Formulations</i>	18
7.3. <i>Binomial Distributions and Probability</i>	18
8. Conclusion	19
9. References	20
Appendix A: X-Charts	21
A.1. Food	21
A.2. Luxury	21
A.3. Household	22
A.4. Gifts	22
A.5. Technology	23
APPENDIX B: S – Chart	24
B.1. Food	24
B.2. Luxury	24
B.3. Household	25
B.4. Gifts	25
B.5. Technology	26
Appendix C: Controlled X&s Charts	27
C.1. Food	27

C.2. Sweets	28
APPENDIX D: Sample Controlled Data	29
D.1. Clothing	29
D.2. Food	29
D.3. Sweets	30
D.4. Luxury	30
APPENDIX E: Boxplots.....	31
E.1. Age	31
E.2. Price	31

1. Introduction

Standard descriptive statistics has two useful implications that benefits large businesses or online businesses when given the data to analyzed. It provides the basis information about the various variables or target features within a dataset and highlights the potential relationships between these features (variables).

Client data for an online business is supplied and is required to be analyzed. Upon evaluation, it is noted that in some cases, the data is not valid and as such has to be removed from the dataset to not compromise or invalidate data discussions before complete evaluation. In this report, the process to data wrangling is briefly explained and the purpose thereof is discussed. Here, descriptive statistics techniques will be applied and used to briefly evaluate the valid data set. Furthermore, the process capabilities indices will be briefly evaluated.

A statistical process control analysis will be conducted with accompanying X&s-charts. These charts will be initialized for every class specified based on their process delivery times where various statistical techniques will be computed, tabulated, and be discussed.

In addition, the delivery processes by the manufacturer will be optimized by the sample means method, the sample standard deviations method as well as by hypothesis testing methodologies. Conclusively, this report will further test and discuss concepts such the MANOVA vignette, reliability of the services and products, what conclusions can be drawn from the Taguchi Loss function as well as binomial distribution and probabilities in relation to the client data.

2. Data Wrangling

Data wrangling involves cleaning raw data to appropriate standards for optimization, in such a way that the information becomes easier to interpret. As introduced, the sales data contains incomplete instances that are not useful for analysis. These incomplete instances are then separated from the valid data, creating two separate sales records without changing the variables or target features. Within the valid and incomplete sales records, the original index was preserved. As such, another index was created to account for manipulating the data and adding additional features.

3. Descriptive Statistics

3.1. Analyzing the Valid Data Set

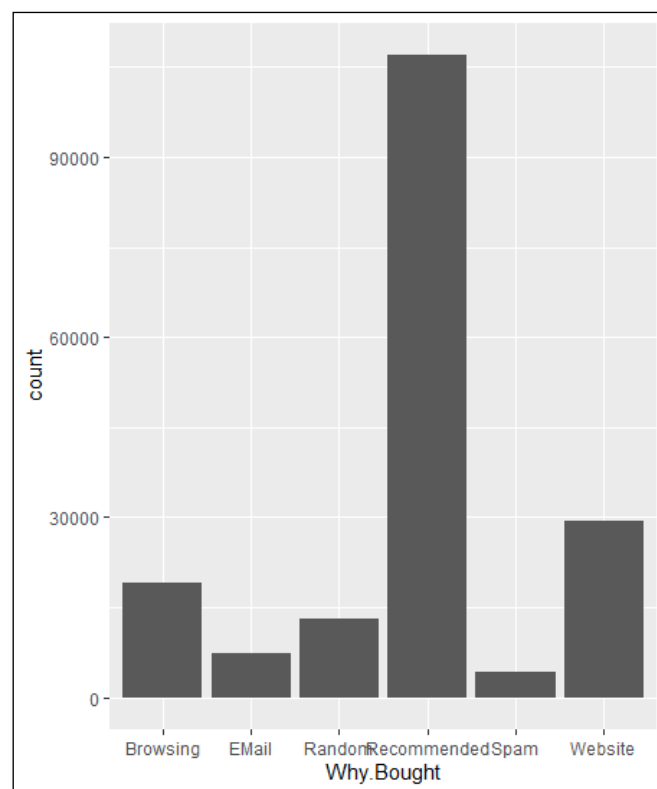


Figure 1: Reasons for buying per class.

Analyzing the sales data provided, we can see from Figure 1 above that recommendation accounts for majority of the purchases / tally counts across the years of operation. This could imply that the online business has made a fair success

and this successful reputation has spread amongst many people – recommending it to other people. Miscellaneous reasoning such as spams and emails has also provided some form of online business interaction. It should also be noted that “browsing and website” purchases provided a significant contribution as well, although not in comparison to the “recommended” option.

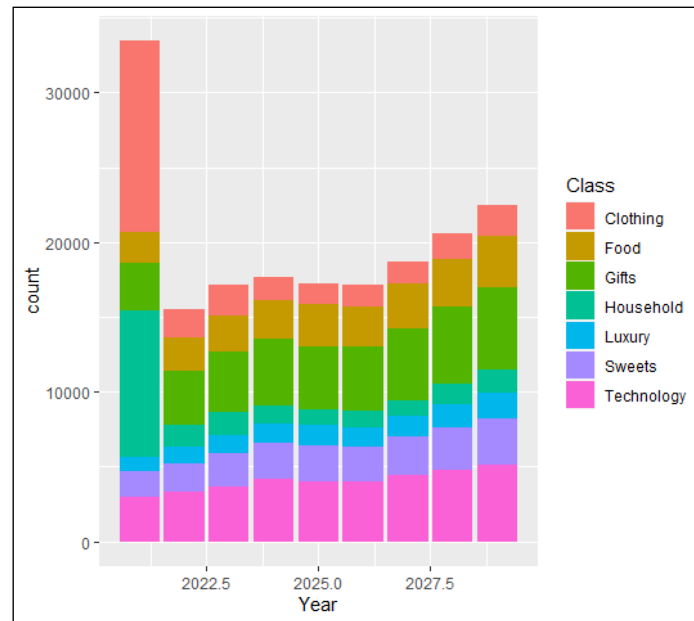


Figure 2: Graph of years of purchasing per class.

Household items and clothing from Figure 2 above, had the highest purchasing units online before 2022. This random variation in comparison to the other years in the data could largely be an implication of realistic external factors such as the COVID-pandemic, forcing users to order online. We do, however, notice a significant decrease in units purchased from the online business post the 2022 year. One significant importance we can also establish from Figure 2, is the general upward trend across the years over most of (if not all) the classes. This upward trend can be linked to the main reasons for buying (Recommended) per class, as the tally count in Figure 1 is averaged across the entire year domain of the data provided.

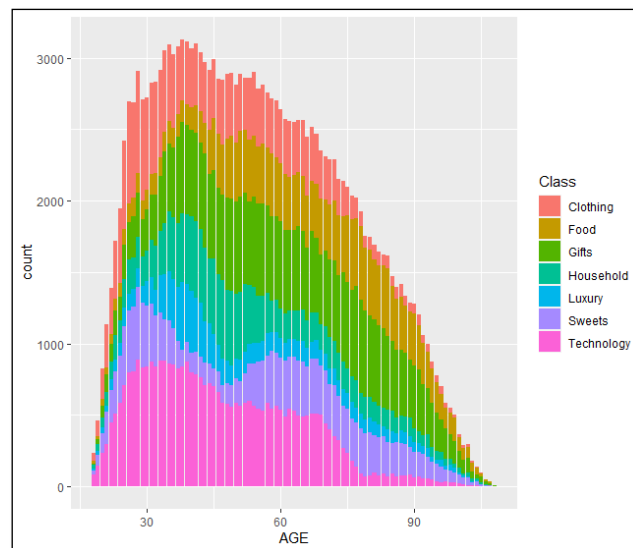


Figure 3: The age distribution per class.

A noticeable feature present in Figure 3 above, is the right skew distribution, moreover a unimodal distribution. The most dominant age group ranges from 18 to 55. Classes such as clothing as well as technology are the dominant class within this dominant age group. A significant decrease of units (count) can be seen on the tail end of the distribution, which is to be expected as the older you are the less likely are to make any general purchases. The statistical summary shown in Figure 4, clarifies the prior statements claimed.

key	X	ID
Min. : 1	Min. : 1	Min. :11126
1st Qu.: 44997	1st Qu.: 45003	1st Qu.:32700
Median : 89992	Median : 90004	Median :55081
Mean : 89992	Mean : 90002	Mean :55235
3rd Qu.:134988	3rd Qu.:135001	3rd Qu.:77637
Max. :179983	Max. :180000	Max. :99992
AGE	Class	Price
Min. : 18.00	Length:179983	Min. : -588.8
1st Qu.: 38.00	Class :character	1st Qu.: 482.3
Median : 53.00	Mode :character	Median : 2259.6
Mean : 54.57		Mean : 12293.7
3rd Qu.: 70.00		3rd Qu.: 15270.7
Max. :108.00		Max. :116619.0
Year	Month	Day
Min. :2021	Min. : 1.000	Min. : 1.00
1st Qu.:2022	1st Qu.: 4.000	1st Qu.: 8.00
Median :2025	Median : 7.000	Median :16.00
Mean :2025	Mean : 6.521	Mean :15.54
3rd Qu.:2027	3rd Qu.:10.000	3rd Qu.:23.00
Max. :2029	Max. :12.000	Max. :30.00
Delivery.time	why.Bought	
Min. : 0.5	Length:179983	
1st Qu.: 3.0	Class :character	
Median :10.0	Mode :character	
Mean :14.5		
3rd Qu.:18.5		
Max. :75.0		

Figure 4: The five number summary for the various classes

On average, it takes about 10 hours for the delivery to be complete. Realistically evaluating this on an economical scale, this fares well amongst other companies. From the figure above, a negative minimum price is shown as a possible result of damaged goods and a refund been placed for the client(s). The five number summary for the age group also justifies the statement *“The most dominant age group ranges from 18 to 55”* made earlier.

