Quality Assurance 344 - ECSA GA4 Report

Prepared for: Engineering Counsel of South Africa

Prepared by: JW Cawood

(Student number 23929871)

Date: 21 October 2022

Abstract

This report attempts to provide documentation of the data wrangling and manipulation procedure for an online business' sales data. The procedure was carried out to transform unprocessed sales data into a useful database. Cleaning the provided raw data was the first step in the data wrangling procedure. The general distribution and underlying trends of several descriptive and target features were examined once the data had been cleaned. As part of the statistical analysis of the data, control limits for the sales delivery times of various item classes were also calculated. After that, control charts were created using the control limits to show which delivery times were out of control. When identifying out-of-control delivery procedures, the likelihood of creating a type 1 or type 2 error was also calculated. Following this, MANOVAs were conducted to see how the delivery time was impacted by the descriptive variables as well as how they interacted with one another. Finally, the reliability of the internet business's suppliers and internal delivery procedures were investigated.

Table of content

1. Introduction 1
2. Data Wrangling 1
   2.1 Features in Data Set 2
3. Descriptive Statistics 3
   3.1 Statistical Summary 3
      3.1.1 Numerical Features: 3
      3.1.2 Categorical Features: 6
   3.2 Analysis by Class 8
      3.2.1 Price Distribution per Class 8
      3.2.1 Age Distribution per Class 10
      3.2.2 Delivery Time Distribution per Class 12
4. Process Capabilities 14
   4.2 Indices: 14
5. Statistical Process Control 15
   5.1 Control Limits 16
   5.2 Control Charts 16
6. Optimizing Delivery Processes 20
   6.1 Control Chart Analysis 20
      6.1.1 Delivery Process Times S-Chart Analysis per Category 20
      6.1.2 Delivery Process Times X-Bar Control Chart Analysis per Category 21
      6.1.3 Consecutive Sample Standard Deviations 21
   6.2 Error Analysis 23
      6.2.1 Type I Error Analysis 23
      6.2.2 Type II Error Analysis 24
   6.3 Optimal Delivery Cost 24
7. MANOVA 25
8. Reliability of the Service and Products 25
   8.1 Lafrideradora Supplier 25
   8.2 Magnaplex Supplier 26
   8.3 Delivery Process Calculations 27
9. Conclusion 27
10. References 27
11. Appendix A – X-Bar Control Charts 28

Table of Figures

Figure 1:Histogram of Ages of Clients 4
Figure 2: Histogram of prices 5
Figure 3: Histogram of Delivery Times 6
Figure 4: Histogram of Reasons why the Product was Purchased 7
Figure 5: Frequencies of different classes 8
Figure 6: Boxplot for Price per Class 9
Figure 7: Joined Quantity vs Price Histograms 9
Figure 8: Boxplots for Age per Class 10
Figure 9: Joined Quantity vs Age Histograms 11
Figure 10: Boxplots for Delivery Time per Class 12
Figure 11: Joined Quality vs Delivery Time Histograms 13
Figure 12: Control Chart for Technology 17
Figure 13: Control Chart for Clothing 17
Figure 14: Control Chart for Food 18
Figure 15: Control Chart for Gifts 18
Figure 16: Control Chart for Household 19
Figure 17: Control Chart for Luxury 19
Figure 18: Control Chart for Sweets 20
Figure 19:Where clothing delivery process times went out of control 22
Figure 20: Where household delivery process times went out of control 23
Figure 21: Cost Function for increasing/decreasing delivery times 24

Table of Tables

Table 1: Summary of numerical features 3
Table 2: Statistical Summary for Delivery times of technology 14
Table 3: Process Capability Indices 14
Table 4: Control Limits for X-bar Charts for each class' delivery process times 16
Table 5:Control Limits for s-charts for each class' delivery process times 16
Table 6: Analysis of s-chart samples outside UCL/LCL 21
Table 7: Analysis of X-bar chart samples outside of UCL/LCL 21
Table 8: Most consecutive samples between -0.3 - +0.4 sigma-control limits 21
Table 9: Probabilities of making a Type I Error 23

1. Introduction

The quality of an online business is under investigation in this report. Client data is provided by the company and is statistically analyzed in order for the business to expand and improve in the most vital areas efficiently. This expansion will improve the business, allowing it to be more competitive in its market as well as ensuring that it can become more stable and long lasting.

The different approaches used in the statistical analysis will allow the business to identify its strong and weak points clearly. They can therefore improve in their weaker areas and use their strong points as opportunities to take advantage of. Areas that will be analyzed in different manners include their delivery times, their prices of different products, their amounts sold of products that fall under certain categories, who is buying their products and why and lastly the reliability of the services that they provide. The overall analysis will bring a clear view of the quality of the business and will show the vital role that the assurance of quality in a business actually plays.

2. Data Wrangling

Data wrangling is a process of cleaning the data and transforming it so that it only contains valid data as well as making it easy to use. Once data wrangling is complete, it is much easier for the business to analyze and use business intelligence to make decisions that aren't affected by false or invalid data. The data becomes structured and more efficient to use which visualization and statistical applications need so therefore it is vital to incorporate data wrangling into your approach of analysis. (Pearce, 2020)

In the case of this data set, the raw data set contained 180000 instances of which 23 instances contained missing data in the "Price" feature. This was the only feature containing missing data and the explanation to that is the value of the product probably got lost in the systems, or there could have been promotions on these products that the system did not take into account. The only other invalid data that was in the data set included negative values which was also in the price feature. These are invalid instances as there cannot be a negative price on the product. These errors could have occurred due to human error. These instances were all removed from the data set and a new, valid data set was created with a new index that included 179978 instances. This means that about only 0.00013% of the data was invalid. Therefore, removing the data was the best solution in the data handling as it will not make much of a difference to the overall outcome of the statistical analysis.

2.1 Features in Data Set

In the valid data set that was attained by removing incomplete cases and indexing the data, there are certain features used to describe the instances provided. The features are either; categorical - which implies that the variable type contains two or more categories or; they are numerical – which, on the other hand, means the variable type contains a numerical value that can be measured against other values. (Jain, 2020)

These features in this data set include:

- **ID**: Although this feature contains numerical values, it is actually a categorical feature as the numerical values do not have any relevance to one another. This value is a unique number given to each instance in the data set to give the instance an identification.
- **Age**: This is a numerical feature which displays the age of the customer that purchased a certain product in each instance.
- **Class**: This feature is categorical. Each purchased product falls under a certain category and this feature indicates as to which class this product was in. The classes are as follows: technology, clothing, food, gifts, household, luxury, and sweets.
- **Price**: this is a numerical feature that indicates the price of each product purchased.
- **Year, Month and Day**: these are 3 different features that correlate and essentially specify the date of each purchase. These 3 features are seen as categorical features as the ratio between two different days, months or years are not meaningful.
- **Delivery time**: this feature indicates the number of hours taken for the product to be delivered from time of purchase. This is a numerical feature.
- **Why Bought**: this is a categorical feature which shows how the client heard of the online website or product on the website. There are six different categories that the client could choose from and they are: recommended (the client was recommended the website by another person), random (the client randomly found the website or the product on the website or there is no other category that explains the way they found it), email (the client was notified via email about the product or the website), browsing (the client found the website or product on the website while browsing the internet), spam (the client was notified via spam emails sent out by the company about the website or product) and lastly, website (the client found the product while on the company's website).

3. Descriptive Statistics

Descriptive statistics are in the forms of measures of central tendency and measures of variability. Measures of central tendency include statistics such as mean, median and mode. Measures of variability refer to the spread of data and these measures include standard deviation, variance, skewness, and minimum and maximum values. These measures are illustrated in a way that helps summarize and understand the features of large data set. (Hayes, 2022)

In this particular set of data, it was vital to choose the best suited measures in order to gain quantitative insight to such a large set of 179978 instances. These were illustrated in manners such as various histograms, boxplots and various tables containing the values of the measures.

3.1 Statistical Summary

A statistical summary provides the gist of the data set provided. As each feature was defined as either categorical or numerical in section 2.1, different types of features will be summarized and analyzed in different ways. Numerical features are only relevant in a form of measures of tendency and spread whereas with categorical features, it is more sensible to view them in some form of frequency distribution.

3.1.1 Numerical Features:

The overall data set first needs to be understood so there is a statistical summary of all numeric features in table 1 below. From there onwards, the features will be broken up into classes and analyzed in that form. Certain features had numerical variables in their columns but weren't necessarily classified as numerical features as their values weren't relevant to one another (date, month, year, and ID), therefore these were excluded in the statistical summary in table 1:

Table 1: Summary of numerical features

Measure	Age	Price	Delivery Time
Mean	54,57	12294,1	14,5
Minimum	18	35,65	0,5
1st Quartile	38	482,31	3
Median	53	2259,63	10
3rd Quartile	70	15270,97	18,5
Maximum	108	116618,97	75
Range	90	116583,32	74,5

ANALYSIS:

- **Age:**

  The age range of 90 is very large, this shows that the online business has a very large target market. From the the age distribution graph below in figure 1 below, you can see that the clients ages are skewed-right. This could be due to the elderly not being so technologically inclined. In table 1, just looking at the values, it looks evenly spread out. This is due to the frequency of older ages not being too low. It is illustrated that the business does not include the target market of teenagers and younger, only young adults and older. There is a very wide target market.

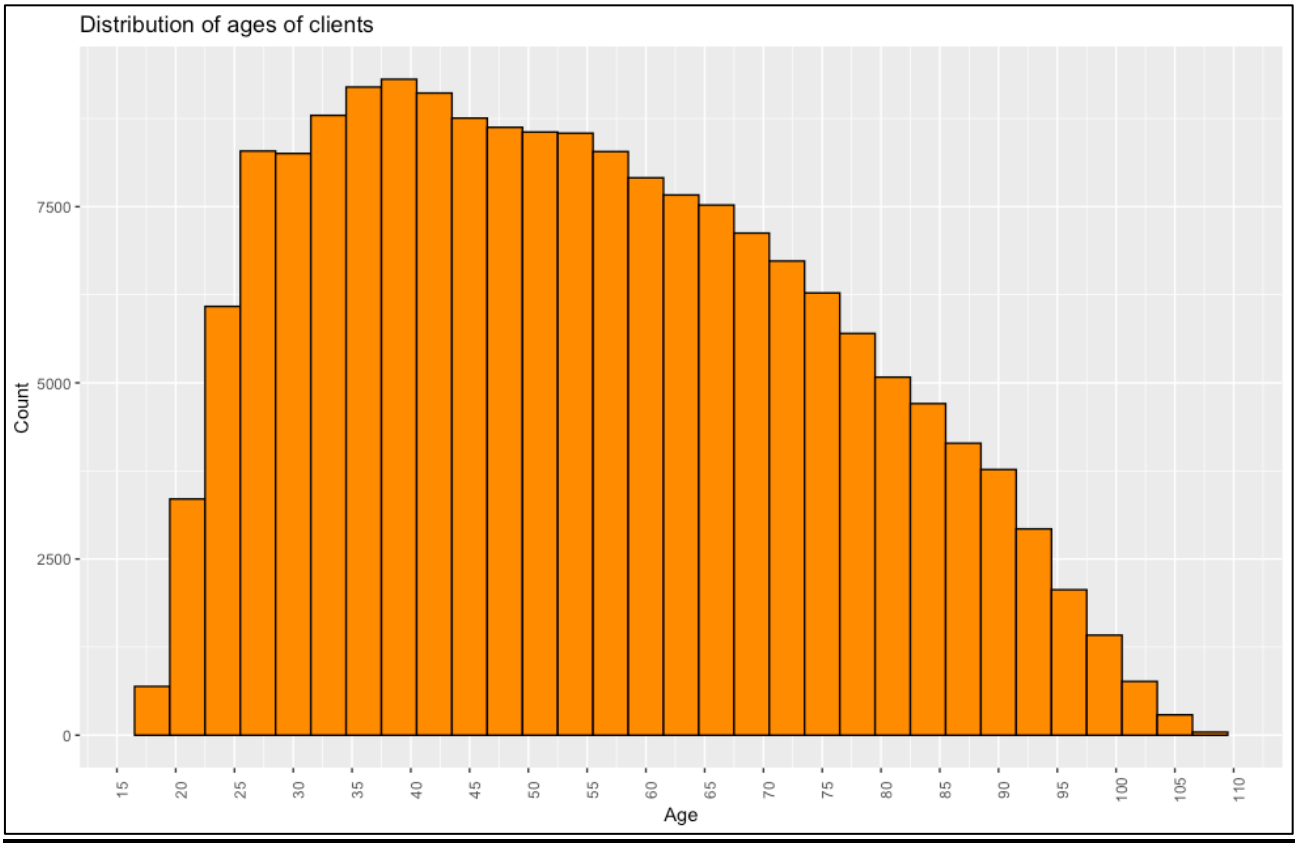


Figure 1:Histogram of Ages of Clients

- **Price:**

  In the statistical summary in Table 1, if you relate it to the histogram plotted in figure 2, you will see that the mean is not a very accurate representation of the data. The data is skewed to the right and the median seems to be the more accurate value. The 3rd quartile indicates that 75% of the prices are below R15270,97 whereas the rest of the 25% a range of from R15270,97 to R116618,97. This is due to there being few instances with very high prices

compared to most of them which drives the mean to be so high. This therefore indicates that majority of sales are made with items in a lower price range.

