Department of Industrial Engineering

Quality Assurance 344

ECSA GA4 Project

Y De Raay (23618221)

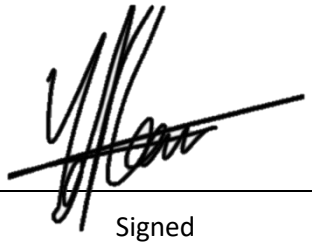
10 October 2022

© Stellenbosch University

Plagiarism Declaration

I, Y de Raay (23618221), declare that:

- I have read and understand the Stellenbosch University Policy on Plagiarism and the definitions of plagiarism and self-plagiarism contained in the Policy [Plagiarism: The use of the ideas or material of others without acknowledgement, or the re-use of one's own previously evaluated or published material without acknowledgement or indication thereof (self-plagiarism or text recycling)].
- I also understand that direct translations are plagiarism, unless accompanied by an appropriate acknowledgement of the source. I also know that verbatim copy that has not been explicitly indicated as such, is plagiarism.
- I know that plagiarism is a punishable offence and may be referred to the University's Central Disciplinary Committee (CDC) who has the authority to expel me for such an offence.
- I know that plagiarism is harmful for the academic environment and that it has a negative impact on any profession.
- Accordingly all quotations and contributions from any source whatsoever (including the internet) have been cited fully (acknowledged); further, all verbatim copies have been expressly indicated as such (e.g. through quotation marks) and the sources are cited fully.
- Except where a source has been cited, the work contained in this assignment is my own work and that I have not previously (in its entirety or in part) submitted it for grading in this module/assignment or another module/assignment.



Signed

10 October 2022

Date

Abstract

Data analysis helps a business identify risks and anomalies and discover complex patterns that allows the business to pick up future trends which provides meaningful insights into your business that can ultimately improve the overall profitability

This report aims to find key issues in the business and solve them through data wrangling of the salesData2022 data set to eliminate any noise such as missing values and negative values in order to generate insightful statistical models that will help the business solve high impact problems.

Once the models have been generated, analysed and interpreted, process capability measures will be calculated to quantify the capability of the process to meet the design specifications so the business can build a product to meet customer expectations. Statistical Process Control will then be performed which is a methodology for monitoring a process to identify special causesof variation within a process and signal the need to take corrective action (Evans & Lindsay, 2018). SPC thus offers a way for the online business to show off its capacity for excellence. Once statistical process control has successfully been carried out, MANOVA analysis will be done to compare one or more dependent values across two or more groups. Lastly reliability of the service, products and the delivery process will be calculated using the Taguchi loss function and various reliability calculations. Overall, after analysing the sales data, value will be added to the business which will create confidence and assurance ultimately paving a way for exponential growth and profitability.

Table of Contents

List of Figures

List of Tables

Nomenclature

Symbols	
C_p	Process Capability
C_{p_u}	Upper One-Sided-Index
C_{p_l}	Lower One-Sided-Index
$L(x)$	Monetary loss (assumed to increase quadratically)
k	Scaled distance of the mean from the target value
T	Target specification assumed to be centred between USL and LSL
Greek symbols	
σ	Standard Deviation of the Process
μ	Actual Process Mean
Abbreviation	
LSL	Lower Specification Limit
USL	Upper Specification Limit

1 Introduction

The goal of this report is to use data analytics tools to solve core business problems and add value by providing actionable recommendations. Through the use of tools and techniques we are able to turn raw sales data into meaningful business insights that will aid in maximizing profits, minimizing variations and improving the overall market strategy by segmenting customers based on age and tailoring product classes to them in order to predict future customer preferences based on past sales. Markets are becoming more competitive and demanding decisions faster. If the business wants to remain a competitor in the market, insights on consumer behaviour and organizational processes need to be developed faster which can be done by analysing descriptive data from the past and using it to predict future trends. The techniques that will be used in this report are statistical models and control charts that will be generated using R studios.