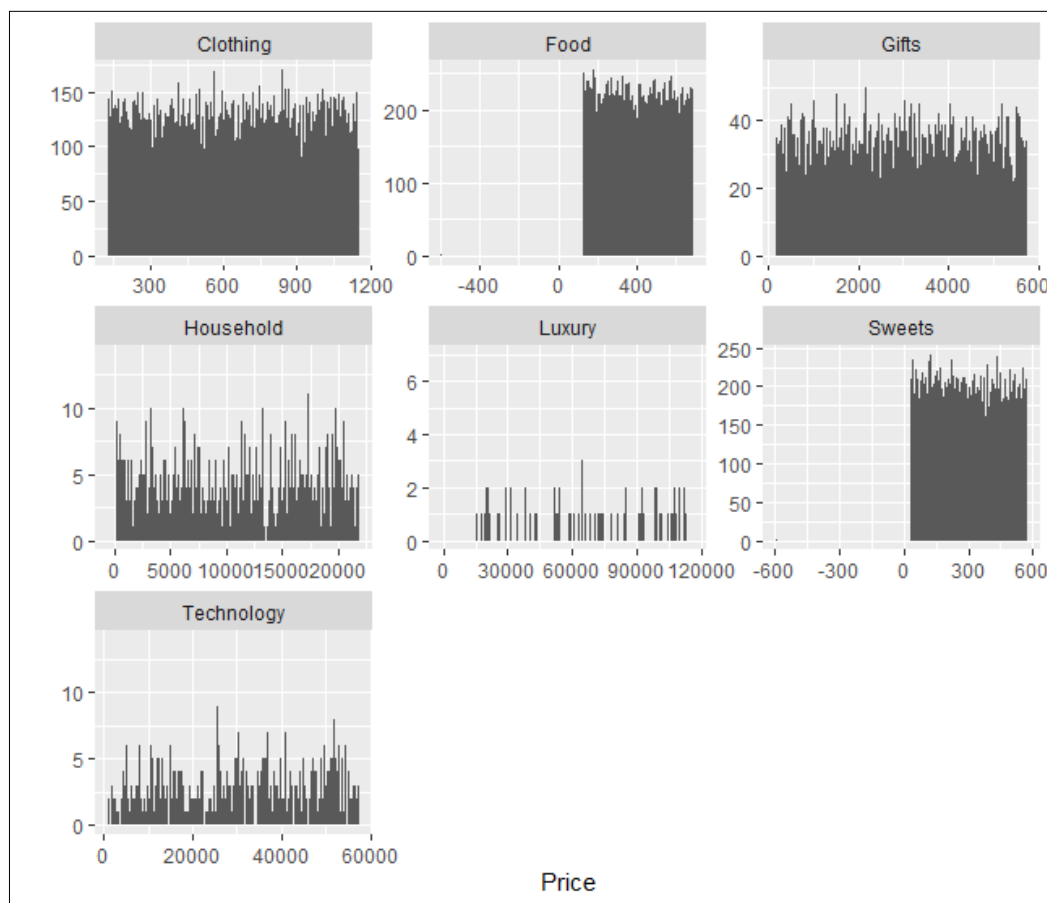
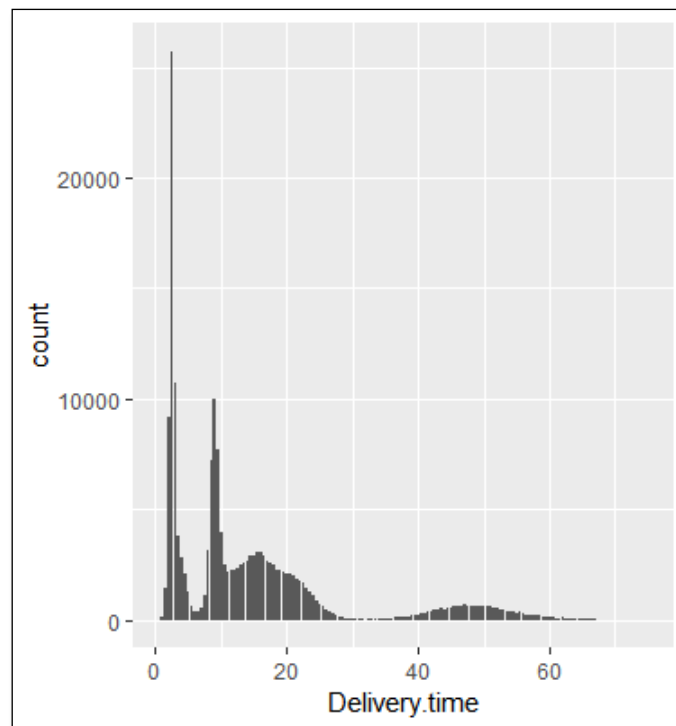


Figure 5: Distribution of different prices for different classes

The diagrams within Figure 5 above represent uniform histograms. Household, technology, and luxury classes significantly contribute to the cost accumulated, despite being showcased as less densely populated plots in relation to the other classes, with prices ranging to magnitudes of times 10 to the power of 5. Although a significant contribution, not many people choose to engage with luxury items.



*Figure 6: Delivery processing time (in hours)
across all classes*

The distribution of delivery processing times as seen in Figure 6 above, is classified as a bimodal distribution. The two peaks are a representation of the most frequent or rather, most delivery process times – the first ranging between 10 to 25 hours and the second between greater 40 hours which could imply that deliveries time accounted for potential long-distance shipment. Evaluating Figure 4, for delivery times, the mean is greater than the median which is represented in the positive right skewness of the graph in Figure 6.

3.2. Analysis of Process Capability Indices

We primarily assume the Upper and Lower Specification Limits to be 24 hours and 0 hours respectively. A lower specification limit equivalent to zero is logical as it is considered a natural boundary and showcases the lowest possible limit that a measurement will reach and be considered acceptable by the online business client. Using the R software, the process capability calculations are summarized in Table 1 below, rounded to 3 decimal places:

Table 1: Process capability indices

Statistical Property	Numeric Value
Upper Specification Limit	24 hours
Lower Specification Limit	0 hours
Mean	20.011 hours
Standard Deviation	3.502 hours
Capability Potential	2.284
Average Capability Potential	2.664
Capability Performance	1.905

The capability potential showcases how well the process spread fits into the specification range. The spread of the technology class figures fits into the limits 2.284 times. The capability performance is used to estimate how close to a given target. The performance for the spread of the process delivery times of the technology class figures is relatively close to the target feature and shows that this process is capable ($C_{pk} > 1.33$) to run without the instance of major defects occurring.

4. Statistical Process Control

Statistical Process Control (SPC) is defined as the use of statistical methods to control a process or production method. SPC tools and procedures help monitor process behavior, identify internal system problems, and find solutions to production problems.

4.1. Control Charts

The valid client data received was utilized to initialize both an X-Chart and an s-chart, which monitors the process variability when measuring the samples. Ordering the delivery process times by year, month, day and Row index, the first 30 samples of sample size 15, were used to determine the center lines, outer control limits and the 2- σ -1-sigma control charts and is tabulated in Figures 7 & 8 below.

	Class	UCL	U2Sigma	U1sigma	CL	L1sigma	L2sigma	LCL
1	Technology	22.9746158797126	22.1078920679566	21.2411682562005	20.3744444444444	19.5077206326884	18.6409968209323	17.7742730091763
2	Clothing	9.40493352386633	9.25995568257756	9.11497784128878	8.97	8.82502215871122	8.68004431742245	8.53506647613367
3	Households	50.2483278659662	49.0196259847182	47.7909241034702	46.5622222222222	45.3335203409742	44.1048184597263	42.8761165784783
4	Luxury	5.49396512637278	5.24116193610037	4.98835874582796	4.73555555555556	4.48275236528315	4.22994917501074	3.97714598473833
5	Food	2.70945773188154	2.63630515458769	2.56315257729385	2.49	2.41684742270615	2.34369484541231	2.27054226811846
6	Gifts	9.48856467334077	9.11274681926422	8.73692896518766	8.36111111111111	7.98529325703456	7.60947540295801	7.23365754888145
7	Sweets	2.89704150965879	2.75728693236512	2.61753235507145	2.47777777777778	2.33802320048411	2.19826862319044	2.05851404589677

Figure 7: X- Chart

The X-Chart showcased in Figure 7 above, monitors the means of successive samples at a constant sample size of 15.

	Class	UCL	U2Sigma	U1sigma	CL	L1sigma	L2sigma	LCL
1	Technology	5.18056970372824	4.55222240293678	3.92387510214531	3.29552780135385	2.66718050056238	2.03883319977091	1.41048589897945
2	Clothing	0.866559568463719	0.761455227250562	0.656350886037405	0.551246544824249	0.446142203611092	0.341037862397935	0.235933521184778
3	Households	7.34418006586244	6.45341013420991	5.56264020255739	4.67187027090486	3.78110033925233	2.89033040759981	1.99956047594728
4	Luxury	3.23598408014972	2.84349406873207	2.45100405731442	2.05851404589677	1.66602403447912	1.27353402306147	0.881044011643818
5	Food	0.437246583672721	0.384213283023697	0.331179982374673	0.278146681725649	0.225113381076626	0.172080080427602	0.119046779778578
6	Gifts	2.24633333311156	1.9738772969496	1.70142126078763	1.42896522462567	1.15650918846371	0.884053152301749	0.611597116139788
7	Sweets	0.835339146409308	0.734021506089943	0.632703865770579	0.531386225451214	0.430068585131849	0.328750944812484	0.22743330449312

Figure 8: S-Chart

The s-chart shown in Figure 8, monitors the process variability when measuring the samples.

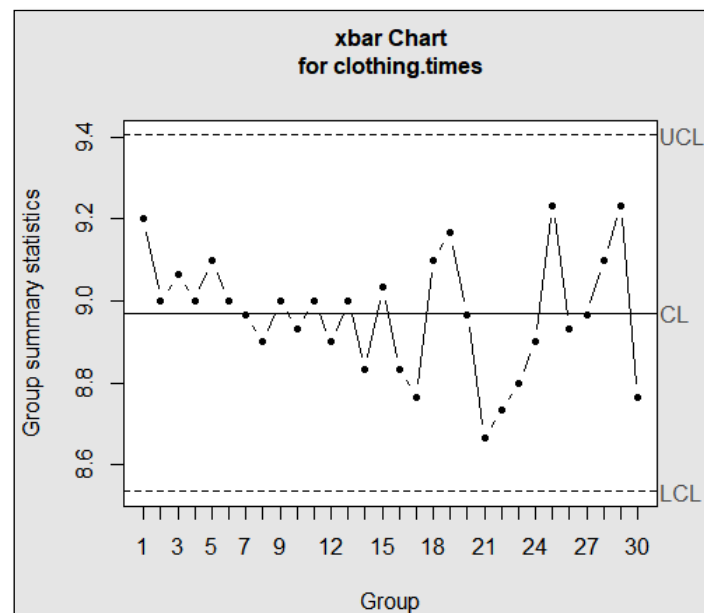


Figure 9: X-Chart for Clothing

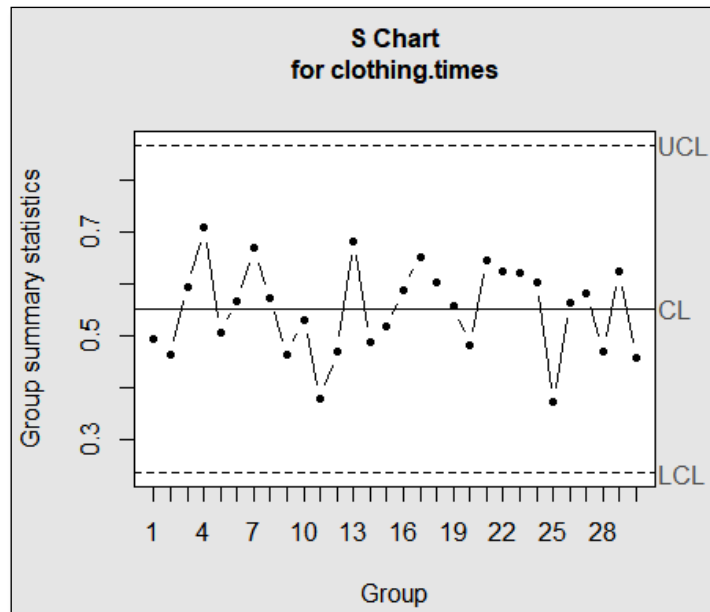


Figure 10: S-Chart for Clothing

The various x-bar charts and s-charts represented in the Figures above as well as those in Appendix A and B respectively, remain within the confines of the Upper Center Line and Lower Center Line limits. Sweets, however, upon initial examination had a sample group breach the upper limit of the S Chart. As such, the violated data had to be removed, as seen in Figure 12 below. This now presents us with controls chart all within their set boundaries.