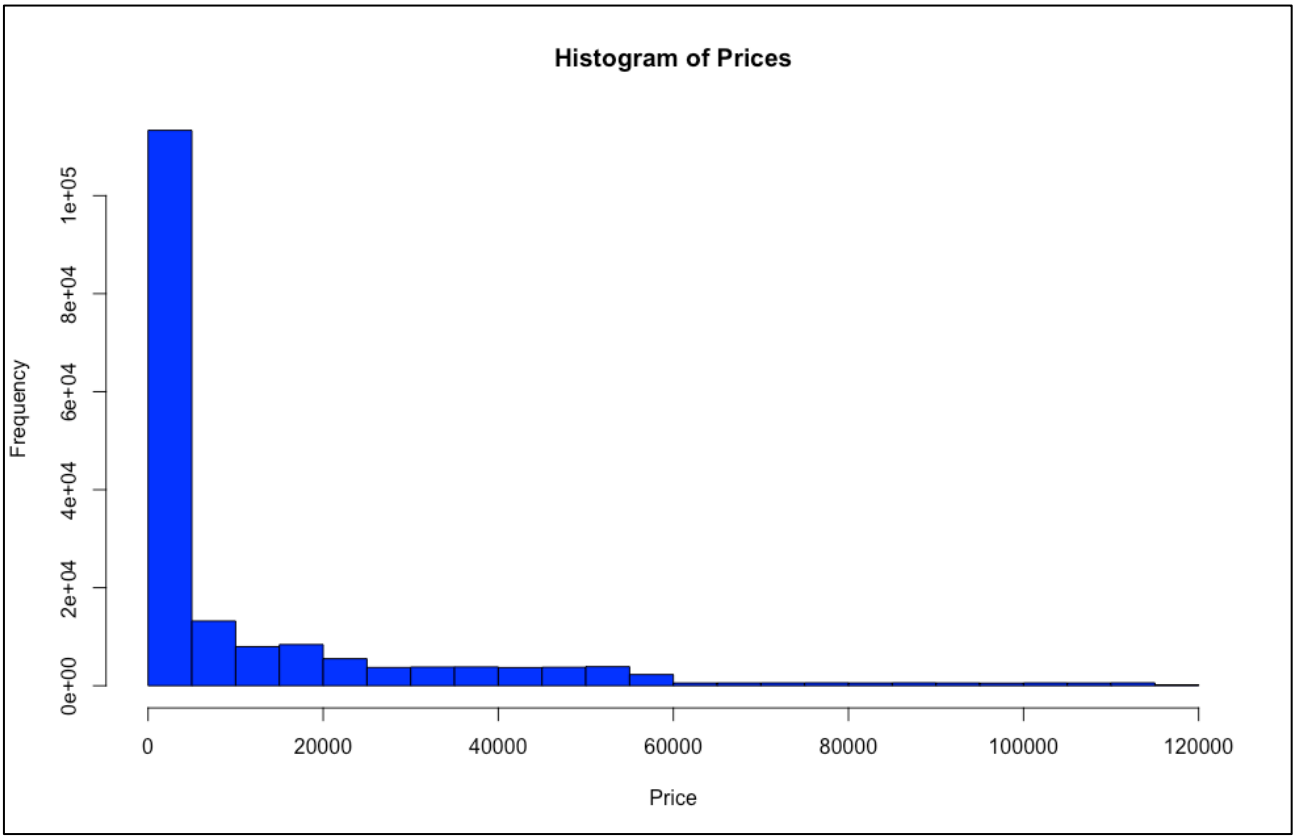


Figure 2: Histogram of prices

- **Delivery time:**

  The histogram for delivery time frequencies in hours in Figure 3 below as well as the 5 point summary for delivery times indicates that the data is right skewed because of the 3rd quartile being at 18,5 and the maximum being at 74,5. With the mean being 14.5 it means that there are a few outliers that have taken longer than normal to deliver. The median is still probably a truer representation of the spread of the data when looking at the histogram. These outliers need to be investigated as to why the delivery times took so long.

Figure 3: Histogram of Delivery Times

3.1.2 Categorical Features:

There are only two categorical features which have any relevance in terms of statistical analysis. These include the why bought feature and the class feature. The frequency of each category shall be compared and illustrated within the feature.

ANALYSIS:

- **Why Bought:**

  This feature contains categories which explain why or how the client came across their product that they purchased in each instant. Figure 4 below clearly illustrates that the method of customer exposure that dominates them all is clients recommending the online website to each other. This indicates that customer engagement for this online business is at a good standard. The voice of customer must have positive opinions if they are recommending the product to new clients. It shows that customer service, the business operations and products that are delivered are positive in the customer engagement side of things. There are opportunities in other categories such as the website as it is the second highest frequency in the boxplot. Them finding their product because of the website means if they could improve their website, they could utilize the opportunity and have more than

one main marketing method of client exposure. Emails and spam seem to be the lowest two categories, and this could be because they are most generally seen as annoying in the eyes of the customer. These two need to be investigated and costs and effort could be reduced in this area and rather utilized elsewhere.

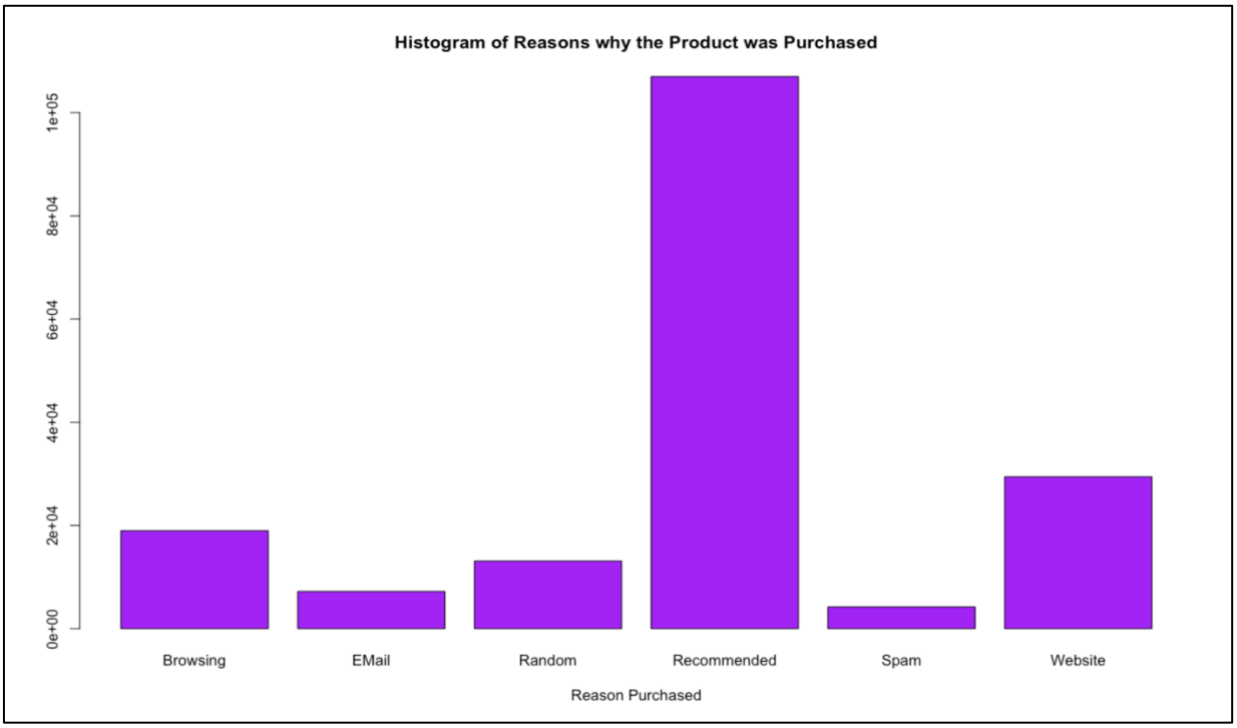


Figure 4: Histogram of Reasons why the Product was Purchased

- **Class:**

  As seen in figure 5 below, gifts and technology are purchased most frequently. The fact that it is an online business makes a large impact on the frequencies below. For example, food is a necessity but majority of the time you buy that in-person at a store – the same goes for sweets. Luxury items are clearly the lowest and this is due to a large quantity of the items having a high price, so they are purchased infrequently. A statistical analysis per class will be performed further on in the document for a more accurate representation of certain measures.

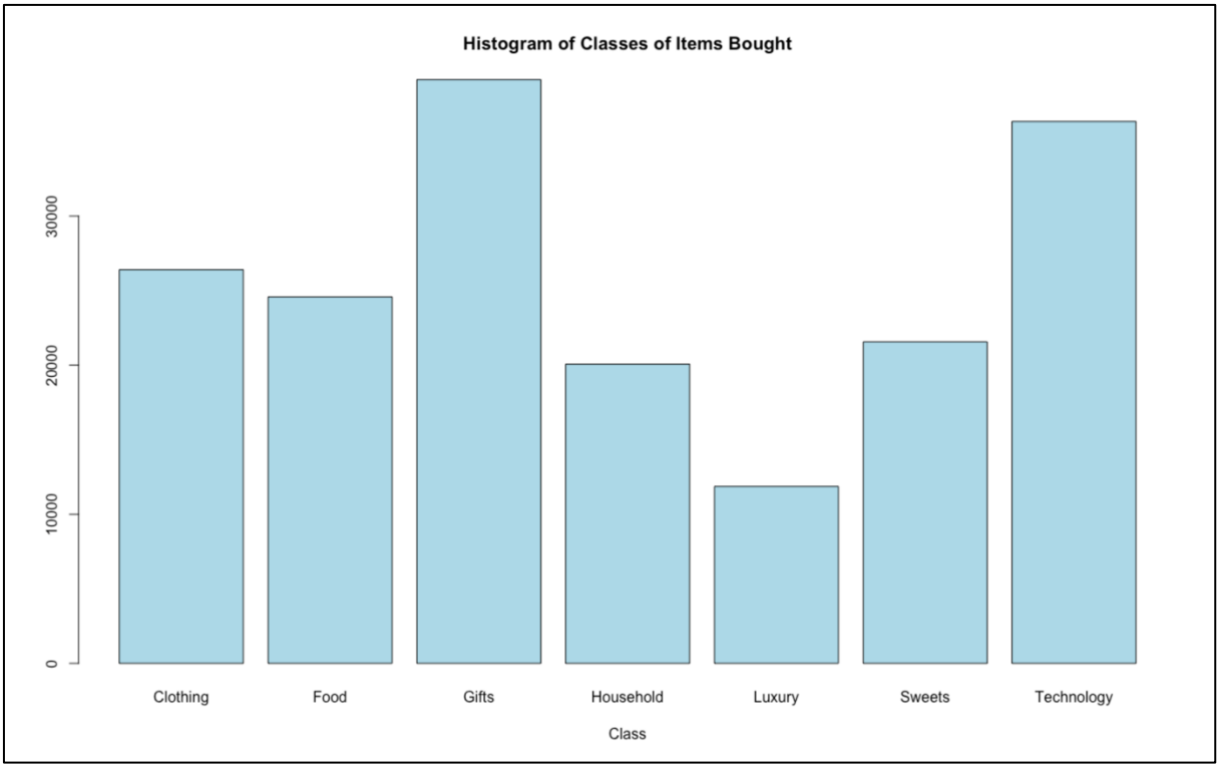


Figure 5: Frequencies of different classes

3.2 Analysis by Class

The analysis of the data by class will allow there to be deeper investigations into the online business because when the instances are grouped into their respective classes, the data is significantly more relevant than illustrating the values from the whole data set. There will be different types of visual representations as compared to when the features of the data set were being individually analyzed. For example, boxplots will be included in this section of the descriptive statistics as they are able to provide an indication of the data's symmetry and skewness.

3.2.1 Price Distribution per Class

The various boxplots combined in Figure 6 below illustrate the measures of central tendency visually. The range for each class varies drastically between some of them. You can see that luxury has the biggest range due to the prices being high for luxury items. Technology also has a high range and as seen in figure 5. Clothing, food, sweets, and gifts have a very low-price range as there are no major items of quality that have a high cost in those sections. Household items fall in between the high-priced and low-priced ranges, and this is due to the cost of things like furniture.