2 Data Wrangling

2.1 Getting to know the dataset parameters

Before the sales data can be analysed it is important to know the form of your data in order to effectively interpret and analyse it. The data can be identified as either numerical, categorical, descriptive, or target features. Data types that are stored and identified based on names or labels assigned to them are referred to as categorical features. This means each individual sales instance is assigned to a particular group or nominal category on the basis of some qualitive property captured by this feature (Kelleher, et al., 2015). Numerical data refers to the data that is in the form of numbers and can either be discrete that assumes a limited number of different responses or continuous which is an unlimited number of responses. Descriptive features describe or summarize the data in a constructive way and are used in analysis as well as predicting the target feature. The target feature is the feature in your data set that you want to focus on.

The descriptive features used to describe the sales data instances:

- **ID:** The ID displays the sales number. A distinct sales number is assigned to each of the 7000 sales instances, making it possible to find or even look up a specific transaction. This qualifies as a numerical feature.
- **AGE:** When a customer completes an online transaction, the AGE feature displays the customer's age.
- **CLASS:** This feature helps the business know what sort of goods a consumer purchased from the online shop. The class feature comprises products in the following categories: clothing, luxury, gifts, food, household, technology, and sweets. This feature may be used to categorize all occurrences of valid sales data into seven groups based on the type of item that was purchased, making it a categorical feature.

- **PRICE:** The PRICE feature indicates the total monetary value of each specific sales instance. It is thus the total amount of money that the online store made for that specific sales instance. This is a numerical feature since the price value of every sales instance will differ and is expressed as a numerical value.
- **YEAR:** The YEAR feature indicates the year in which the product was bought.
- **MONTH:** The MONTH feature indicates the month in which the product was bought.
- **DAY:** The DAY feature indicates the on which day of the month the product was bought.

Target features used to analyse the sales data instances:

- **DELIVERY TIME:** This feature shows how long (in days) it will take to get a product after purchasing it from an online retailer. Since the delivery time of the online business must be agreeable to clients in order to guarantee that the firm achieves quality from their perspective, this may be regarded a target feature. Thus, this is one of the characteristics that the statistical analysis should emphasize. Given that it is given as a number value, delivery time is a numerical feature.
- **WHY BOUGHT:** The WHY BOUGHT section describes through which marketing channel the customer discovered the online website before making a purchase. This qualifies as a target feature since the business may utilize the findings of the statistical analysis to create a more effective marketing strategy. This feature is categorical because it can be used to divide all occurrences of valid sales data into six groups according to how the client discovered the online domain. The six categories are website, random, recommended, spam, email, and browsing. The vast quantity of accessible data must be pre-processed before analysis can begin once the distinct features in the dataset have been recognized and comprehended.

2.2 Defining Data Wrangling

The salesData2022 provided is raw data and therefore in order to get useful information in the data to improve the organisations processes the data needs to be cleaned up and converted to a form that makes information accessible. Data wrangling involves processing the data in various ways for the purpose of analysing the data in order to get useful information from it.

2.3 Performing Data Wrangling

SalesTable2022's data was examined and divided into valid and invalid data. If a row in the original data set included any NA values, any negative values, or any values larger than the highest priced product, the data was rendered invalid. The invalid data frame was then constructed from these instances. After the invalid data was removed the valid data frame was constructed using the remaining data.

2.3.1 Invalid data

The table below shows the invalid data that was removed from the sales data set as it contains missing values and negative values in the price feature.

	IncompleteIndex	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
13	13	98765	64288	25	Clothing	*NA*	2021	1	24	8.5	Browsing
17	17	177777	68698	30	Food	*NA*	2023	8	14	2.5	Recommended
10	10	65432	51904	31	Gifts	*NA*	2027	7	24	14.5	Recommended
12	12	87654	40983	33	Food	*NA*	2024	8	27	2.0	Recommended
8	8	54321	62209	34	Clothing	*NA*	2021	3	24	9.5	Recommended
16	16	166666	60188	37	Technology	*NA*	2024	10	9	21.5	Website
3	3	19541	71169	42	Technology	*NA*	2025	1	19	20.5	Recommended
2	2	16321	81959	43	Technology	*NA*	2029	9	6	22.0	Recommended
6	6	34567	18748	48	Clothing	*NA*	2021	4	9	8.0	Recommended
9	9	56789	63849	51	Gifts	*NA*	2024	5	3	10.5	Website
15	15	155555	33583	56	Gifts	*NA*	2022	12	9	10.0	Recommended
7	7	45678	89095	65	Sweets	*NA*	2029	11	6	2.0	Recommended
14	14	144444	70761	70	Food	*NA*	2027	9	28	2.5	Recommended
5	5	23456	88622	71	Food	*NA*	2027	4	18	2.5	Random
11	11	76543	79732	71	Food	*NA*	2028	9	24	2.5	Recommended
4	4	19999	67228	89	Gifts	*NA*	2026	2	4	15.0	Recommended
1	1	12345	18973	93	Gifts	*NA*	2026	6	11	15.5	Website