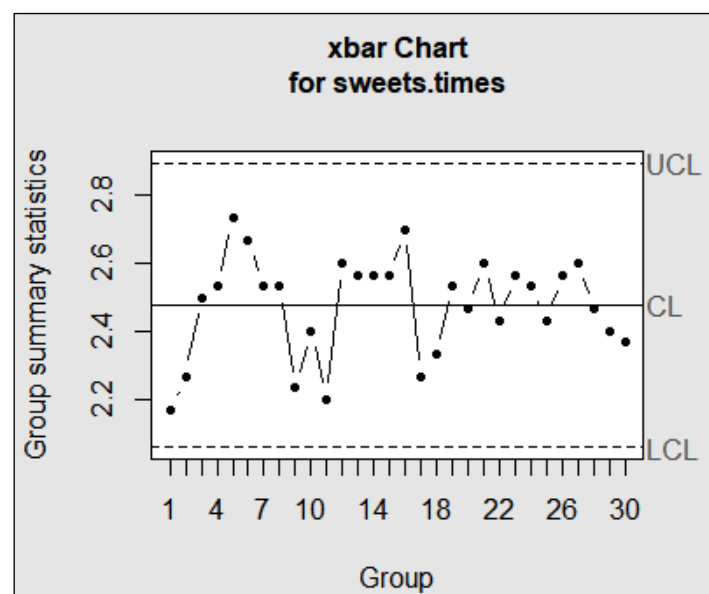


Figure 11: X-Chart for Sweets

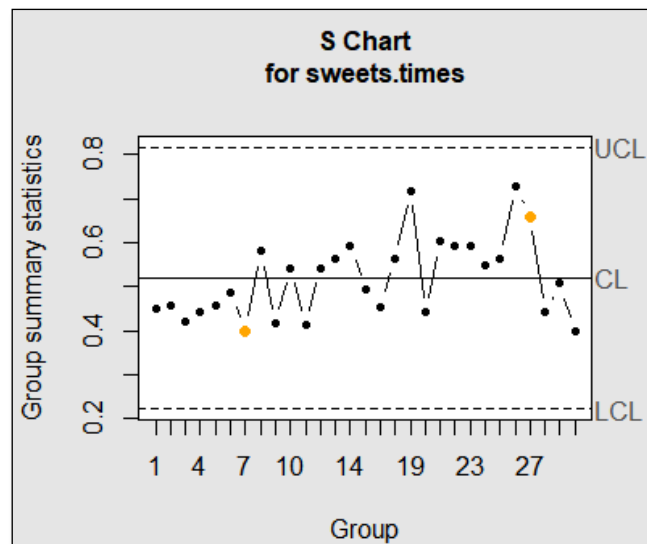


Figure 12: S-Chart for Sweets

4.2. Controlled Samples of Delivery Process Times

After completing the control chart initialization and understanding the process, 15 delivery time samples were drawn, from sample 31 to the last instance. These sample charts were created using the control limits calculated when the charts were initialized on the first 30 samples in sub-section 4.1.

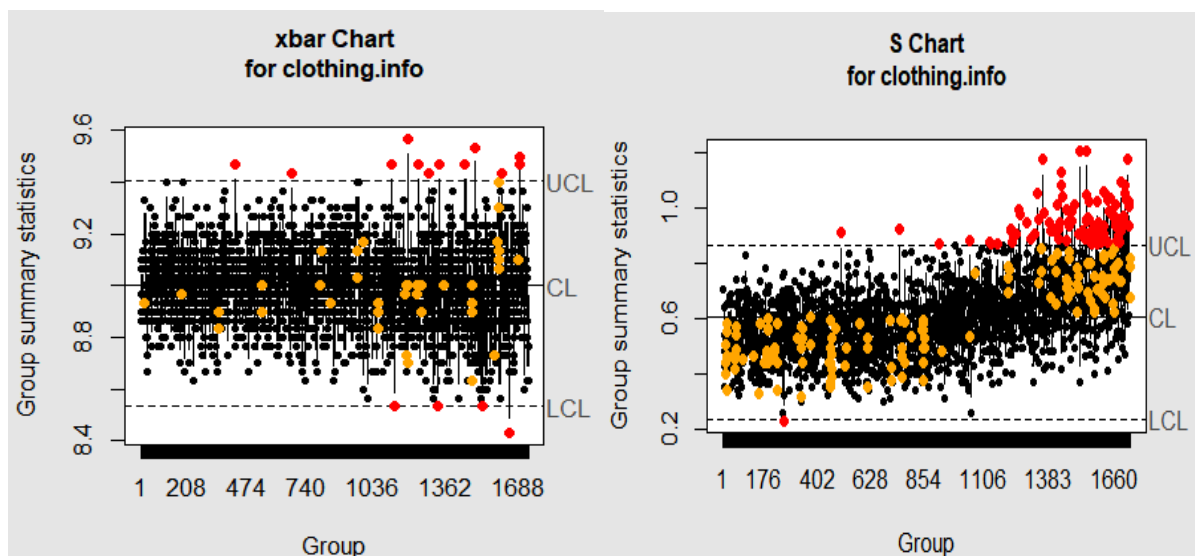


Figure 13: Controlled X&s Chart for Clothing

Figure 13, much like the charts to follow, is a densely populated graph. Upon observation, samples greater than group 1000 have data that violates the UCL

threshold. Action must be taken to find the special causes and permanently remove it from the process.

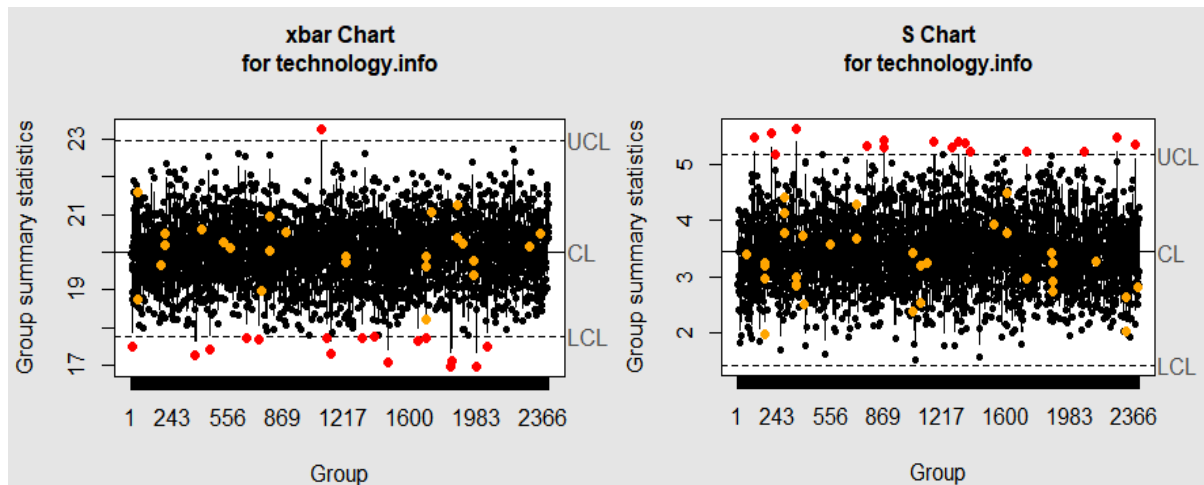


Figure 14: Controlled X&s Chart for Technology

The technology samples represent that of a process that is almost completely stable. A few data points exceed the control limit threshold, which needs to be removed, however most points in both charts are densely populated towards the center line. The Food and Sweets classes, follow a similar trend and observation, as supported by the figures in Appendix C.

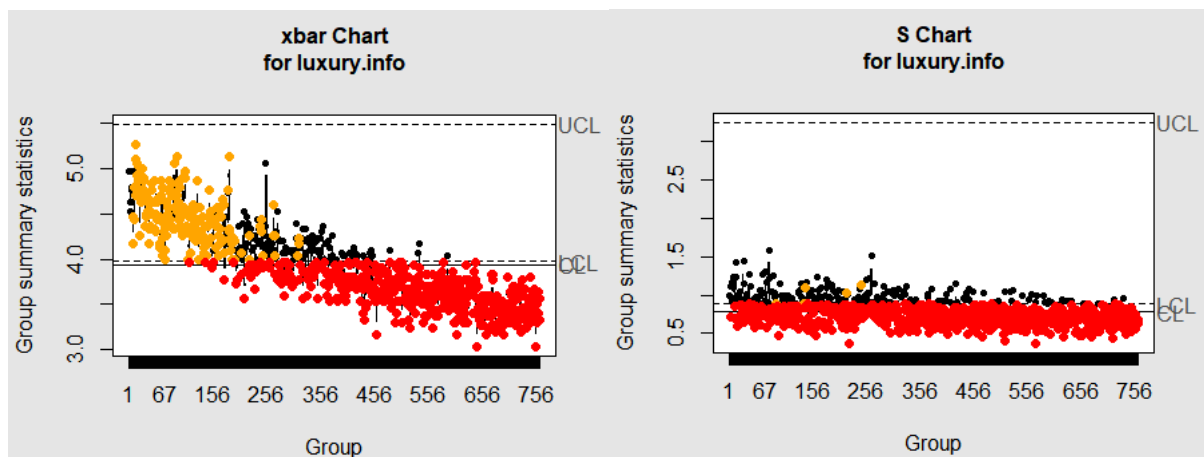


Figure 15: Controlled X&s Charts for Luxury

The luxury charts are a prime example of out-of-control situations. The vast special causes present need to be removed. Another noticeable feature in Figure 15 can be observed by the positioning of the CL and LCL limit – which is responsible for processing most points in violating the control limits. The “Beyond Limits” rule further justifies this statement, in addition the downward trend seen in the X-Chart.

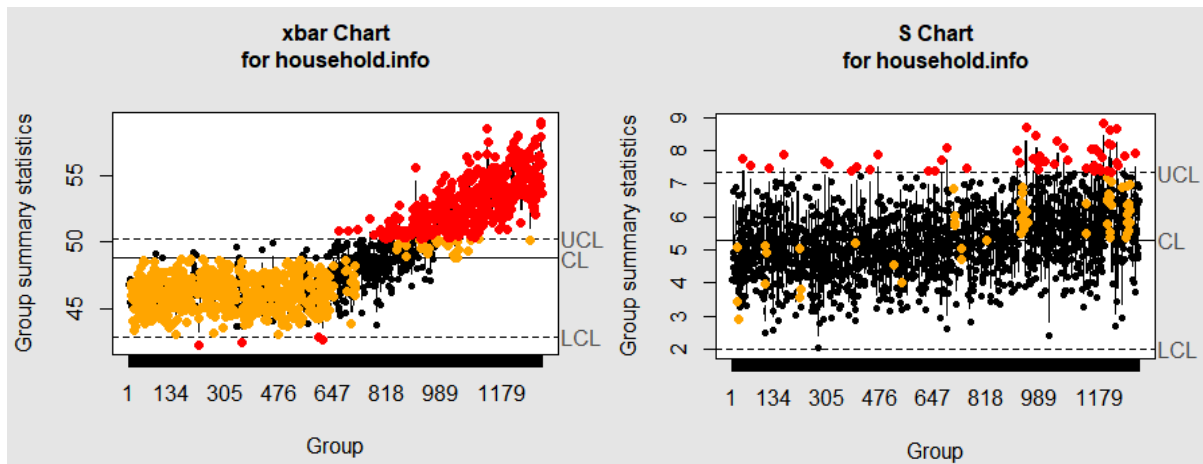


Figure 16: Controlled X&s Charts for Household

The Household class has a substantial increase in data point that exceed the outer control limits. This statement is supported by both charts in Figure 16, moreover the positive increase in red points for groups larger than 800. Thus, in such situations the special causes need to be resolved promptly.