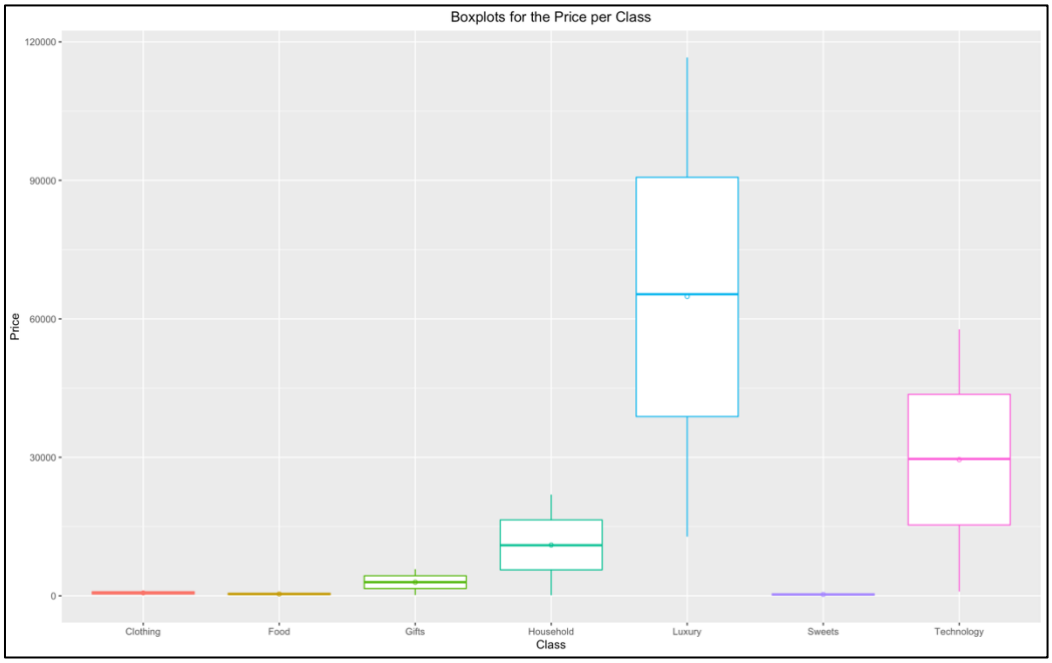


Figure 6: Boxplot for Price per Class

Figure 7 below is all the histograms of the quantities vs prices per each class joined together. When analyzing and relating figure 6 and 7 together, there are certain conclusions that arise for each class. Technology is one of the most frequently bought items, even though the quantities look small, the summation of all those small quantities adds up to a majority. Therefore, company needs to ensure that they utilize the situation of technology and manage it well as it is their biggest source of income. Clothing, food, sweets, and gifts have quantities in smaller ranges. It is vital that the company ensures a high and constant frequency for these classes for them to maintain a reasonable income in these classes. Luxury is difficult to see due to the scale of the graphs but it has small quantities at a very high price range.

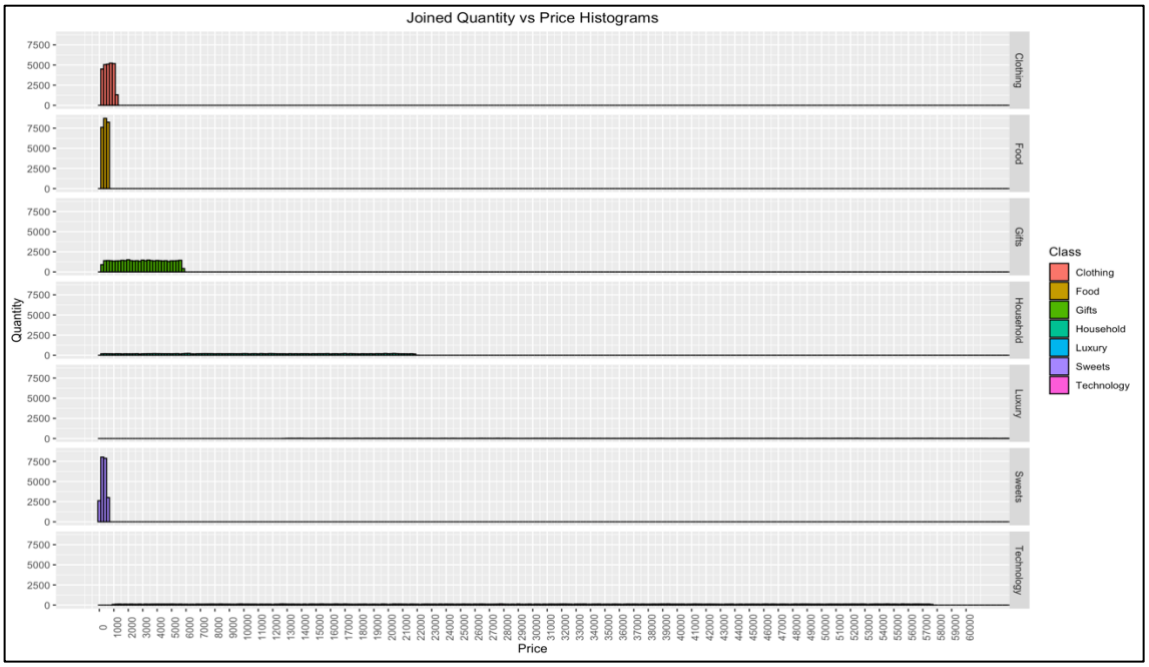


Figure 7: Joined Quantity vs Price Histograms

3.2.1 Age Distribution per Class

Figure 8, below, is a box plot that represents the age distribution for each class for the purchases made. This visual indicates that clothing, household, luxury and especially technology is purchased by younger clients. Food, gifts, and sweets seem to be purchased by older clients. The means are balanced throughout the classes, but this is due to outliers balancing them out in each class, the median is a more accurate representation.

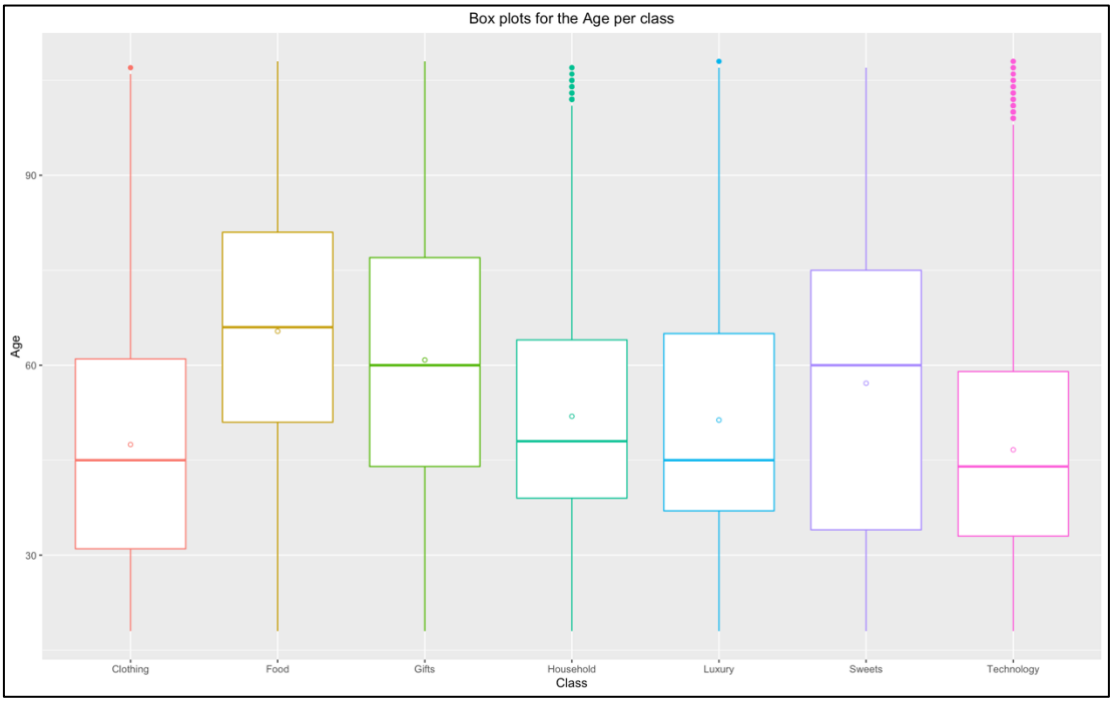


Figure 8: Boxplots for Age per Class

The following analyses are based on the histograms for quantity vs age in figure 9 below:

- **Clothing:**

  The clothing class age distribution is right-skewed. This shows that their clothing items are being purchased more frequently by the younger ages. The business should investigate this situation as they might only have clothing that appeals to younger clients. They will need to expand their clothing range in order to have the age more evenly distributed throughout the ages.

- **Food:**

  This distribution slightly is skewed to the left. Clients under the age of 40 rarely purchase food from the online business. This skewness needs to be investigated by the company as it is difficult to explain the reasoning for this. Food is an essential need so it could be that younger generations are rather purchasing their food in stores rather than online.

- **Gifts:**

  The histogram for gifts is uniformly distributed. This is a positive outcome as it indicates that all ages are purchasing gifts. The distribution does not need to be investigated further.

- **Household:**

  This age distribution is skewed to the right. Younger age groups are purchasing household items more frequently. Clients of ages between 25 and 55 are the main customers of household items which is explainable due to these years of a humans life being the years in which people are buying and moving into households where new household items are needed.

- **Luxury:**

  Luxury items are normally purchased in years of your life where you are earning a source of income as you have spare money to spend in some situations. This explains why the graph is skewed to the right and there is a peak between the ages of 30 and 50.

- **Sweets:**

  The distribution of purchasing frequencies of ages for sweets is bimodal. There are two peaks in the graph, at the ages between 20 and 35 and then another peak at the ages between 55 and 90. The company needs to investigate why there is a gap in sweet purchasing between the ages of 35 and 55.

- **Technology:**

  The technology age distribution is right-skewed. This implies that younger aged clients are purchasing more technology items. The company could increase the frequency of the elderly purchasing more technology items by introducing programs that are more helpful in terms of understanding technology as they are less inclined.

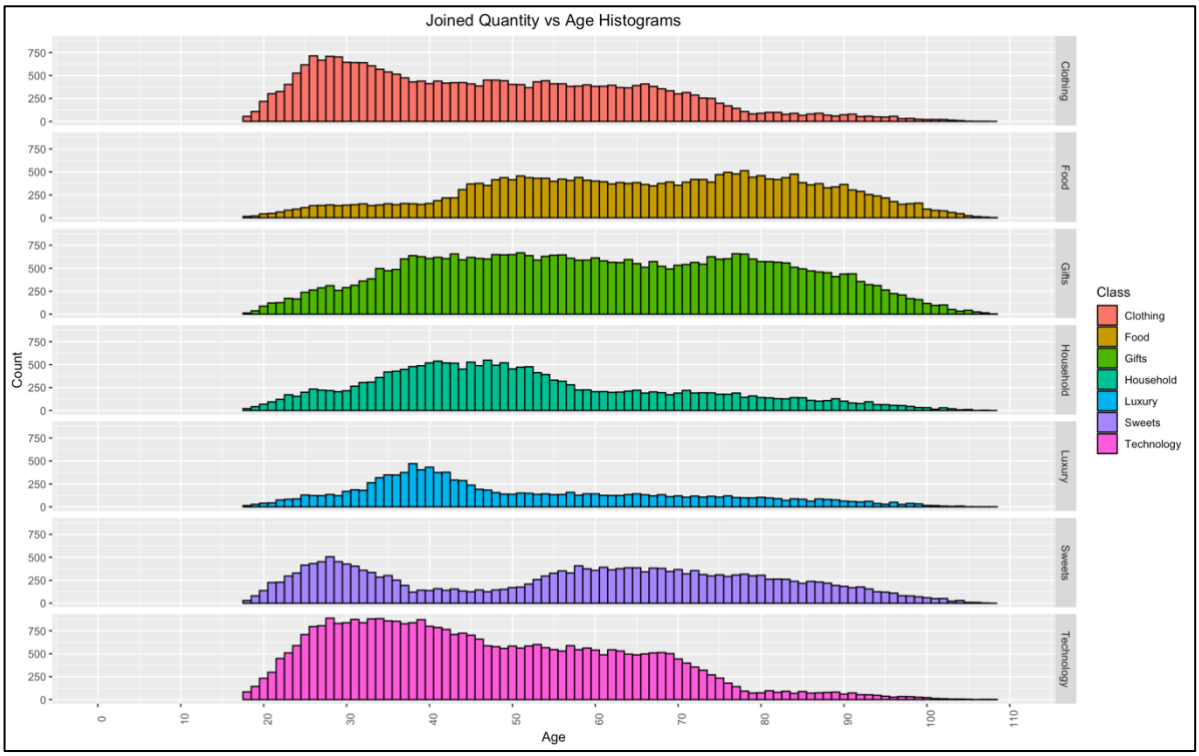


Figure 9: Joined Quantity vs Age Histograms

3.2.2 Delivery Time Distribution per Class

There is a large variation in distribution between certain classes, as seen in the boxplots in Figure 10 below. The delivery times for household items is much larger than any other class, some deliveries took up to 75 hours. Technology is second highest but significantly lower than household. The rest of the classes' delivery times are low, except for gifts and clothes having slightly higher means.