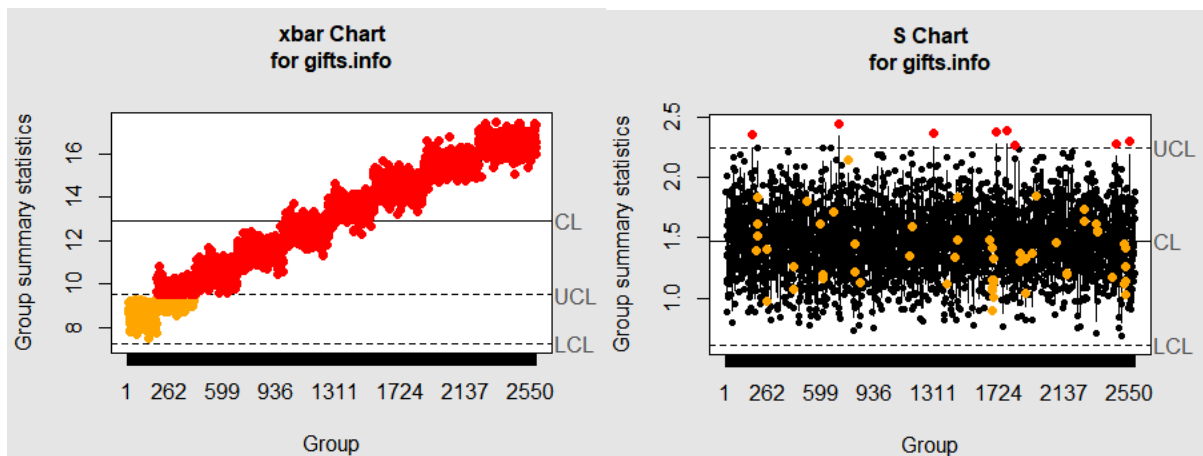


Figure 17: Controlled X&s Chart

Observing the S-chart, a large majority of the points are centered around the center line with a few outliers. However, it is not possible to assume that the process is stable because the X-chart has a larger portion of points in the danger (red) zone significantly increases and exceeding the UCL. Similarly, to Figure 15, the CL exceeds the UCL, which indicates that for the Gifts Class, the process is out-of-control.

5. Optimizing the delivery processes

For this section of the report, the concept of Statistical Process Control was re-evaluated, showcasing which data classes were out of control processes. An evaluation on the delivery process mean delivery time will be discussed and a conclusion will be provided on how it relates the Taguchi Loss.

5.1. Sampled controlled data

To optimize the delivery process, an analysis on the respective classes was conducted to investigate the quantity of points that lie above the UCL and quantity of points that lie below the LCL, represented in the figures below by green dots (above) and red dots (below) respectively.

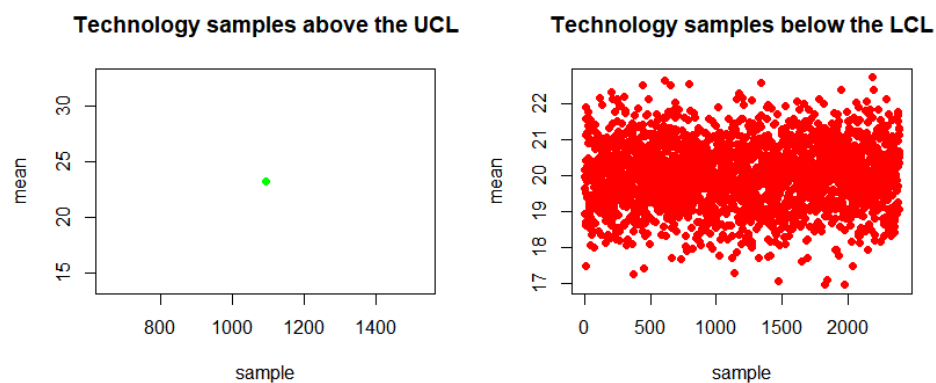


Figure 18: Technology samples

A vast majority of samples have the tendency below the LCL. This observation is a common occurrence found amongst classes, as seen in the Appendix D. Although the data has been controlled by sample means limits or sample standard deviation limits, the processes are found to be out-of-control processes. If there is an uncontrolled condition present, more research should be done. It is crucial to identify the causes that can be removed and to take action to do so. The process will become stable and return to an under-control situation once assignable causes are eliminated, leaving only random fluctuation resulting from common causes.

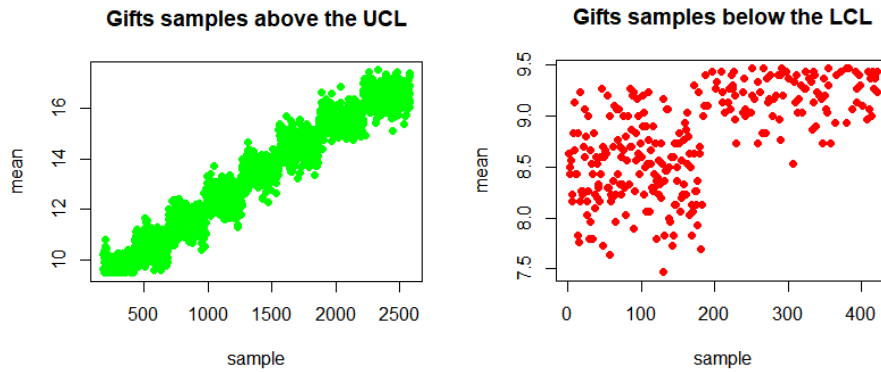


Figure 19: Gifts samples

The “Household” and “Gift” classes proved to be the most unstable as both the figures for samples above the UCL and samples below the LCL are densely populated figures.

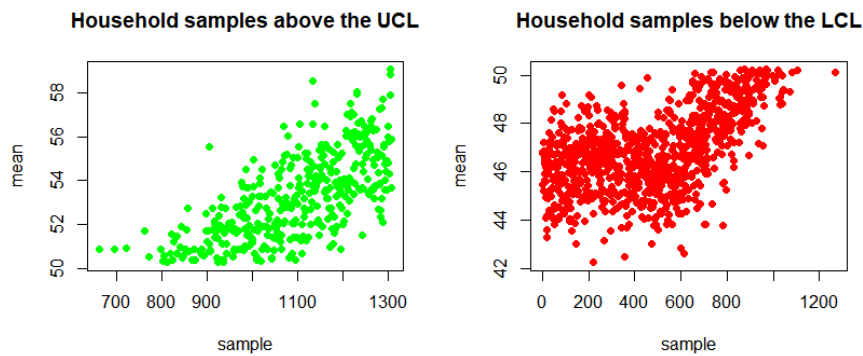


Figure 20: Household samples

5.2. Hypothesis testing: Type I

The likelihood of making a manufacturers error is described by the null and hypothesis testing metrics. The null hypothesis states that the process is in control and is centered on the centerline calculated shown in the table below. The alternative hypothesis states that the process is not in control and has moves from the centerline, in addition to either increasing or decreasing its variation. The probability that the null hypothesis test will hold true although incorrectly rejected is equivalent to 0.2699%.

5.3. Costs involved

The current additional costs for individual delivery times of the technology class are priced at R150 752 900. To obtain the best profit to center the delivery process, the

delivery time should be reduced by 34 hours for every item to achieve a total of 19.49952 hours for the technology class on the account that the lowest cost for said class remains at R16 104 770.

5.4. Hypothesis Testing: Type II

The likelihood of making a consumer's error is also described by the null and alternative hypothesis testing metrics. For type II errors, the alternative hypothesis, as stated in sub-section 5.2, is true however identification due to the delivery process average changing to 23 was not made known failed, assuming the null hypothesis to be true instead. The probability that the alternative hypothesis test was true although incorrectly rejected, is equivalent to 199.73 %.

6. MANOVA Results

Finding out if various levels of independent variables have an impact on the dependent variables individually or in combination with one another is the main goal of multivariate analysis of variance (MANOVA). The null hypothesis for the MANOVA states that the Age, Price, Year and Delivery times having no significant difference to the reasons for purchasing. As such, the alternative hypothesis states that for at least one of the factors aforementioned, there is a difference to the reasons for purchasing. The MANOVA computed is summarized below in Figure 21.

Descriptive:					
why.Bought	n	AGE	Year	Price	
1 Browsing	18994	53.849	2025.142	16130.561	
2 Email	7225	55.755	2024.716	6661.072	
3 Random	13121	56.963	2024.656	4288.261	
4 Recommended	106988	54.481	2024.885	13440.539	
5 Spam	4208	54.659	2024.841	9360.900	
6 Website	29447	53.965	2024.684	11020.505	
Delivery.time					
1	14.739				
2	14.422				
3	14.179				
4	13.226				
5	15.235				
6	19.033				
Wald-Type Statistic (WTS):					
	Test statistic	df	p-value		
why.Bought	"12957.913"	"20"	"<0.001"		
modified ANOVA-Type Statistic (MATS):					
	Test statistic				
why.Bought	12419.45				
p-values resampling:					
	paramBS (WTS)	paramBS (MATS)			
why.Bought	"<0.001"	"<0.001"			

Figure 21: MANOVA Summary

The significance value choose for the data is 0.05. Comparing this to the probability value obtained, that being less than 0.001, this implies that the data has rejected the null hypothesis specified and that there is in fact one factor that produces a difference to the reasons for purchasing.

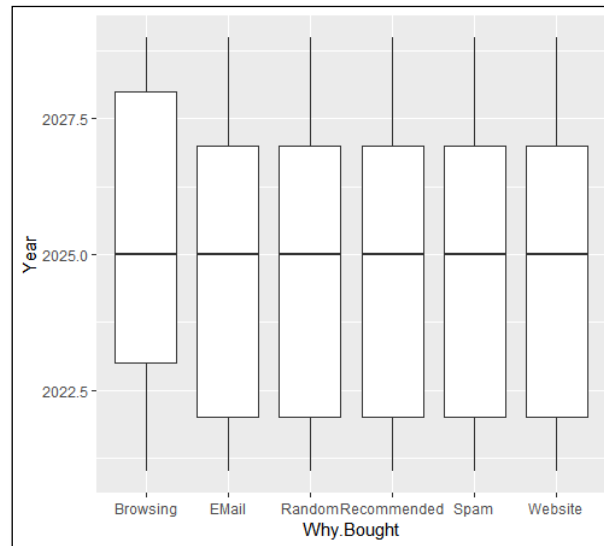


Figure 22: Boxplot of Year vs Why.Bought

The Year Class, has a central limit that is relatively similar across all the years observed. Much like the other classes, as seen in Appendix E, the Year class has little to no variation between its means shown in Figure 22 above. The Price Class however, has a major variation between its means for the different reasons for purchasing. This justifies the rejection of the null hypothesis as it satisfies the alternative hypothesis.

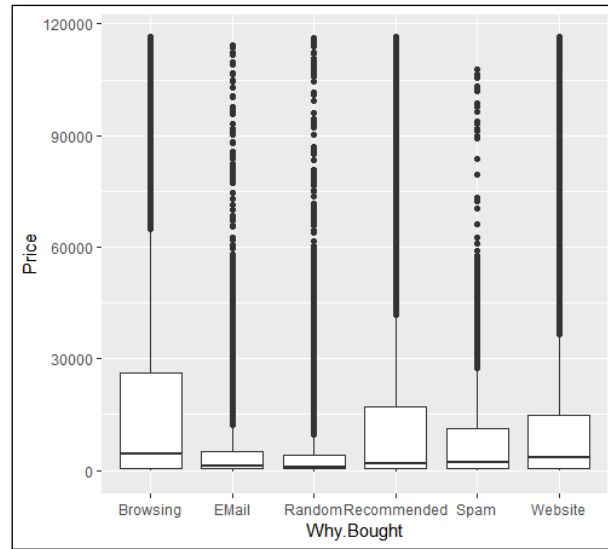


Figure 23: Boxplot of Prices vs Why.Bought

7. Reliability of the service and products

To evaluate the reliability of services and products, statistical properties such as the Taguchi loss Function and Binomial Probability can be used to draw useful conclusions for management reports.

7.1. **Taguchi Loss Function**

Taguchi defines loss as a product that is not accepted irrespective of the specification having been met or not. The Taguchi Loss Function has tremendously fueled the continuous improvement movement in reducing the variation of a product. It is defined by the following equation:

$$L = k(y - m)^2 \quad (1)$$

Food deliveries are required to be kept cool during transit. Cool Food, along with its subsidiary Lafrideradora, are involved in ensuring that this condition is met. In this situation, with a blueprint specification of 0.06 ± 0.04 cm in thickness for a refrigerator part that costs \$45 to scrap, the Taguchi Loss Function was determined to be represented by the equation 2 below:

$$L = 28125(y - 0.06)^2 \quad (2)$$

However, the team at Cool Food were able to reduce the scrap cost \$35. As such the process deviated from the target by 0.027 cm, thus resulting in a Taguchi Loss Function (represented in the equation 3 below) and a loss of $L = 15.95$.

$$L = 21875(y - 0.06)^2 \quad (3)$$

7.2. Reliability Formulations

A reliability block diagram may be used to demonstrate the interconnection between individual components. Magnaplex was appointed as another subsidiary to solve its waste issue for some of the technology items that are unnecessarily kept as safety stock due to operations and production failure. Three operations (two identical machines per operation) are required for evaluation with reliability issued for operation A, operation B and operation C as 0.85, 0.92 and 0.90 respectively.

For one machine working in operations A, B, C respectively, the total reliability results in 0.7038, however, having two machines at each operation substantially increased the reliability performance by 26% to 0.9615, where the higher reliability is said to be the preferred level for this situation.

7.3. Binomial Distributions and Probability

For the delivery process, it was proclaimed that 19 out of the 20 vehicles available for operations at any time yields in a reliable service. In addition, 21 delivery drivers who each work 8 hour shifts per day are also available.

The number of days per year we should expect reliable delivery times (with additional features), is tabulated in Table 2 below

Table 2: Expected Reliability of delivery times

Statistical Probabilities	Numeric Value
Vehicle - Reliability (20 vehicles)	19.851
Driver - Reliability (20 vehicles)	20.932
Vehicle - Reliability (22 vehicles)	21.837
Driver - Reliability (22 vehicles)	20.932

Since, under the assumption that, the probability of failure per vehicle remains constant, the reliability of drivers for either 20 or 22 vehicles remain the same. As such, the estimated number of days for a reliable delivery time is 361.1221 days. That equates to roughly 98.94% of the year.

8. Conclusion

Analysis of the client data has shown that recommendations are the most significant reason customers buy from the business, with the Technology class contributing the most purchases overall. Management should continue to adhere to the standard set out for the Technology Class and find ways to bring other class to that standard as well.

Miscellaneous marketing techniques should be removed in order to decrease the amount of unnecessary expenditure and increase the overall profit for a more successful business. In addition, management should implement techniques that will reduce the amount of out-of-control process. Stability has promising results for the profit and that should always be the target.

Conclusively, observing the Taguchi Loss Function, lower scrap costs and incorporating two machines at each operation has the most feasible outcome. Adjusting the probability of failure, will spur major benefits for management.

9. References

Dauber, D. 2022. *R for Non-Programmers: A guide for social scientists*. [Online]. Available: https://bookdown.org/daniel_dauber_io/r4np_book/data-wrangling.html [13 October 2022]

Grolemund, G. 2016. *Data Wrangling with R and RStudio*. [Online]. Available: <https://www.rstudio.com/resources/webinars/data-wrangling-with-r-and-rstudio/> [12 October 2022]

Hessing, T. 2019. *What are Xbars and S Charts*. [online]. Available: <https://sixsigmastudyguide.com/x-bar-s-chart/> [16 October 2022]

ISIXSIGMA. 2022. *Control charts*.

Jardine, J. 2022. *Trend Analysis on Quality Management: Improving Decision-Making With Timely Data*. [Online]. Available: <https://www.mastercontrol.com/gxp-lifeline/trend-analysis-quality-management/> [17 October 2022]

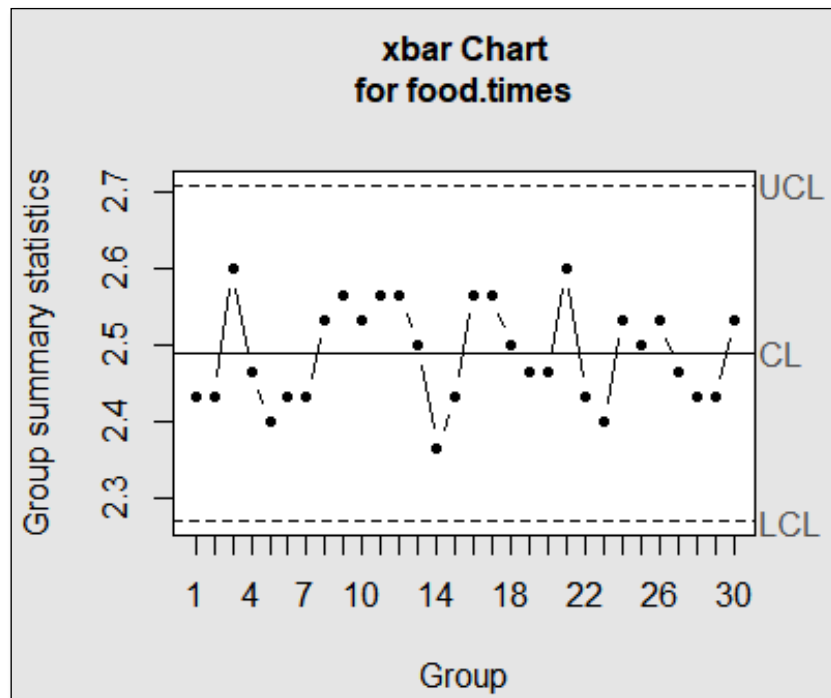
SPCExcel.2004. *Interpreting Control Charts*. [Online]. Available: <https://www.spcforexcel.com/knowledge/control-charts-basics/interpreting-control-charts#points-beyond-the-control-limits> [16 October 2022]

Total Quality Management. 2017. Lower Control Limit. [Online]. Available: <https://www.sciencedirect.com/topics/engineering/lower-control-limit#:~:text=LCL%2C%20lower%20control%20limit%3B%20UCL%2C%20upper%20control%20limit,corrected%2C%20such%20as%20a%20new%20operator%20in%20training.> [18 October 2022]

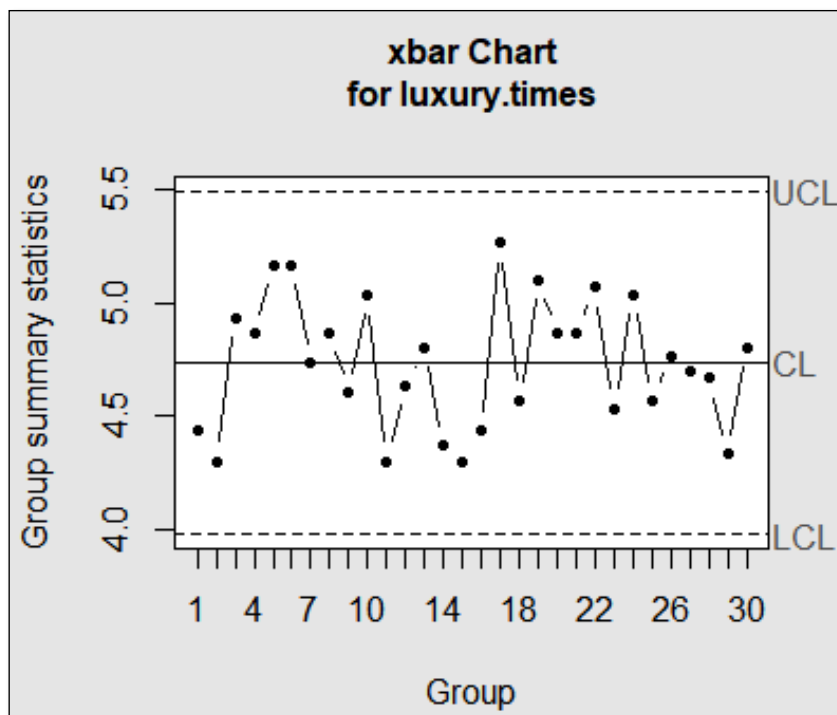
Wach, S. 2016. What does an out-of-control process indicate. [Online]. Available: <https://www.winspc.com/what-does-an-out-of-control-process-indicate/> [20 October 2022]