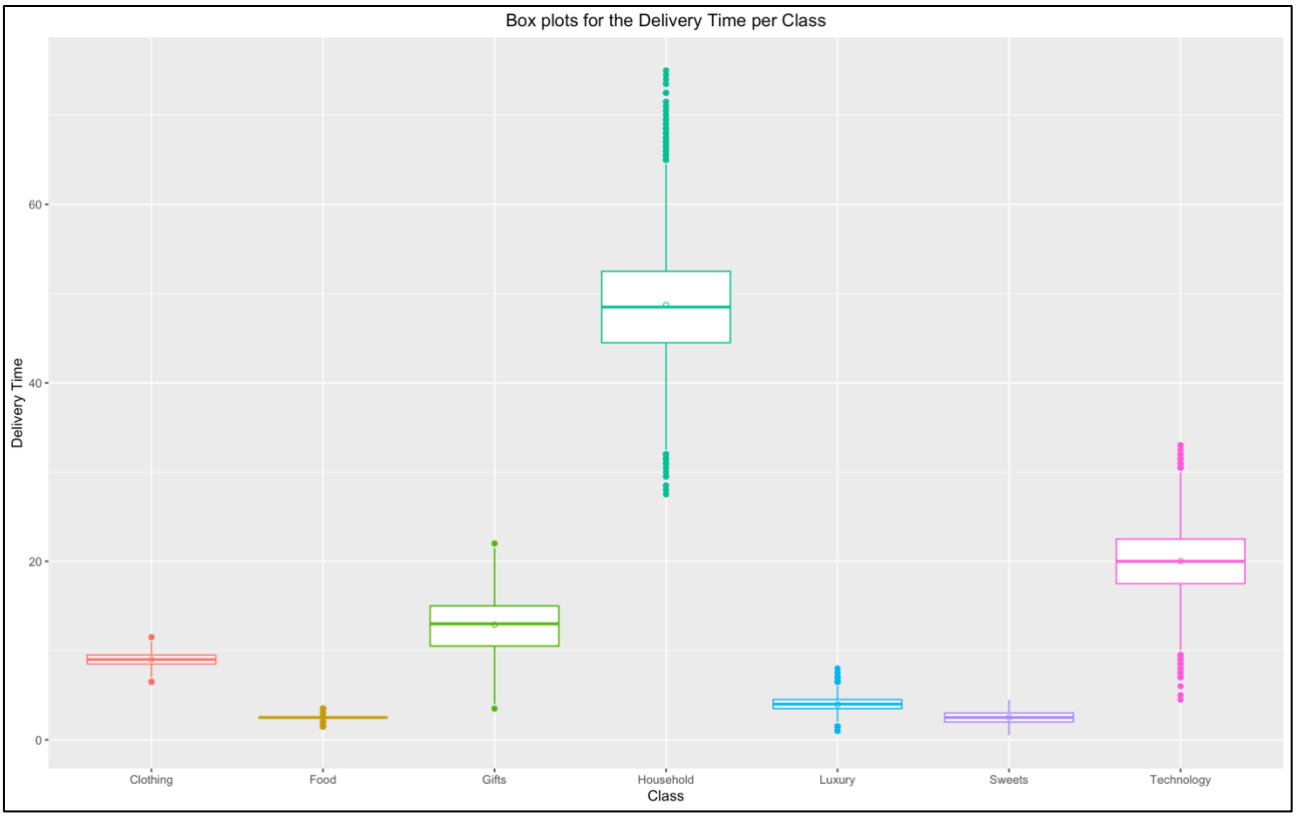


Figure 10: Boxplots for Delivery Time per Class

Looking at figure 10 and 11, it is easy to see the correlation between the two graphic illustrations. Below are the analyses of each class on their delivery time distribution:

- **Household:**

  Household items' delivery times stand out compared to the rest. The items are more distributed over a larger range, and all the values are significantly higher than the rest of the classes. This could only be explainable if it were to be customized orders on household items such as furniture. If there are no customized options on the website, it means that the business does not keep the furniture and other items in stock. The company needs to investigate this and possibly look at larger inventory capacity to be able to keep larger items in stock.

- **Technology:**

  Technology items' delivery items have a mean of 20 hours. The standard deviation is not as large as household items. The mean could be high due to certain technology items not being

available in the country consistently. The business needs to investigate how they can decrease the delivery times for technology.

- **Food, Luxury, and Sweets:**

  These three classes have very similar distributions, they all have short delivery times with small standard deviations. For food and sweets, it is expected as they are perishables and they need to be prioritized in terms of delivery in order for a customer to have a longer shelf life on their products. Luxury has a good delivery time distribution and does don't need to be investigated.

- **Gifts:**

  Gift items' delivery times are uniformly distributed but have a range from 5 hours to 25 hours. This could be decreased and needs to be investigated as there is no reason why gifts and luxury should differ that much.

- **Clothing:**

  Clothing has a small standard deviation which is good, but its mean is higher than luxury, food, and sweets. The mean isn't a large difference and is not a critical problem but can be investigated by the business to be potentially improved.

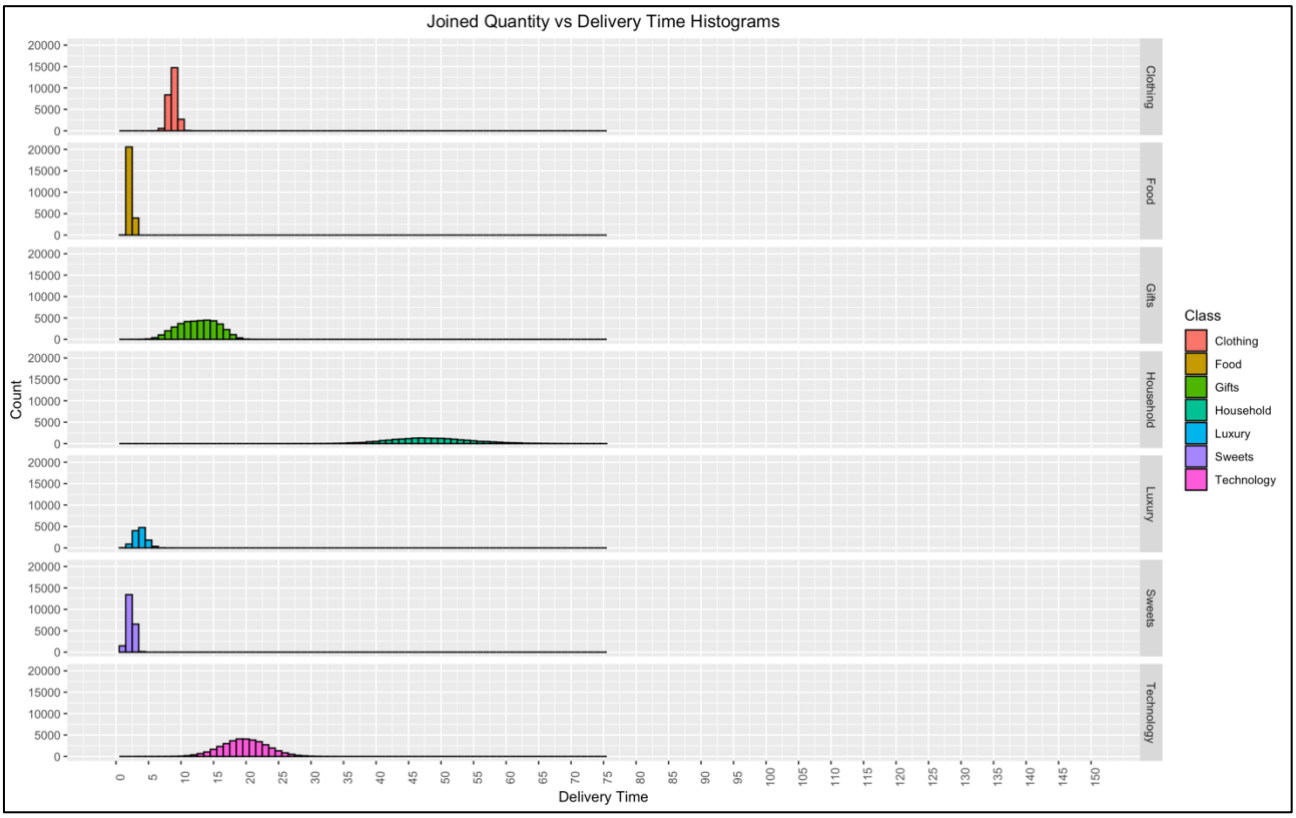


Figure 11: Joined Quality vs Delivery Time Histograms

4. Process Capabilities

The process capability of the delivery time of technology classed items will be measured. This mean that indices will measure how capable the process is to be able to delivery technology items within the specified limits on a consistent basis. These indices will predict the ability of the delivery process by using the subset of technology class instances from the valid data set. Four indices will be calculated as a measurement, namely: Process Capability (C_p), Upper One-sided Index (C_{pu}), Lower One-sided Index (C_{pl}), and Process Capability Index (C_{pk}). (Process Capability Analysis Cp, Cpk, Pp, Ppk - A Guide, n.d.)

Table 2 to the right shows a statistical summary of the delivery times of the technology class of which the values will be used in calculations.

Table 2: Statistical Summary for Delivery times of technology

Measures	Delivery Time Values
Standard Deviation	3,5
Mean	20,01
Minimum	4,5
1st Quartile	17,5
Median	20
3rd Quartile	22,5
Maximum	33

4.2 Indices:

The specification limits given were as follows:

- Upper Specification Limit (USL): 24 hours
- Lower Specification Limit (LSL): 0 hours

A lower specification limit is logical due to the fact that the delivery times given are in hours. It is specified as zero because if it is possible for a customer to make an order, have it processed and delivered all within an hour, it would result in a delivery of 0 hours.

From these limits and certain values in table 2, table 3 below indicates all the indices that needed to be calculated to analyze the systems process capability.

Table 3: Process Capability Indices

Cp	Cpu	Cpl	Cpk
1.142207	1.90472	0.37969	0.37969

- **C_p:** The process capability in table 3 of 1.142207 indicates that the spread of the delivery times is less than the width of the specification limits. Essentially this means that the delivery time process can fit inside the limits 1.142207 times. (Graham & Clearly)
- **C_{pu}:** The upper one-sided index in table 3 of 1.90472 indicates that the mean of the delivery times for technology items is very close to the upper specification limit as it is significantly greater than one. This shows that the process is very capable of meeting the upper specification limit. (Montgomery, 2012)
- **C_{pl}:** The lower one-sided index in table 3 of 0.37969 indicates that the mean of the delivery times for technology items is not close to the lower specification limit as it is much smaller than one. This shows that the process is not capable of meeting the lower specification limit. (Montgomery, 2012)
- **C_{pk}:** The process capability index for delivery times for technology items is 0.37969 as seen in table 3 above. It is the minimum value between the C_{pu} and C_{pl}. Since the C_p and C_{pk} are not equal and the C_{pk} is significantly less than one, it indicates that the delivery time mean is not centered between the specification limits. The ratio of C_{pk}/ C_p indicates how significantly off the delivery times were centered. 0.33242 is the calculated ratio and this means that the process is off target by 66.76%. The technology delivery time process needs to be heavily investigated.

5. Statistical Process Control

Statistical process control is a method using certain visual techniques to measure, control and improve a process. This method will benefit the business as having control over their processes will give them a competitive advantage. This process eliminates special cause variations. (Hessing, n.d.)

Delivery process times will undergo this method for each class using X-bar and s-charts for the statistical process control. X-bar control charts graphically illustrate the average change of the process over time and can evaluate whether each class' delivery process time is out of control or not. S-charts graphically illustrate the standard deviation of the process over time. These charts will use samples from data that has been ordered from the oldest instances to the newest. The first 30 samples of the ordered data set will be used with each sample size being 15 instances.

5.1 Control Limits

Tables 4 and 5 below show the values of the control limits for the X-bar and s-charts respectively.

Table 4: Control Limits for X-bar Charts for each class' delivery process times

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22,9746	22,1079	21,2412	20,3744	19,5077	18,6410	17,7743
Clothing	9,4049	9,2600	9,1150	8,9700	8,8250	8,6800	8,5351
Household	50,2483	49,0196	47,7909	46,5622	45,3335	44,1048	42,8761
Luxury	5,4940	5,2412	4,9884	4,7356	4,4828	4,2299	3,9771
Food	2,7095	2,6363	2,5632	2,4900	2,4168	2,3437	2,2705
Gifts	9,4886	9,1127	8,7369	8,7369	7,9853	7,6095	7,2337
Sweets	2,8970	2,7573	2,6175	2,4778	2,3380	2,1983	2,0585

Table 5:Control Limits for s-charts for each class' delivery process times

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5,1806	4,5522	3,9239	3,2955	2,6672	2,0388	1,4105
Clothing	0,8666	0,7615	0,6564	0,5512	0,4461	0,3410	0,2359
Household	7,3442	6,4534	5,5626	4,6719	3,7811	2,8903	1,9996
Luxury	1,5111	1,3278	1,1445	0,9612	0,7780	0,5947	0,4114
Food	0,4372	0,3842	0,3312	0,2781	0,2251	0,1721	0,1190
Gifts	2,2463	1,9739	1,7014	1,4290	1,1565	0,8841	0,6116
Sweets	0,8353	0,7340	0,6327	0,5314	0,4301	0,3288	0,2274

5.2 Control Charts

The values in the tables in section 5.2 will be incorporated in the control charts that are to be plotted to determine which processes are out of control. S-charts generated below will be evaluated for each class to determine which class' delivery process times will need to go under investigation. For the s-charts, the remaining samples that occur after the first 30 samples used for the limits will be used to plot the control charts. X-bar charts that were used to initialize statistical process control method can be found in appendix A. The Figures 12 to 18 below show the control charts for each class.