Appendix A: X-Charts

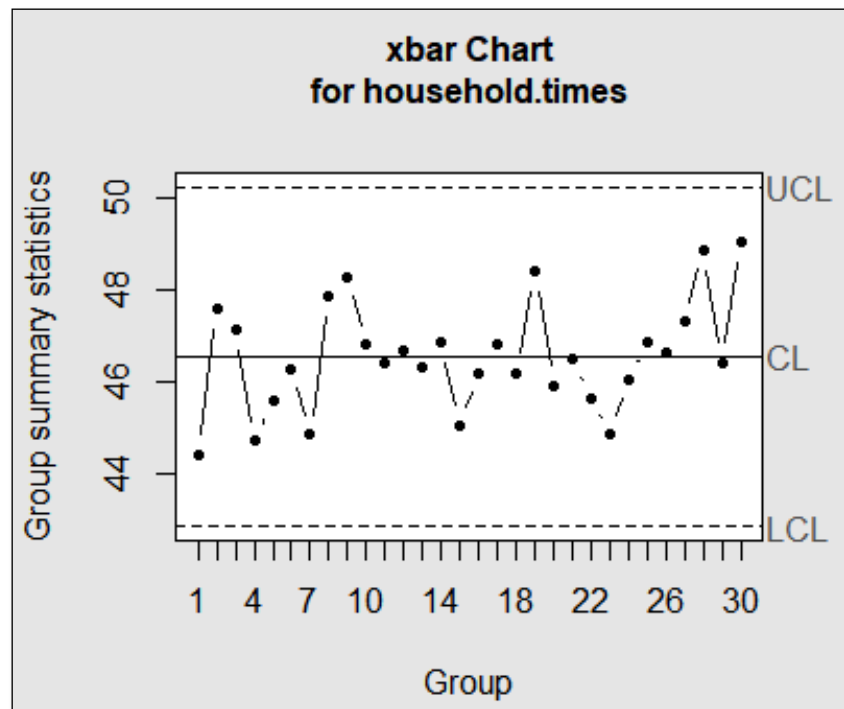
A.1. Food



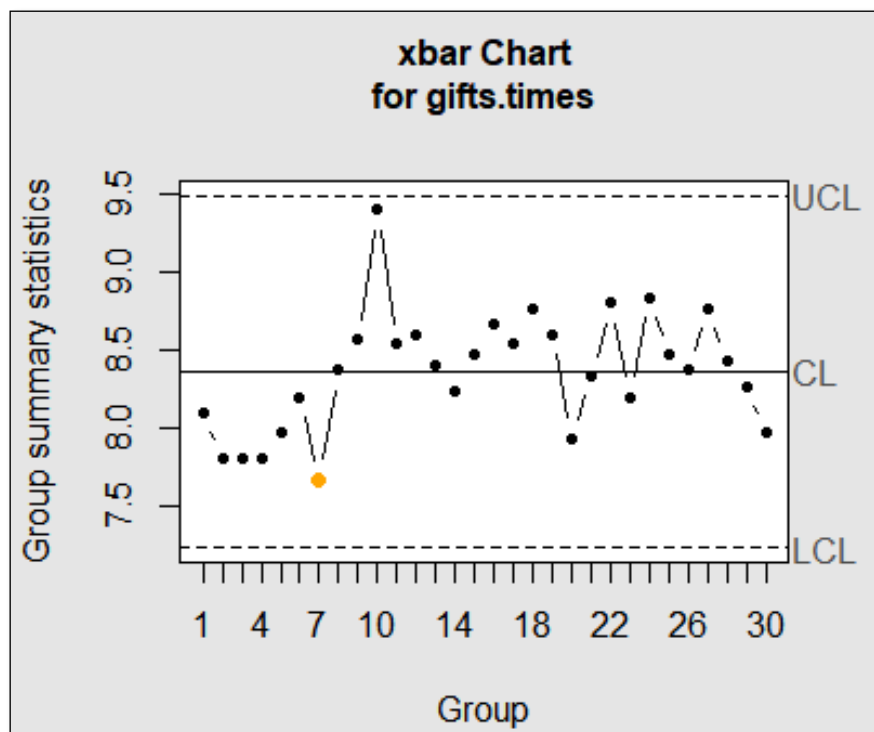
A.2. Luxury



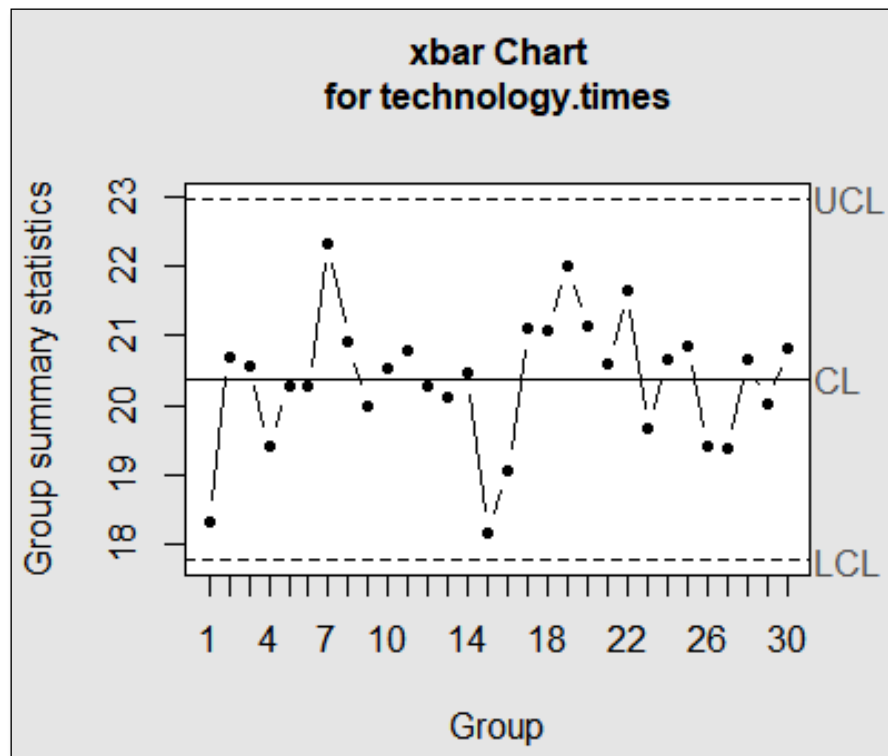
A.3. Household



A.4. Gifts

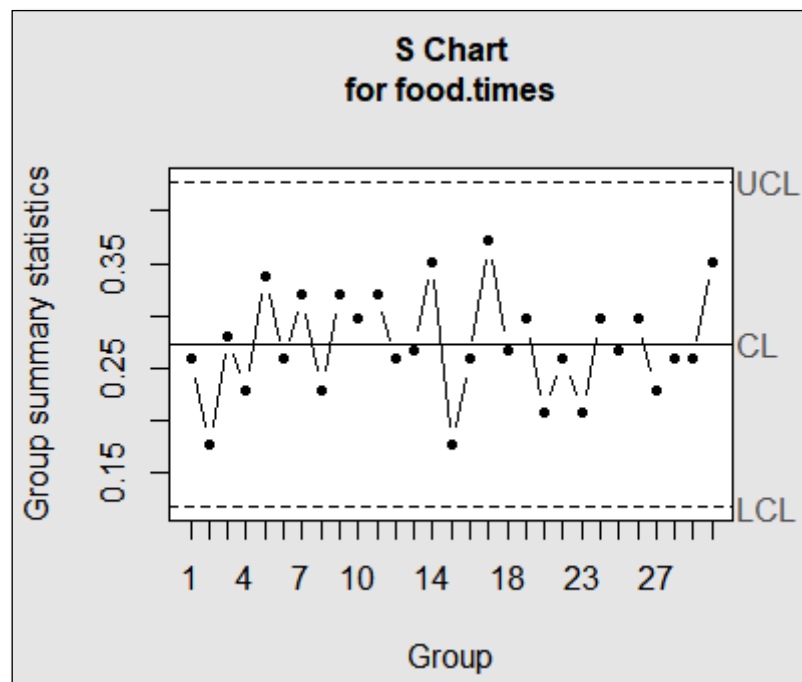


A.5. Technology

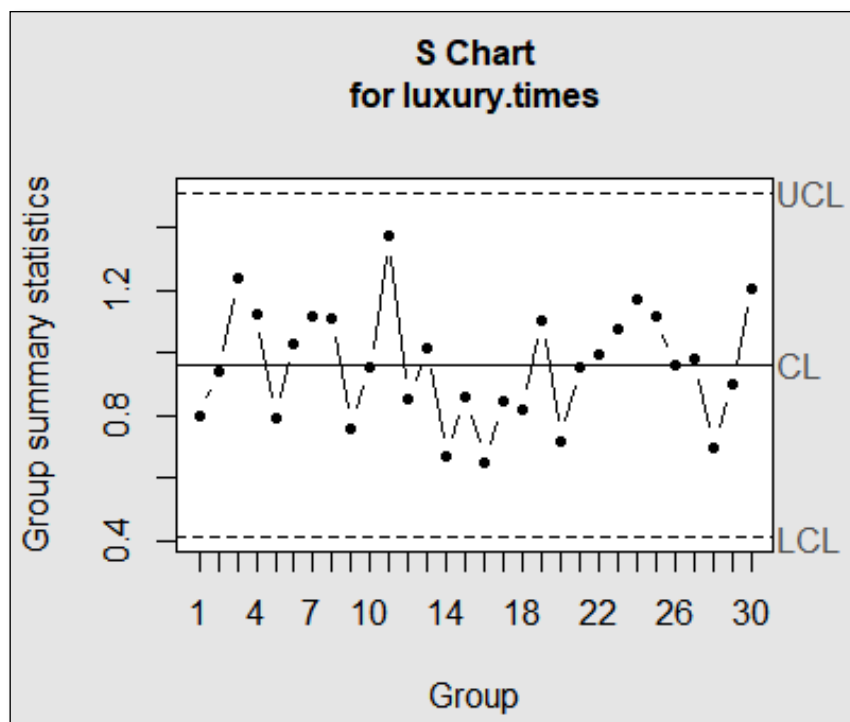


APPENDIX B: S – Chart