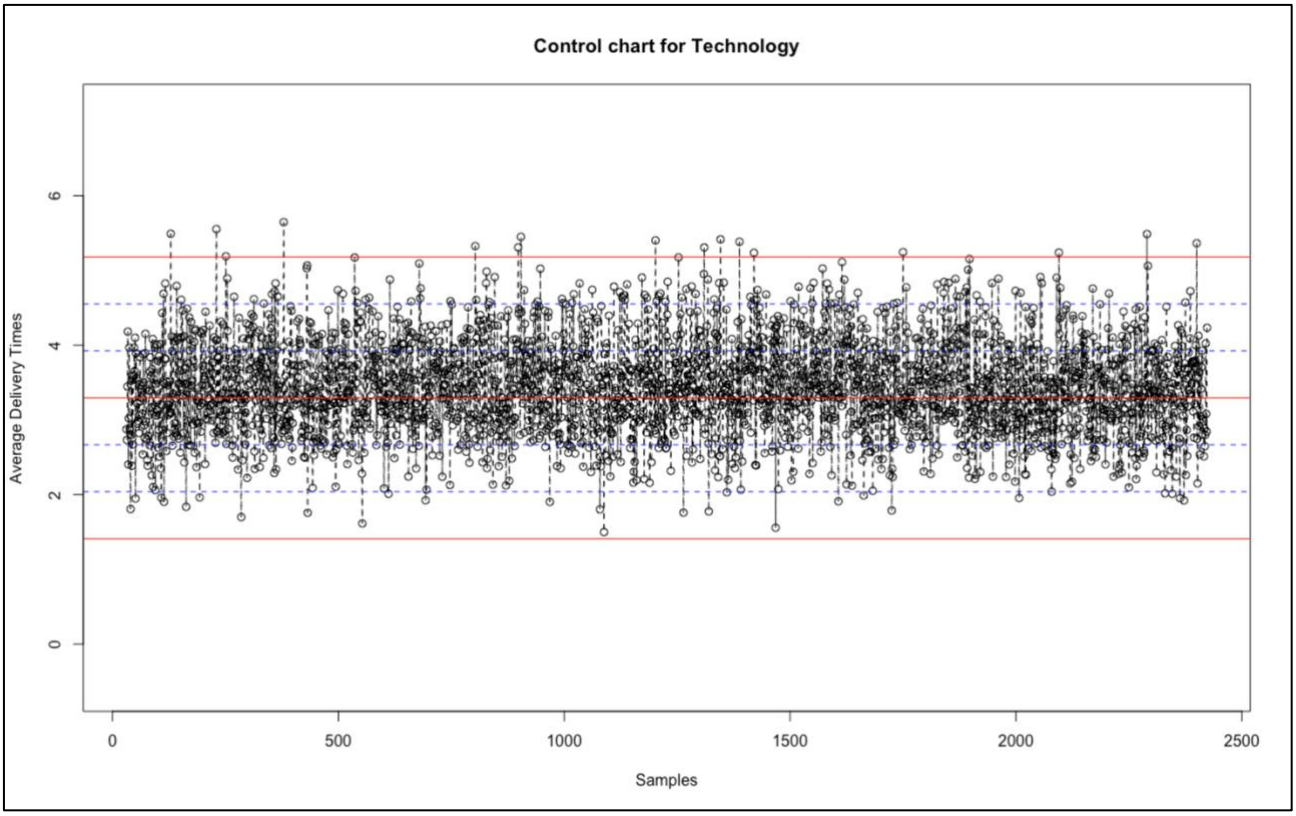


Figure 12: Control Chart for Technology

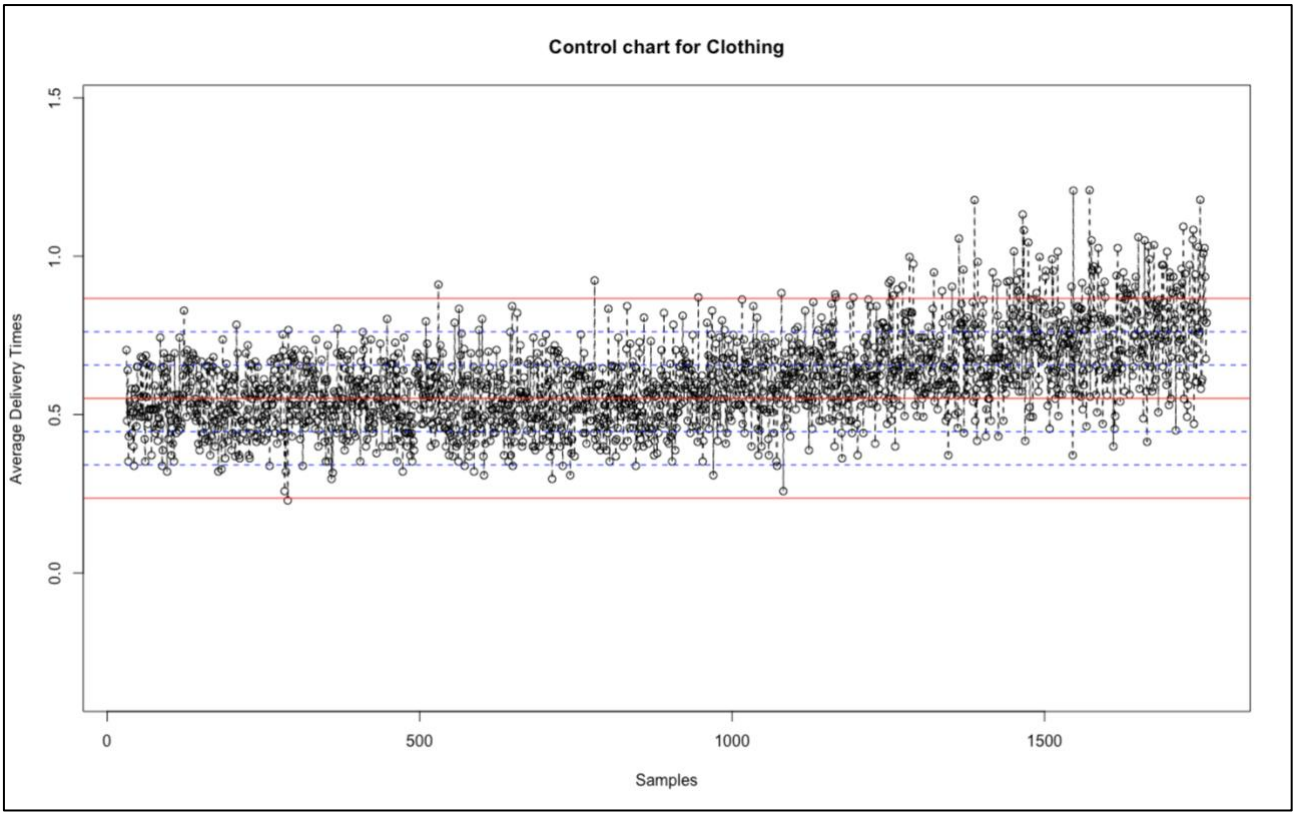


Figure 13: Control Chart for Clothing

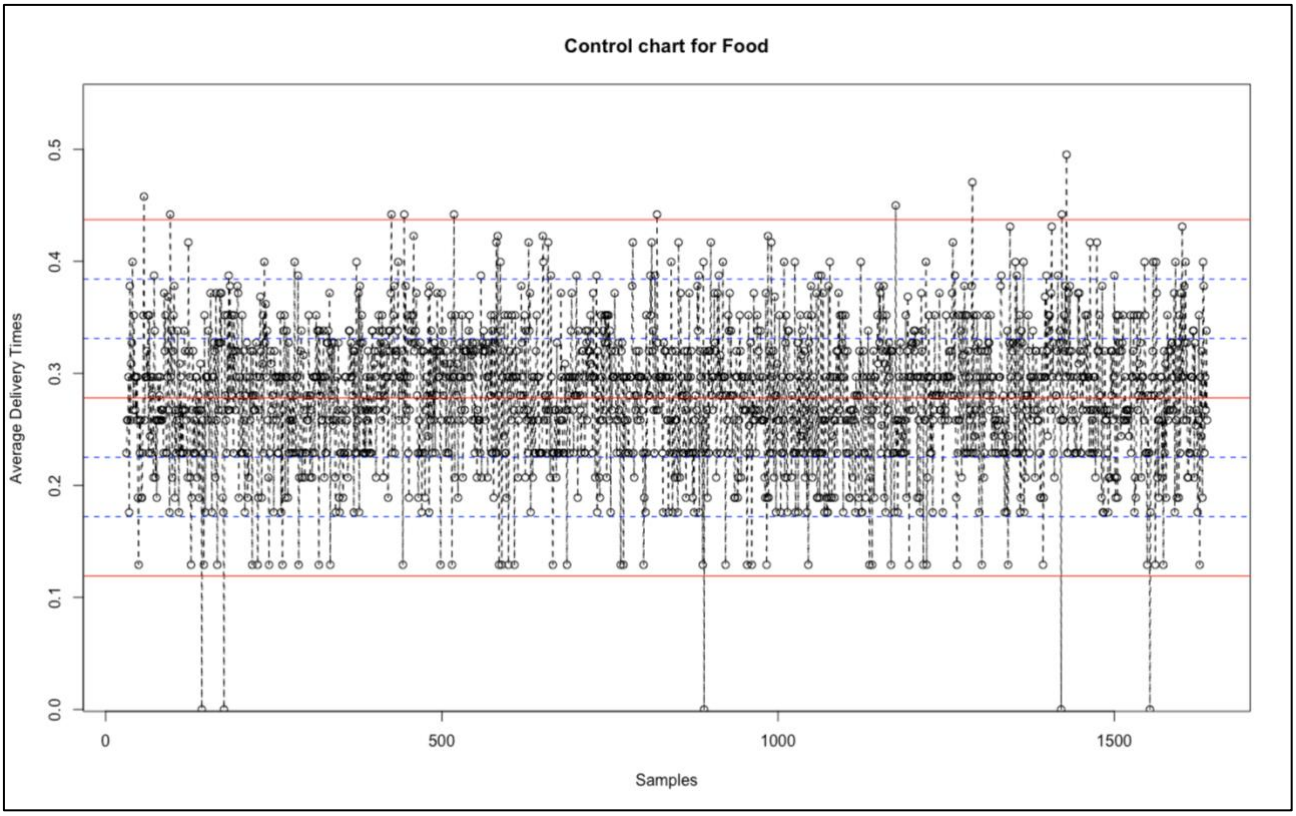


Figure 14: Control Chart for Food

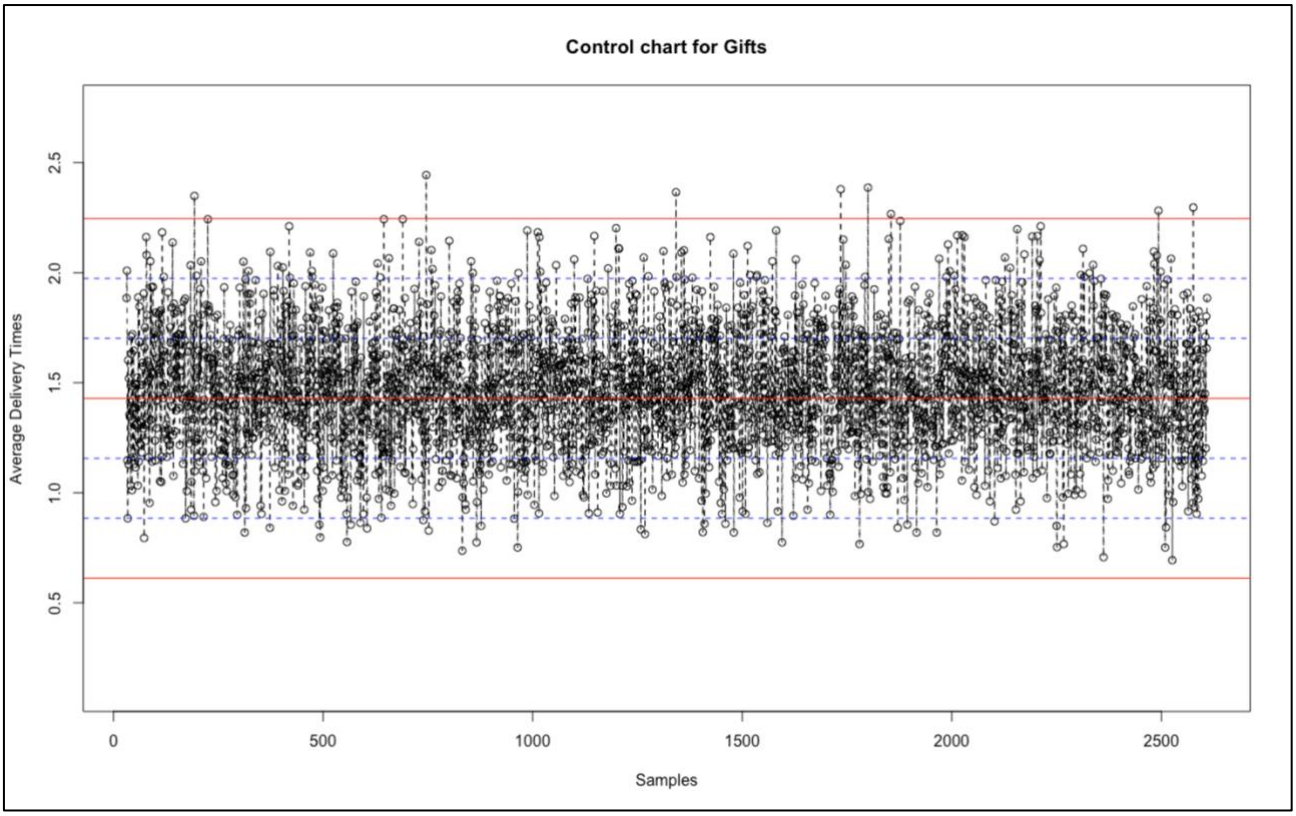


Figure 15: Control Chart for Gifts

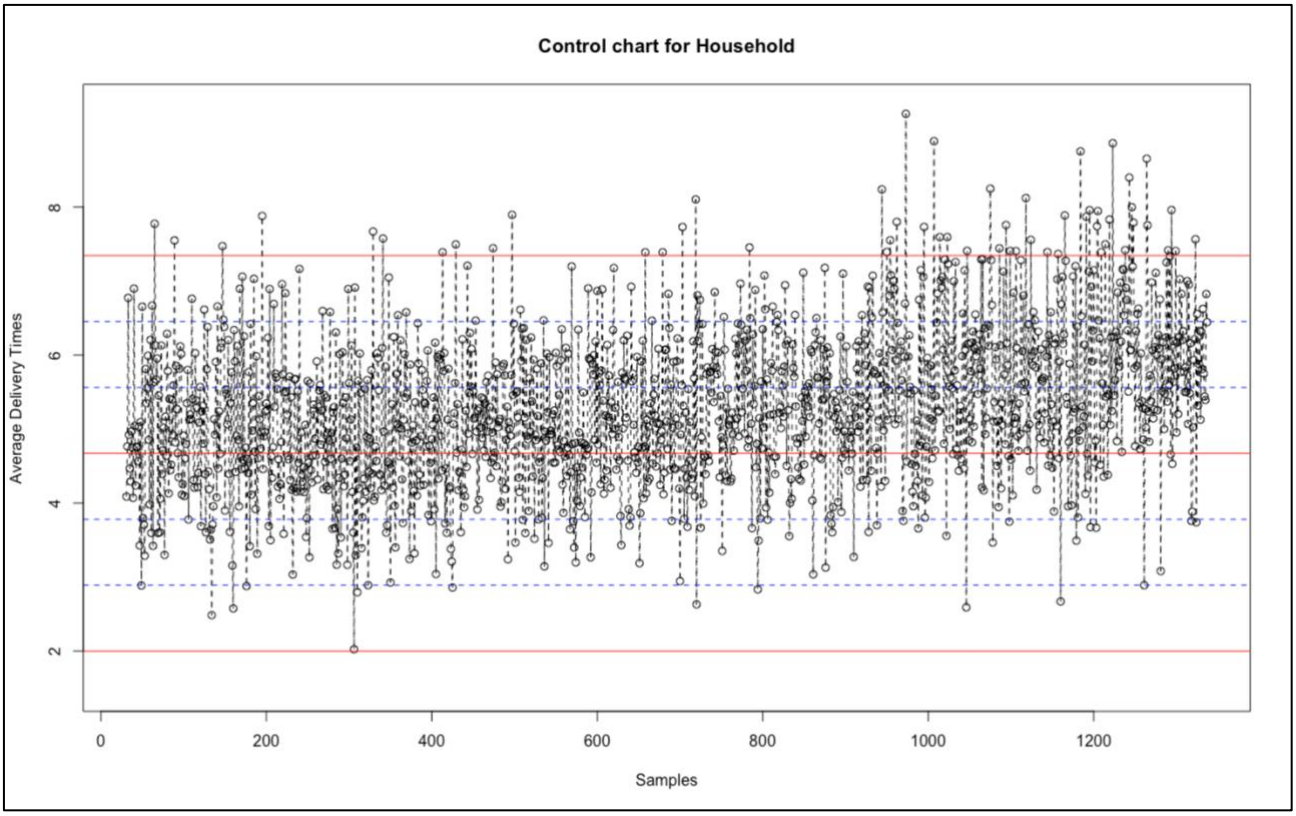


Figure 16: Control Chart for Household

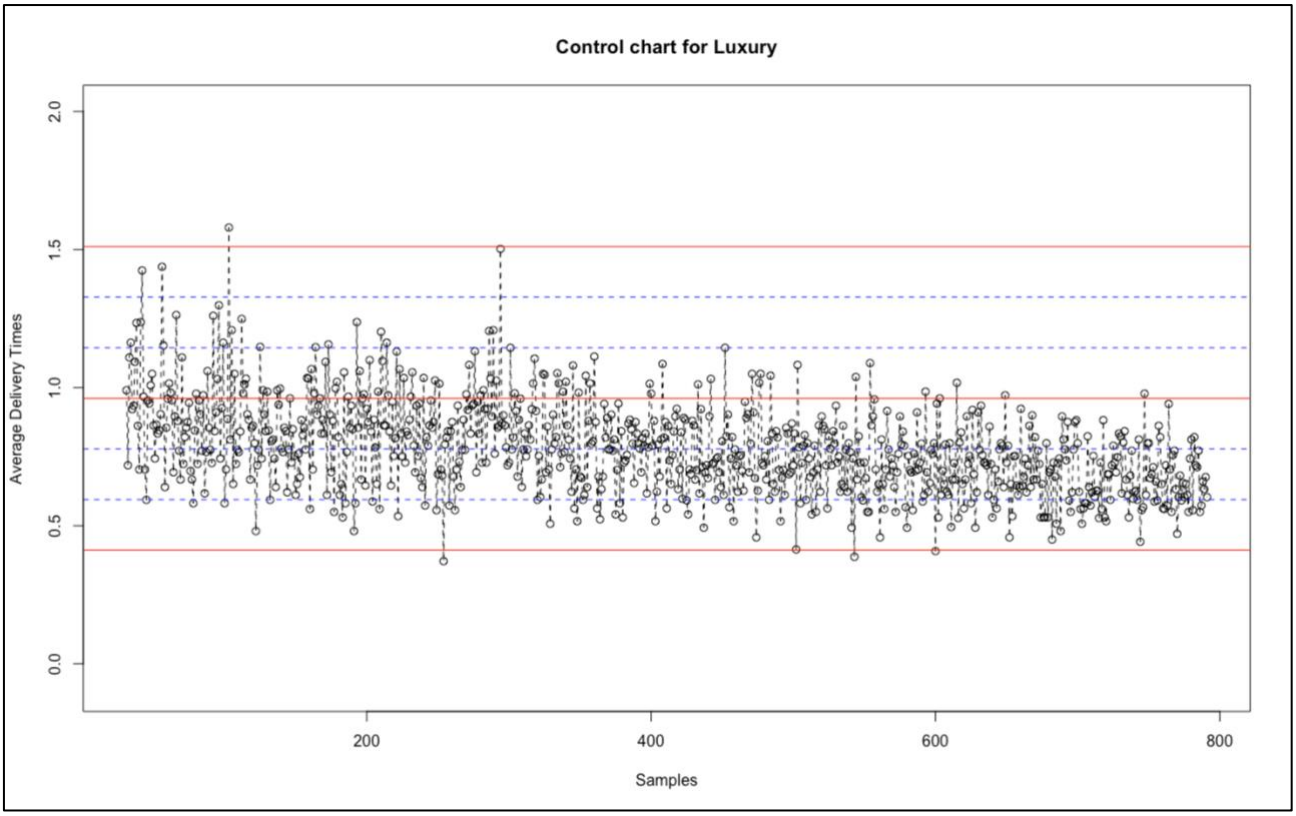


Figure 17: Control Chart for Luxury

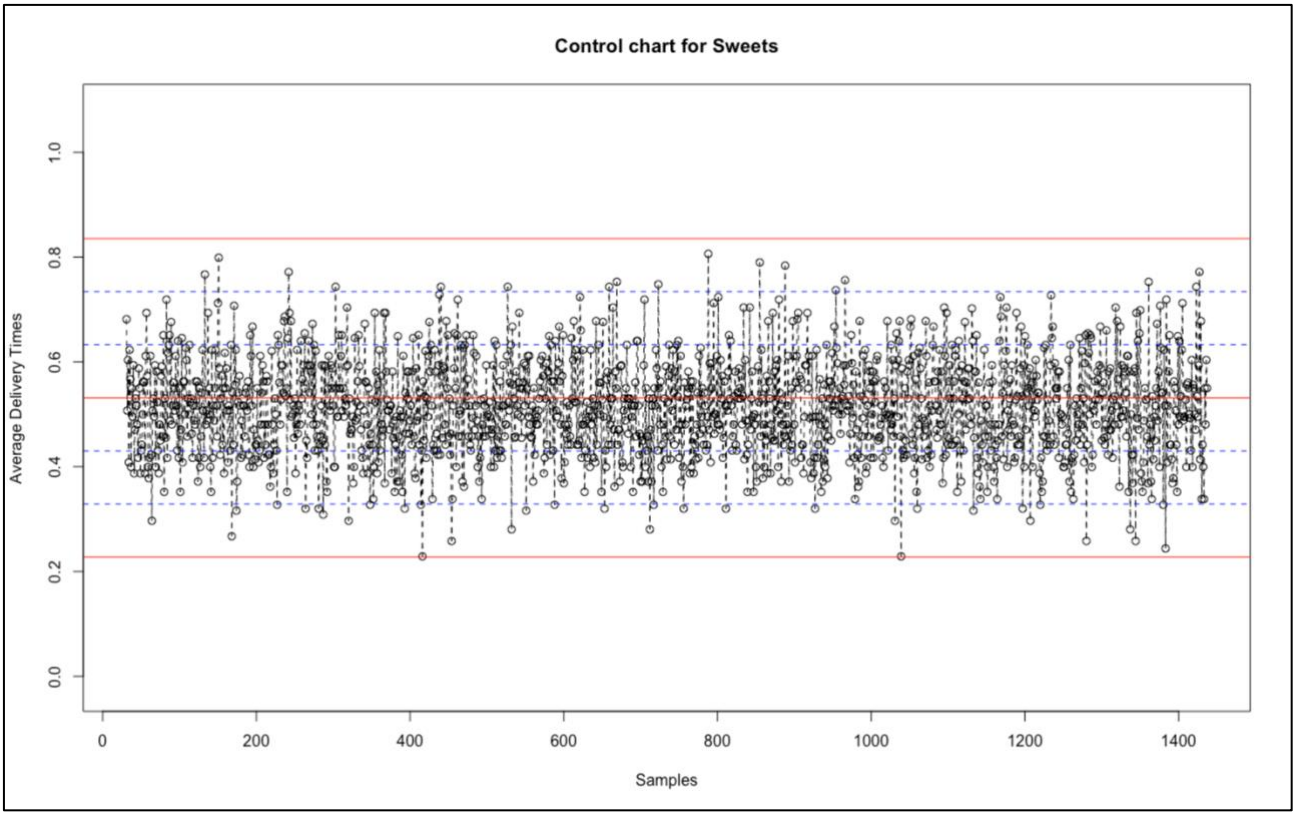


Figure 18: Control Chart for Sweets

6. Optimizing Delivery Processes

6.1 Control Chart Analysis

The control charts shown in section 5.2 can now be analyzed and the delivery process times for each class will be determined to either be in control or out of control. This is determined by 3 factors, namely, that the points on the control charts lie within the upper and lower control limits, that points are evenly distributed above and below the center line, and that there should be more points densely positioned closer to the center line rather than at the upper and lower control limits.

6.1.1 Delivery Process Times S-Chart Analysis per Category

The s-charts need to be analyzed by identifying samples that indicate an out-of-control process. The first 3 and last 3 positions of samples that are outside of the UCL or LCL of each class as well as the total number of samples that are outside of the UCL or LCL for each class (last column) are in table 6 below.

Table 6: Analysis of s-chart samples outside UCL/LCL

	Samples Outside of UCL/LCL						
Class	**1st**	**2nd**	**3rd**	**3rd Last**	**2nd Last**	**Last**	**Total**
Clothing	289	530	780	1754	1756	1757	98
Household	65	89	147	1294	1299	1323	54
Food	19	57	96	1422	1429	1553	16
Technology	129	230	251	2 095	2290	2400	16
Sweets	18	NA	NA	NA	NA	NA	1
Gifts	193	746	1342	1855	2493	2576	8
Luxury	103	254	543	NA	NA	600	4

6.1.2 Delivery Process Times X-Bar Control Chart Analysis per Category

Table 7 below illustrates the same as table 6, except the values shown that gives details about the samples is for the X-bar charts instead of the s-charts.

Table 7: Analysis of X-bar chart samples outside of UCL/LCL

	Samples Outside of UCL/LCL						
Class	**1st**	**2nd**	**3rd**	**3rd Last**	**2nd Last**	**Last**	**Total**
Clothing	455	702	1152	1677	1723	1724	17
Household	252	387	629	1335	1336	1337	400
Food	75	633	1203	NA	1467	1515	5
Technology	37	398	483	1872	2009	2071	17
Sweets	942	1104	1243	NA	1294	1403	5
Gifts	213	216	218	2607	2608	2609	2290
Luxury	142	171	184	789	790	791	434

6.1.3 Consecutive Sample Standard Deviations

Table 8 below identifies the most consecutive samples in the s-charts between the range of -0.3 and +0.4 sigma-control limits.