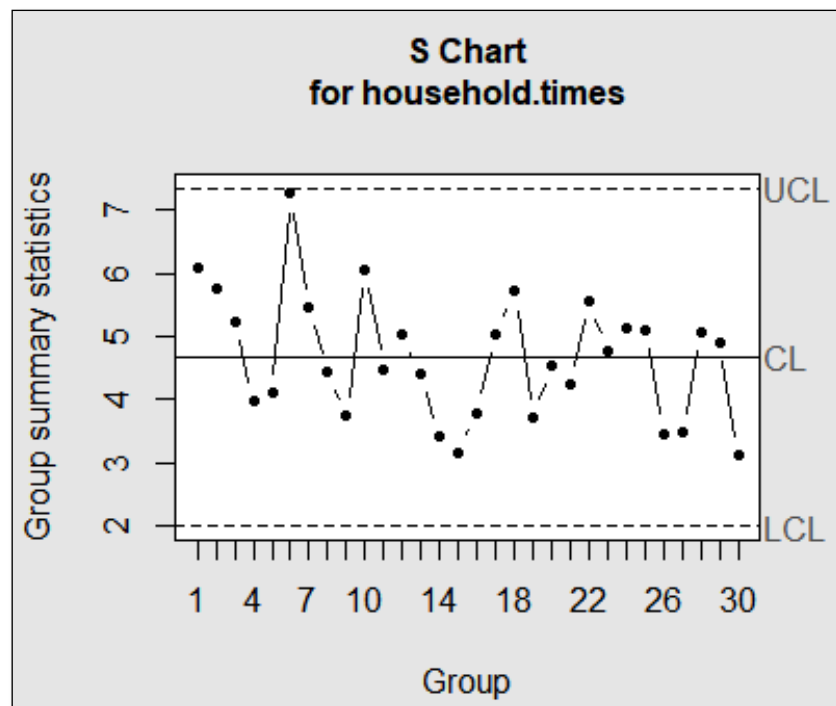
B.1. Food



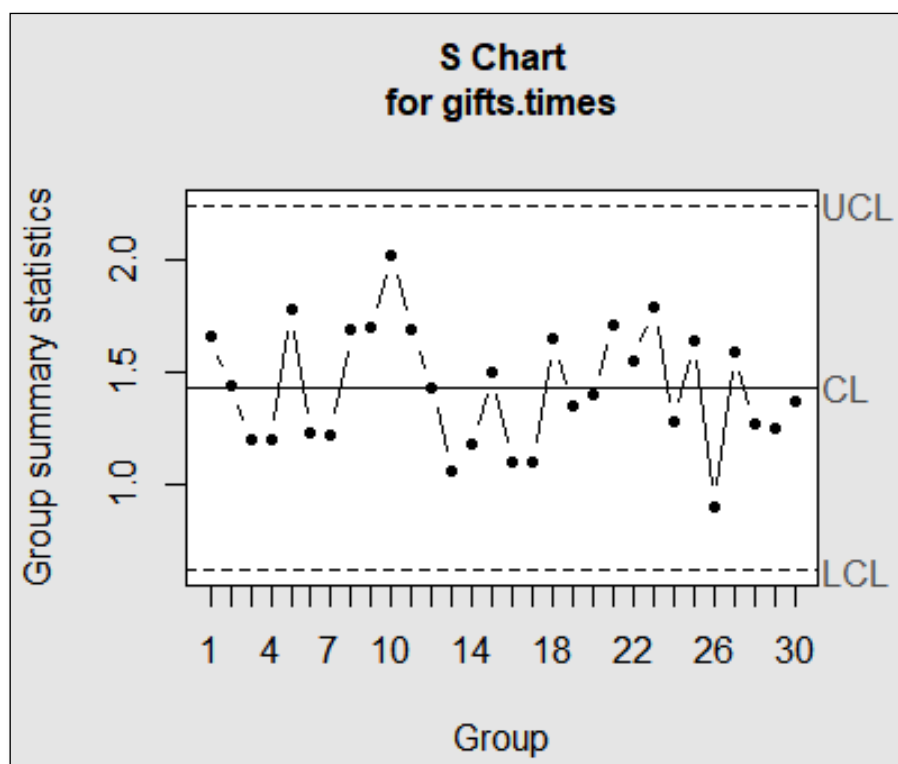
B.2. Luxury



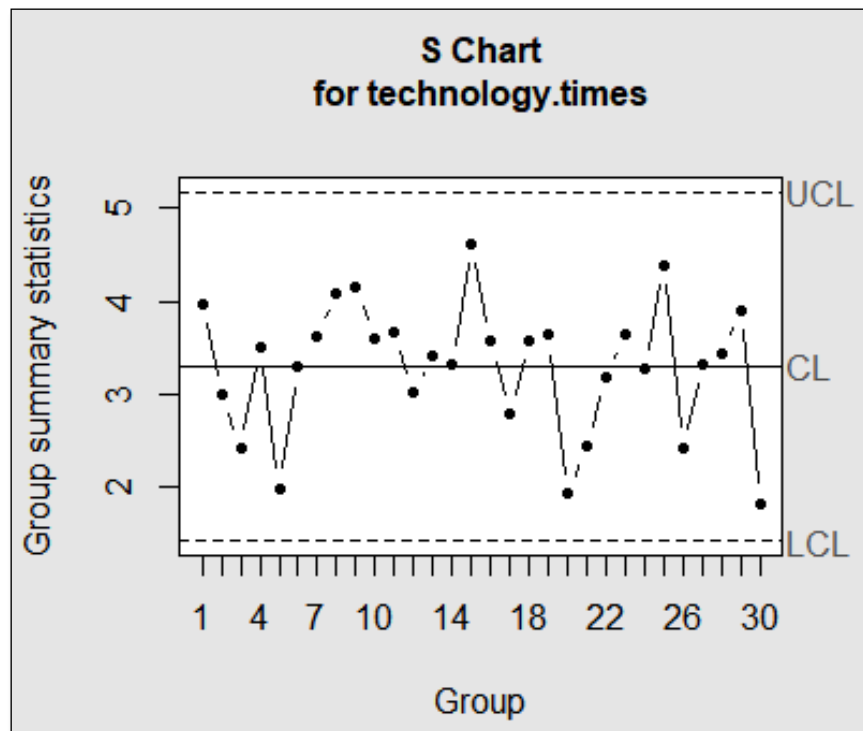
B.3. Household



B.4. Gifts

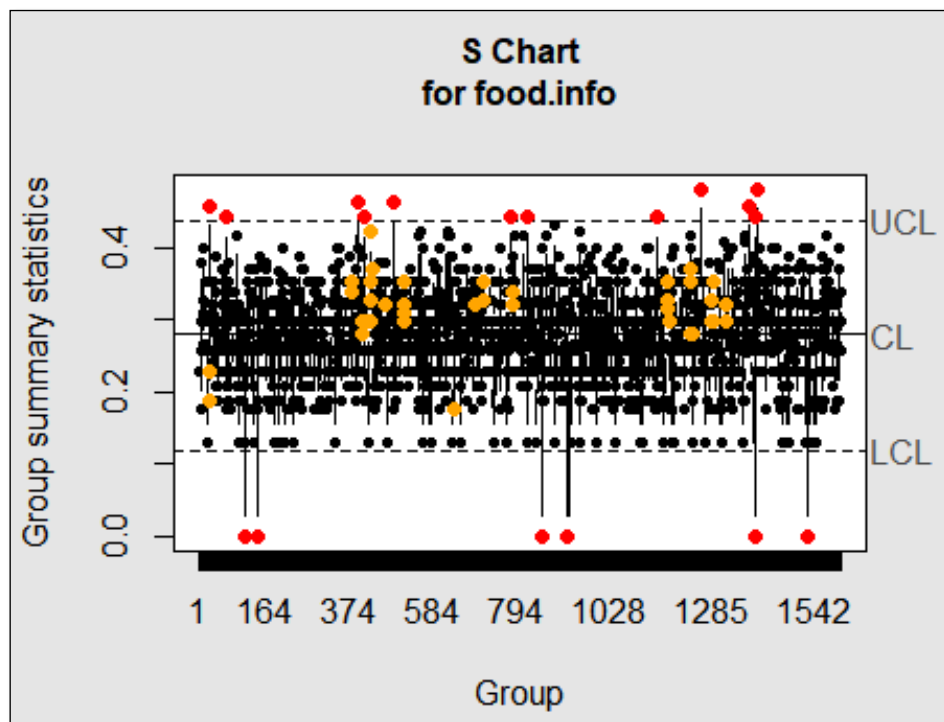
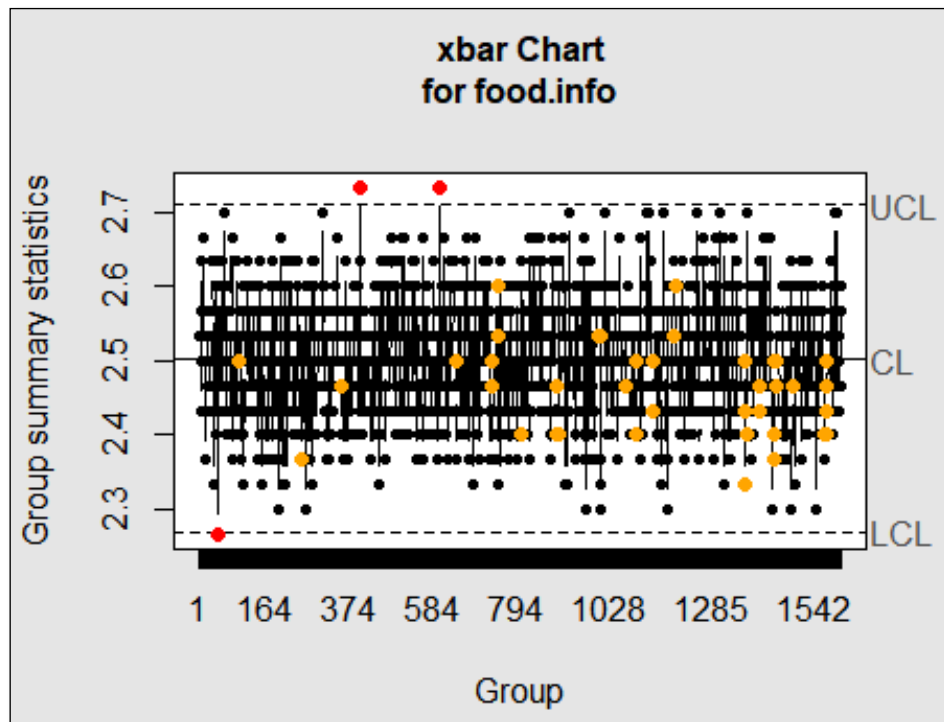


B.5. Technology

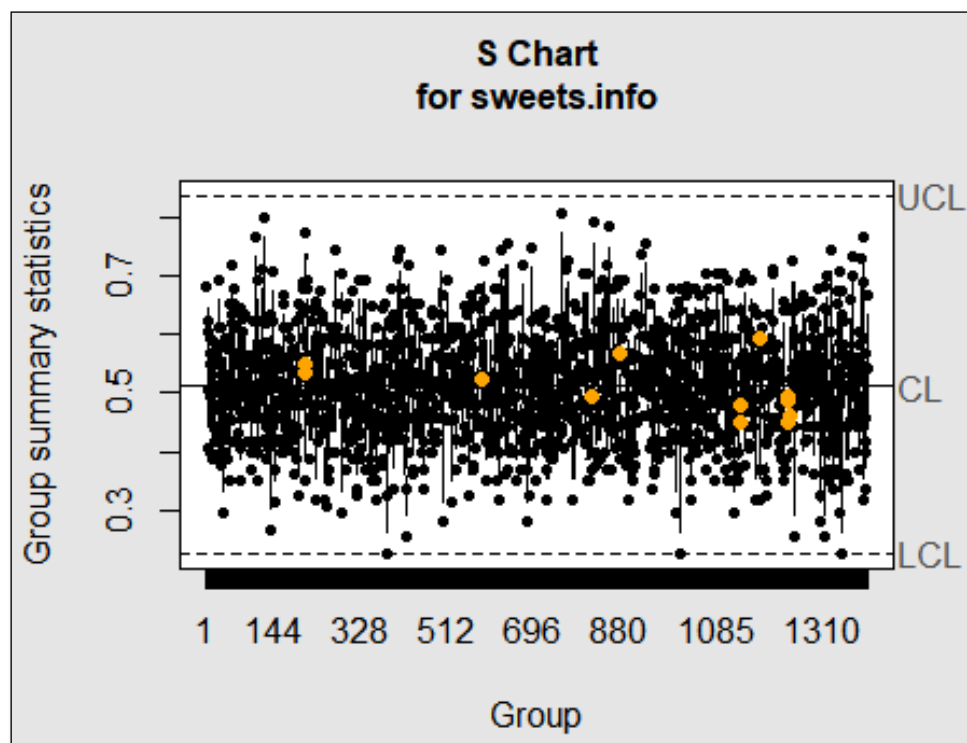
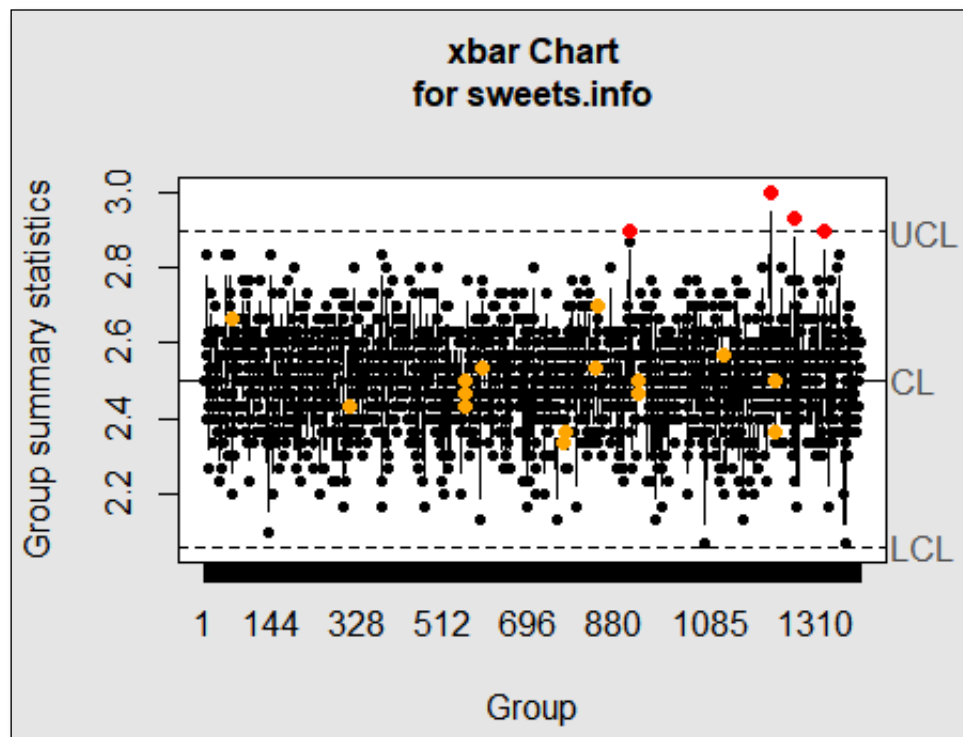