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

Class	**Most Consecutive Samples**
Technology	6
Clothing	4
Food	7
Gifts	5
Household	3
Luxury	4
Sweets	4

ANALYSIS:

By making use of the s-charts in figures 12 to 18 as wells as tables 6-8, the classes that had delivery process times that are out of control are analyzed below:

- **Clothing:**

The control chart in figure 13 for clothing delivery process times and table 6 above shows that it is evident that the clothing items delivery process times are out of control. 98 out 1760 samples (5.6%) in total are outside of the upper or lower control limits. You can see in figure 13 that from about sample 1300, the process becomes out of control. The delivery time increases which is not a good reflection on the process. Figure 19 below points out in red where the process becomes out of control on the s-chart. The process needs to be investigated by the business to discover what the reasoning was behind this slight increase and stop the process in order to be able to redesign it to decrease the delivery times.

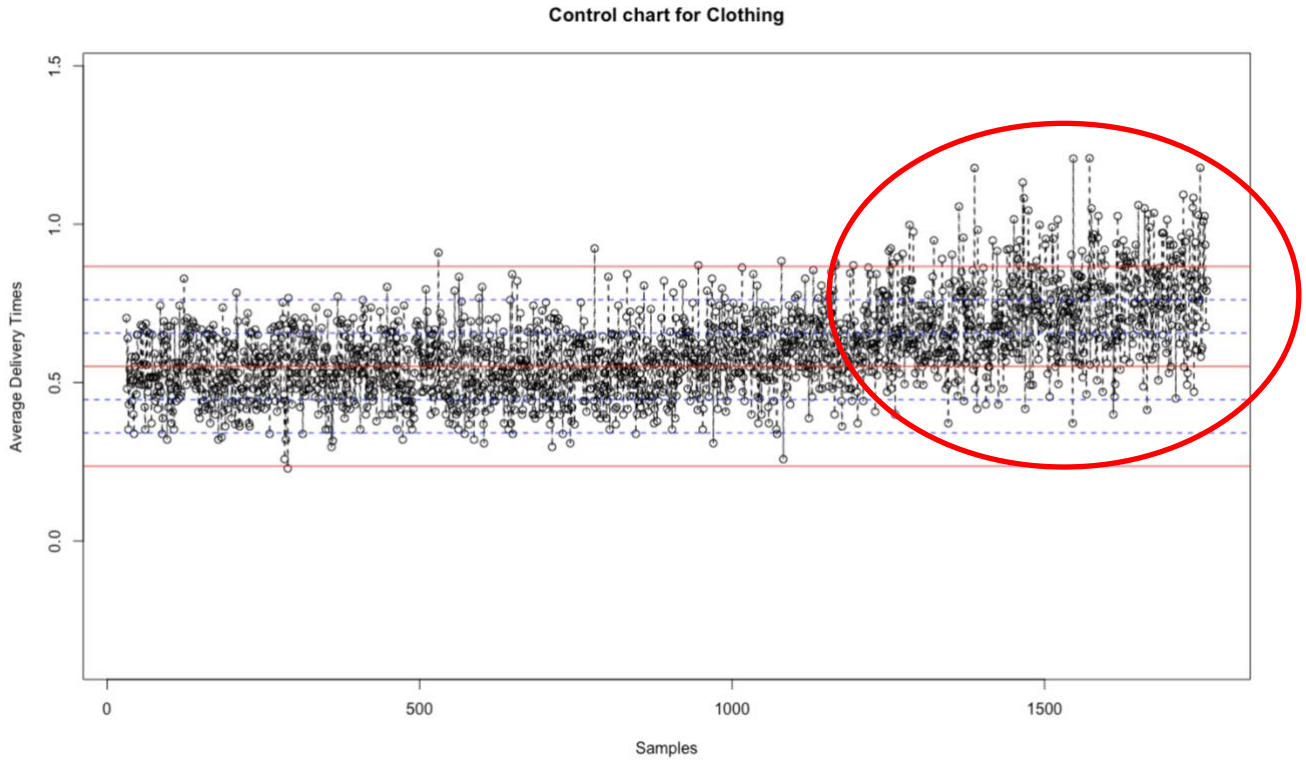


Figure 19:Where clothing delivery process times went out of control

- **Household:**

The household delivery process times s-chart in figure 16 has a slight rise from sample 975 onwards. In table 6, it is shown that 54 out of 1337 samples (4%) are fall outside of the upper or lower control limits. Figure 20 below shows in red where the process becomes out of control. The business needs to stop the process, investigate the reasoning behind the control of the process and find a solution to the delivery time process design for household items.

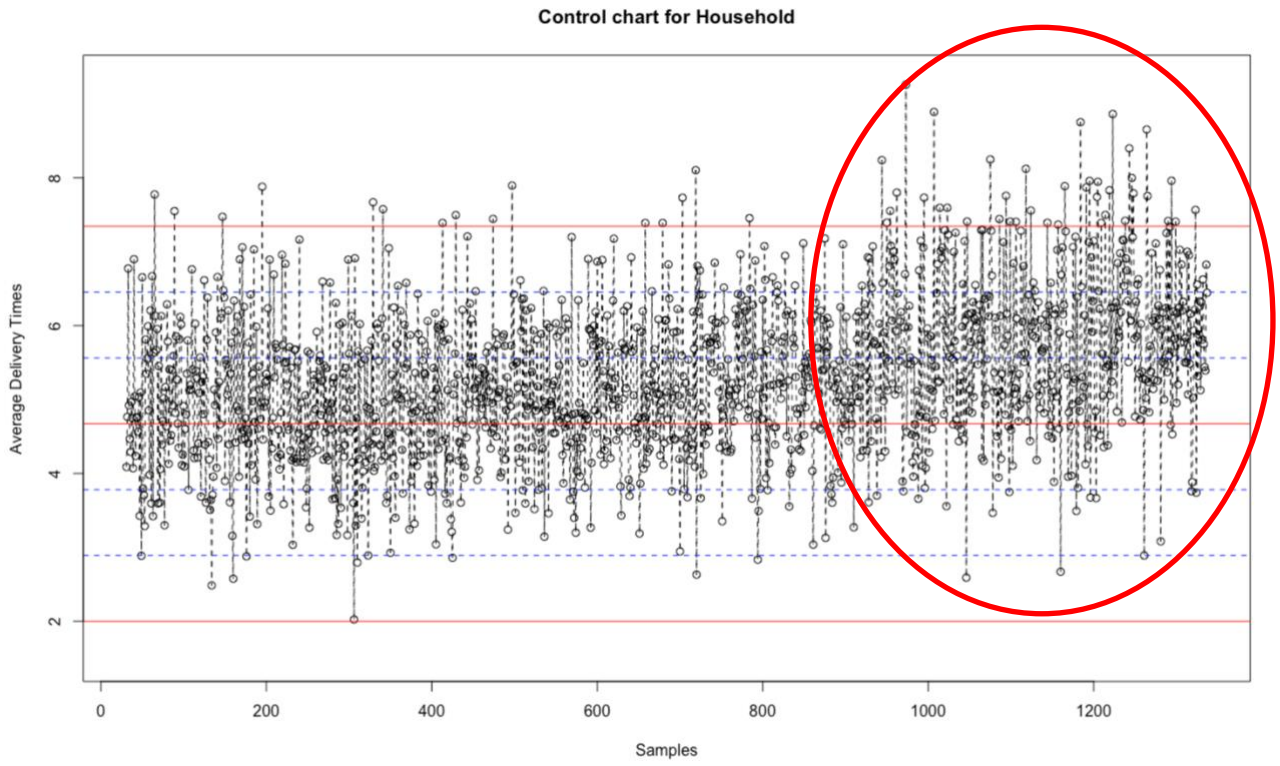


Figure 20: Where household delivery process times went out of control

6.2 Error Analysis

6.2.1 Type I Error Analysis

A type I error is a false positive conclusion to a hypothesis, in terms of this control process, the process is stopped due to an indication that it is out of control but in fact the process was operating sufficiently. (Bhandari, 2021)

Table 6 and 8 are two methods of identifying an out-of-control process, namely method A and method B respectively. Method A identifies samples that are outside the control limits and method B shows most consecutive samples within a specific range of -0.3 and +0.4 sigma-control limits.

The probabilities of making a type I error with method A and B are illustrated in table 9 below.

Table 9: Probabilities of making a Type I Error

Method	Probability of Type I Error
A	0,27%
B	27,33%

There is only a 0.27% chance of making a type I error when using method A and a much higher, 27.33% chance of making a type I error when using method B. This essentially means that when using method B, the process will be stopped 27.33% of the time when there is in fact no reason for

the process to be stopped. Method B will waste money and time for the online business and method A is a much more logical method to use as the risk of making a type I error is significantly lower than the risk for method B.

6.2.2 Type II Error Analysis

Type II error analysis is defined a false negative, in other words, the process will not give indications that it needs to be stopped but in fact it does need to be stopped.
A type II error analysis will be done on the technology class delivery process time. The current mean for the technology delivery process time is 20 hours. If, unknown to the business, the mean were to change to 23 hours, the probability of a type II error occurring is 3,43%. This essentially means that if the mean were to change to 23 hours, 3.43% of the time the process should be stopped but will not.

6.3 Optimal Delivery Cost

It is vital to assess the costs that the late delivery processes incur on the online business to achieve an optimal profit. The business is currently losing R329/item-late-hour in lost sales for technology items if it is delivered slower than 26 hours. It costs the business R2.50/item/hour to reduce the average delivery time by 1 hour. Calculations need to be made to determine how many hours the delivery process time needs to be centered for maximum optimal profit. This will not only achieve minimal costs but also increase customer satisfaction.
The cost function for increasing or decreasing average delivery times is illustrated in figure 21 below.

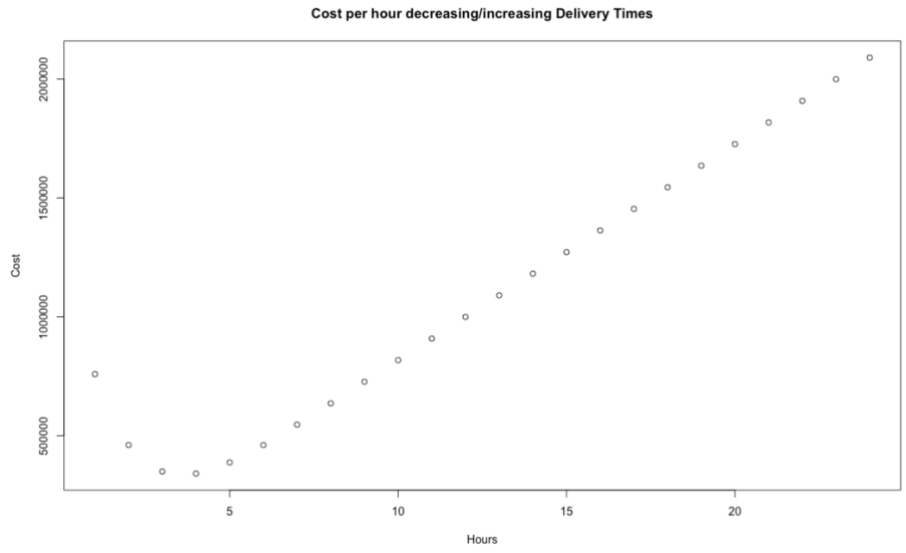


Figure 21: Cost Function for increasing/decreasing delivery times

As seen in the figure above, the average delivery time should hence be decreased by 3 hours to achieve a minimal cost of R 340 870. This means the optimal average delivery time for technology items is 17 hours.

7. MANOVA

Multivariate analysis of variance (MANOVA) is an analysis technique that involves more than one dependent variable simultaneously. MANOVA is able to assess patterns between multiple dependent variables while limiting the error rate.

The relationship between features in the data set will be tested and their level of significance to each other. In this MANOVA, the alpha level is 0.5 and there is a null hypothesis and an alternative hypothesis:

- Null hypothesis: The price and class of an item has a correlation with the delivery time of the item.
- Alternative hypothesis: The price and class of an item does not have any correlation with the delivery time of the item.

Results:

```
              Df  Pillai approx F num Df den Df    Pr(>F)
Class          6 0.81691    20712     12 359942 < 2.2e-16 ***
Residuals 179971
```

The P-Value calculated is significantly smaller than the alpha level of 0.5. This indicates that the null hypothesis deems to be true and there is a correlation between the price and class of an item and the delivery time of it.

8. Reliability of the Service and Products

8.1 Lafrideradora Supplier

Lafrideradora is a supplier of food refrigerator parts for the online business. Management would like to investigate if the requirements are being met and the level of quality it is costing them. The Taguchi Loss Function is used to determine these factors.

Taguchi Loss Function:

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

In problem 6 and problem 7 of chapter 7, we obtain the following calculations:

Initially given: scrap value = 45$ per unit and tolerance of 0.04cm

$$45 = k(0.04)^2$$
$$k = 28125$$

$$\therefore L(x) = 28125(x - 0.04)^2$$

Scrap value changed to 35$ per unit:

$$35 = k(0.04)^2$$
$$k = 21875$$

$$\therefore L(x) = 21875(x - 0.04)^2$$

Process deviation from target can be reduced to 0.27cm:

$$x = 0.027 + 0.04 = 0.067$$
$$\therefore L(0.067) = 21875(0.027)^2 = 15.94688$$

Therefore, the Taguchi loss is at 15.94688 for the Lafrideradora supplier.

8.2 Magnaplex Supplier

Magnaplex is supplier for certain technology items sold on the business's website. The supplier currently has 3 operations that are performed in series in their production process. They use two machines at each stage and feel it is wasteful to use machines as backup in case of failures. Calculations on the reliability of the current process as well as a process where they don't have back upis to be done to see if it is worth it.

- Reliability using single machines:

$Reliability = reliability\ Machine\ A \times reliability\ Machine\ B \times reliability\ Machine\ C$
$Reliability = 0.85 \times 0.92 \times 0.9 = 0.7038$

- Reliability using parallel machines:

$Reliability = 1 - (1 - reliability\ Machine\ A) \times (1 - reliability\ Machine\ B) \times (1 - reliability\ Machine\ C)$
$Reliability = 1 - (1 - 0.85) \times (1 - 0.92) \times (1 - 0.9) = 0.9988$

Magnaplex needs to ensure that they are using their machines in parallel as it is 99.88% reliable. When using the machines in series it is only 70.38% which will not be sufficient for the company.

8.3 Delivery Process Calculations

1560 days of data from the past were used for this investigation. This research showed that there is a 73.47% possibility that there will only be 20 trucks and 21 drivers available, which would mean that the delivery process demands would not be met. Accordingly, 268 days a year are reliable for delivery, leaving 97 days for unreliable delivery. This is not ideal because it affects customer loyalty and happiness, thus it was decided to look into how reliable the delivery process would be if an additional truck were added while keeping the current number of drivers. The reliability of the delivery process increases to 83.16% by adding one more truck, bringing the total to 21. This indicates that 303 days out of the year, the delivery method is trustworthy.

9. Conclusion

The online business was statistically analyzed in many aspects. It has been made clear that there are many areas where the business needs to be investigated and redesigned to fix issues that were picked up in the process of analysis. Delivery times of different products is a vital problem that the company needs to put time, effort, and investment into. General trends observed in the descriptive statistics are not all entirely the business's fault as it must be taken into account that generations of the human population have different desires and behaviors. There are however tasks they can achieve to avoid these trends and increase their numbers in areas needed. The business can now undergo the investigative phase in order to optimize their business.

10. References

Pearce, H. (2020, April 6). *Guide To Data Wrangling: What It Is And Who Should Do It*. Retrieved from bright data: https://brightdata.com/blog/proxy-101/guide-data-wrangling?kw=&cpn=18323184899&cam=aw_all_products-base-search_dsa_blog_base-kw_en-desktop_blog-proxy-101__621567933077&cq_src=google_ads&cq_cmp=18323184899&cq_term=&cq_plac=&cq_net=g&cq_plt=gp&utm_term=&utm

Jain, D. (2020, December 11). *Feature Handling: Categorical and Numerical*. Retrieved from Towards Data Science: https://towardsdatascience.com/feature-handling-3f14c12ecbb8

Hayes, A. (2022, August 1). *Descriptive Statistics: Definition, Overview, Types, Example*. Retrieved from Investopedia: https://www.investopedia.com/terms/d/descriptive_statistics.asp

Statistics > Box Plots. (2010). Retrieved from NetMBA Business Knowledge Center: http://www.netmba.com/statistics/plot/box/

Process Capability Analysis Cp, Cpk, Pp, Ppk - A Guide. (n.d.). Retrieved from 1Factory: https://www.1factory.com/quality-academy/guide-to-process-capability-analysis-cp-cpk-pp-ppk.html

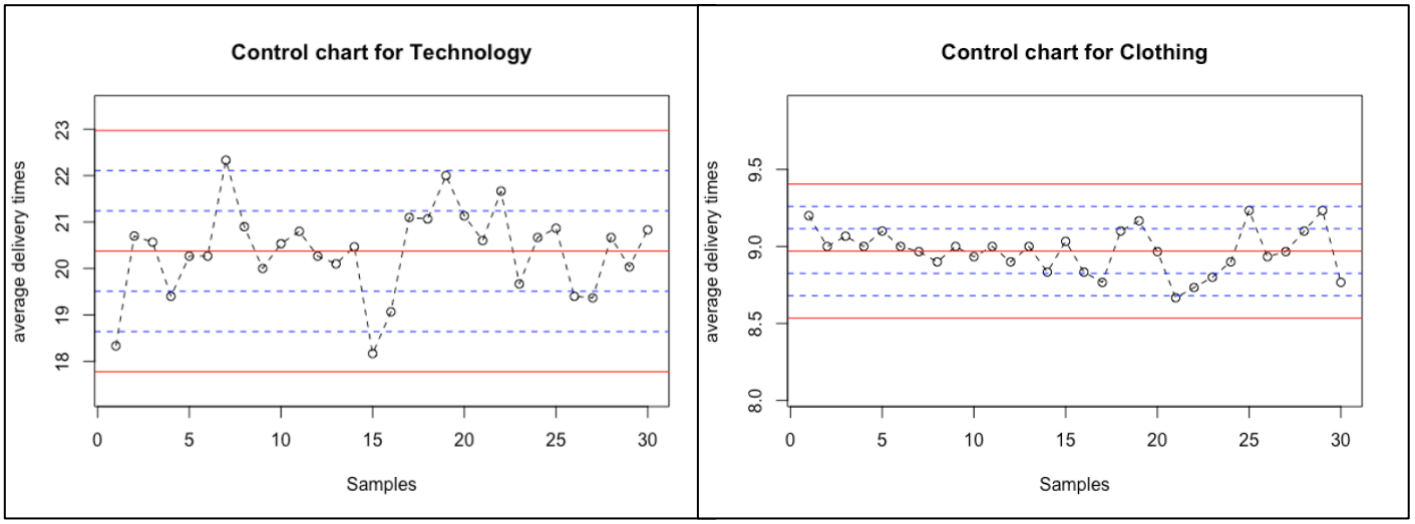
Graham, J., & Clearly, M. (n.d.). *Practical Tools for Continuous Improvement.* PQ Systems.

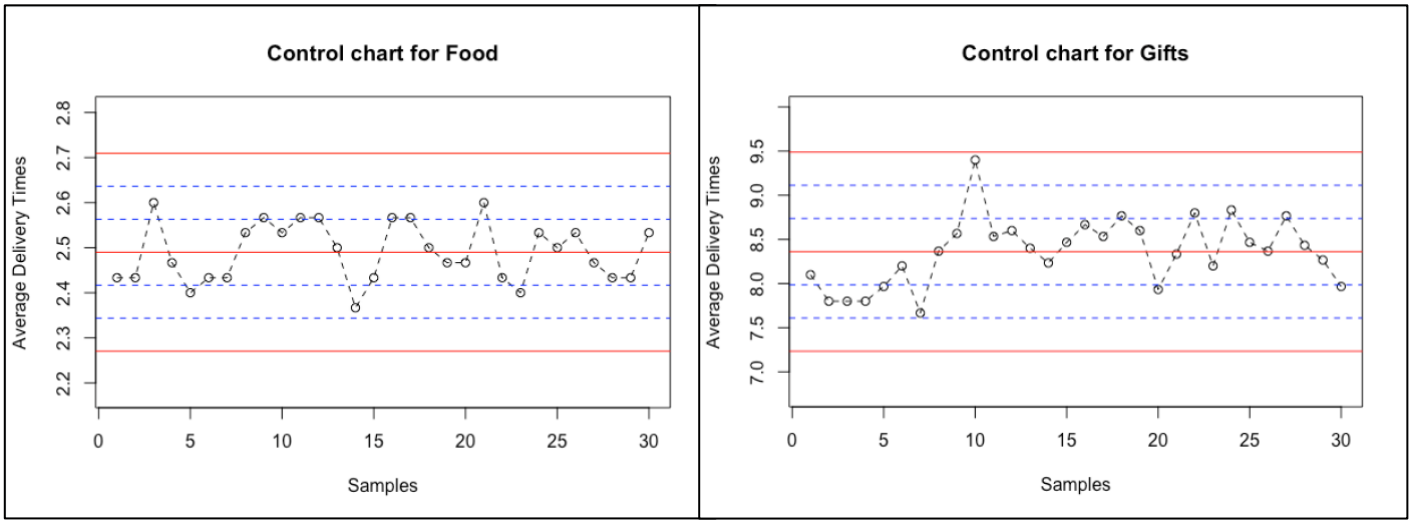
Montgomery, D. (2012). *Introduction to Statistical Quality Control.* John Wiley & Sons.

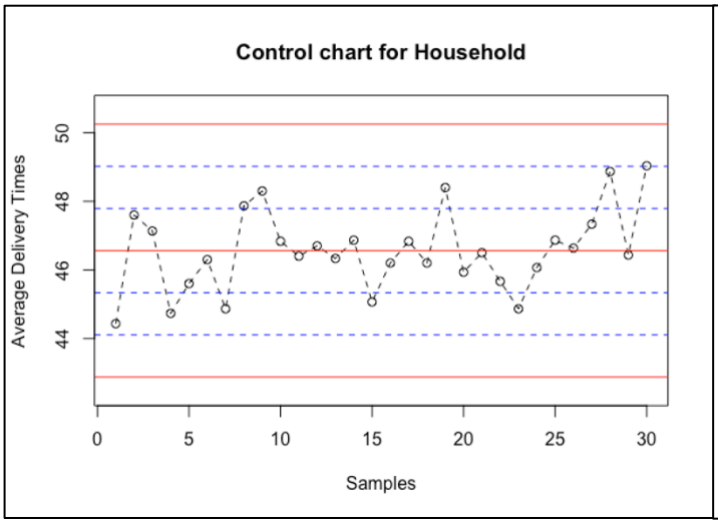
Hessing, T. (n.d.). *Statistical Process Control (SPC)*. Retrieved from Six Sigma Study Guide: https://sixsigmastudyguide.com/statistical-process-control-spc/

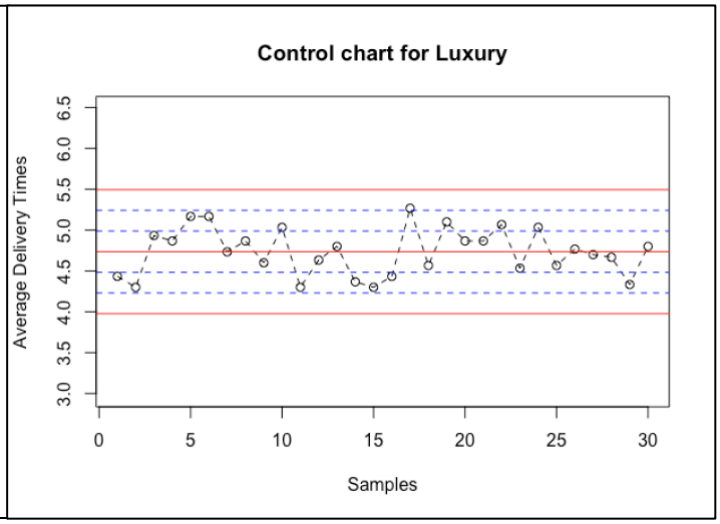
Bhandari, P. (2021, January 18). *Type I & Type II Errors | Differences, Examples, Visualizations*. Retrieved from Scribbr: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

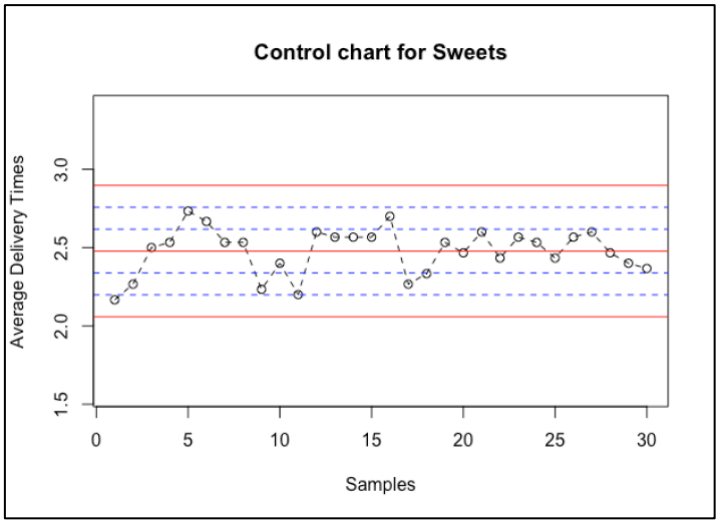
11. Appendix A – X-Bar Control Charts






Control chart for Household
Average Delivery Times
44
46
48
50
0
5
10
15
20
25
30
Samples


Control chart for Luxury
Average Delivery Times
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
0
5
10
15
20
25
30
Samples


Control chart for Sweets
Average Delivery Times
1.5
2.0
2.5
3.0
0
5
10
15
20
25
30
Samples