Appendix C: Controlled X&s Charts

C.1. Food

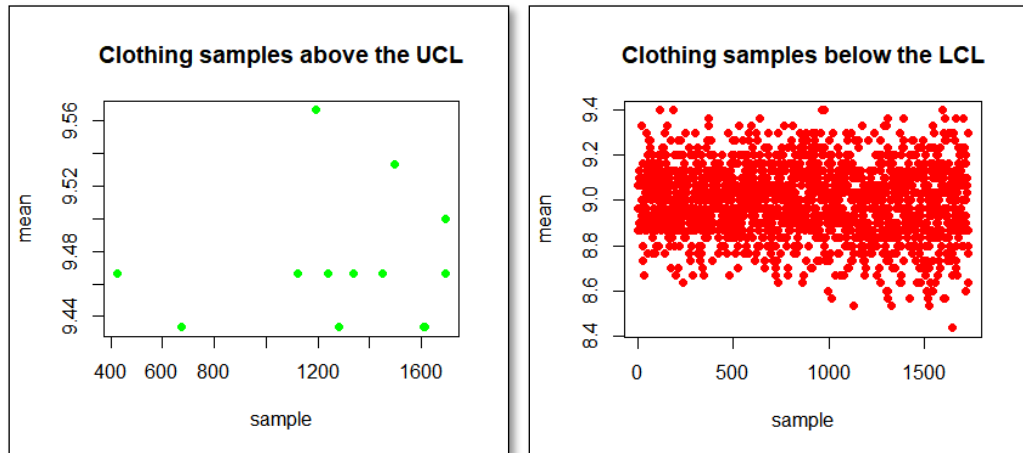


C.2. Sweets

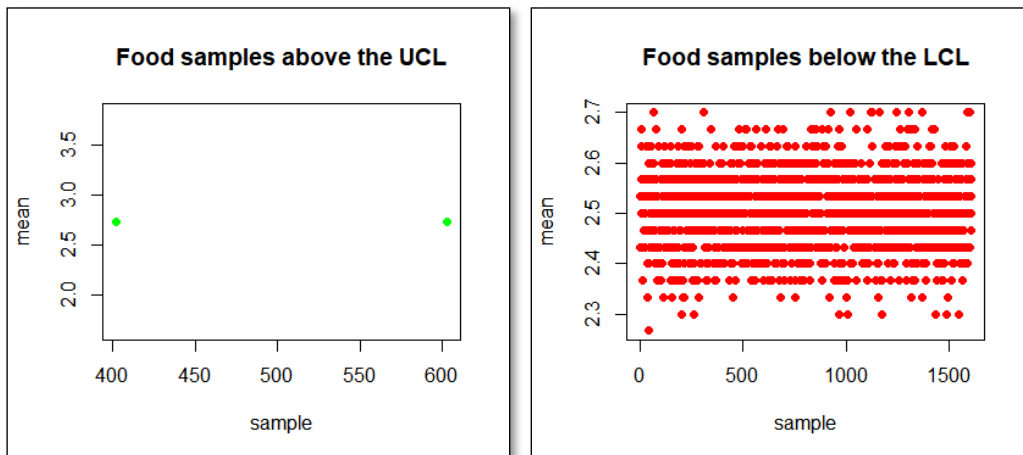


APPENDIX D: Sample Controlled Data

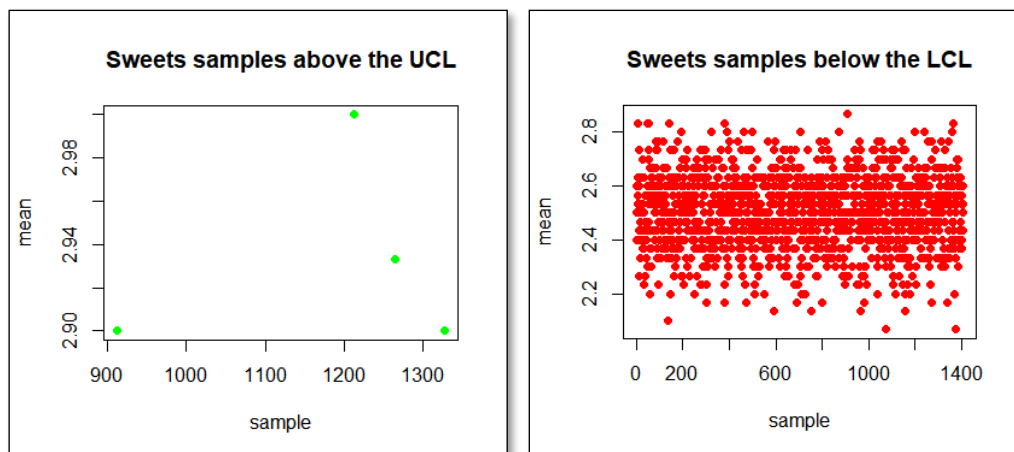
D.1. Clothing



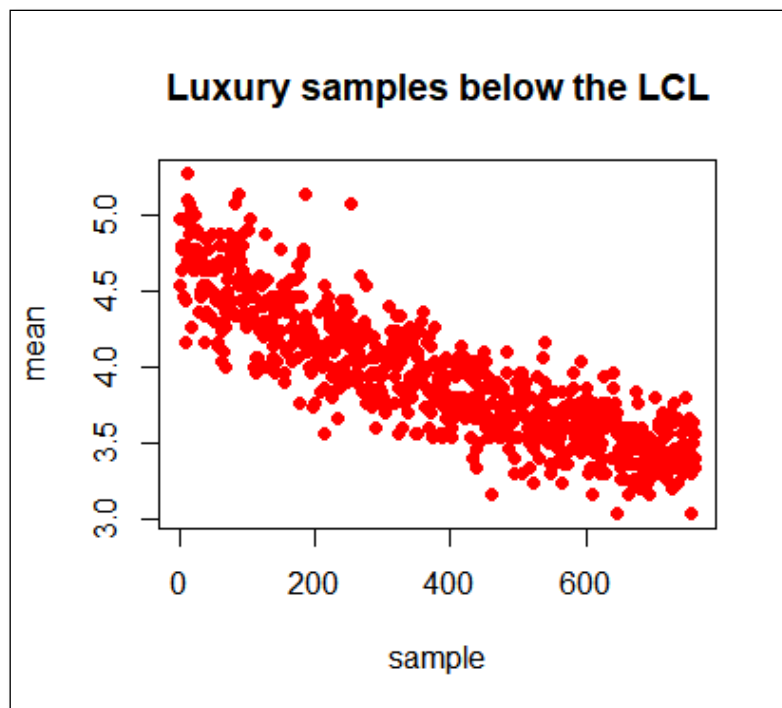
D.2. Food



D.3. Sweets

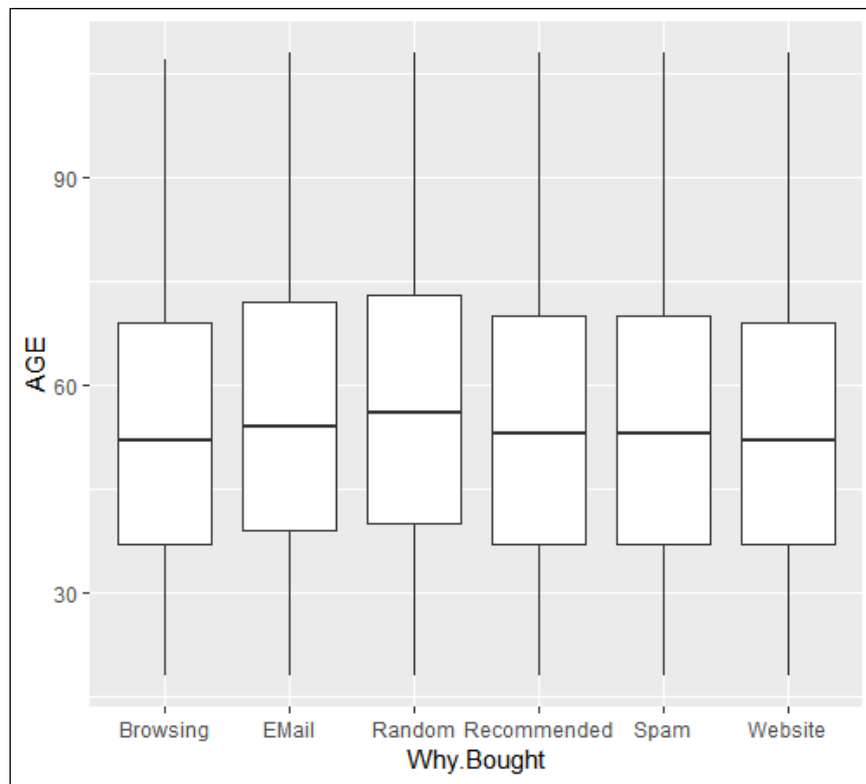


D.4. Luxury



APPENDIX E: Boxplots

E.1. Age



E.2. Price