Figure 1: Figure showing invalid data features

In the above data frame, it is evident that the price feature is the only column with invalid features. The original data set did not include any negative values therefore only the missing values were removed to create the valid data frame. The maximum price for a product in the valid data frame is R116619.

2.3.2 Valid data

The table below shows a snippet of the cleaned, processed data that can now be used to generate statistical models for interpretation of the different features

	ValidIndex	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	1	1	19966	54	Sweets	246.21	2021	7	3	1.5	Recommended
2	2	2	34006	36	Household	1708.21	2026	4	1	58.5	Website
3	3	3	62566	41	Gifts	4050.53	2027	8	10	15.5	Recommended
4	4	4	70731	48	Technology	41843.21	2029	10	22	27.0	Recommended
5	5	5	92178	76	Household	19215.01	2027	11	26	61.5	Recommended
6	6	6	50586	78	Gifts	4929.82	2027	4	24	14.5	Random
7	7	7	73419	35	Luxury	108953.53	2029	11	13	4.0	Recommended
8	8	8	32624	58	Sweets	389.62	2025	7	2	2.0	Recommended
9	9	9	51401	82	Gifts	3312.11	2025	12	18	12.0	Recommended
10	10	10	96430	24	Sweets	176.52	2027	11	4	3.0	Recommended
11	11	11	87530	33	Technology	8515.63	2026	7	15	21.0	Browsing
12	12	12	14607	64	Gifts	3538.66	2026	5	13	13.5	Recommended
13	13	13	24299	52	Technology	27641.97	2024	5	29	17.0	Browsing
14	14	14	77795	92	Food	556.83	2025	6	3	3.0	Random
15	15	15	62567	73	Clothing	347.99	2024	3	29	8.5	Website
16	16	16	14839	47	Technology	54650.41	2027	12	30	18.5	Recommended
17	17	17	96208	44	Technology	14739.09	2028	3	17	13.0	Recommended

Figure 2: Figure showing valid data features

The valid data frame constructed consists of 179983 instances but a snippet of the data is displayed above in order to present the ordered, processed data that can now be used to create descriptive graphs that will be used to analyse the business performance and areas that can be improved.

3 Descriptive Statistics

3.1 Definition

Descriptive statistics refers to historical data that can be processed, visualized and interpreted to provide guidance and governance to the business in order to help in future decision making.

3.1.1 Five-point summary

The five point summary summarizes each feature into the minimum, 1st quartile, median, 3rd quartile, maximum and the mean was added as it is necessary for interpretation purposes as it gives the business an indication of the average value per feature.

Table 1: Table showing the five point summary for age

AGE	
Minimum	18.0
1st Quartile	38.0
Median	53.0
Mean	54.5
3rd Quartile	70.0
Maximum	108.0

Table 2: Table showing the five point summary for price

PRICE	
Minimum	-588.80
1st Quartile	482.31
Median	2259.63
Mean	12293.74
3rd Quartile	15270.74
Maximum	116618.97

Table 3: Table showing the five point summary for delivery time

DELIVERY TIME	
Minimum	0.5
1st Quartile	3.0
Median	10.0
Mean	14.5
3rd Quartile	18.5
Maximum	75.0

3.2 Analysis

3.2.1 Numerical features

Numerical data are values that can be measured and organized logically. Their characteristics are numbers that describe an object's various properties(Javaid, 2022). Therefore, histograms of the delivery time frequencies and price frequencies were generated to help the business further understand these features.

3.2.1.1 Delivery time frequencies

The histogram of delivery time frequencies in days is shown below:

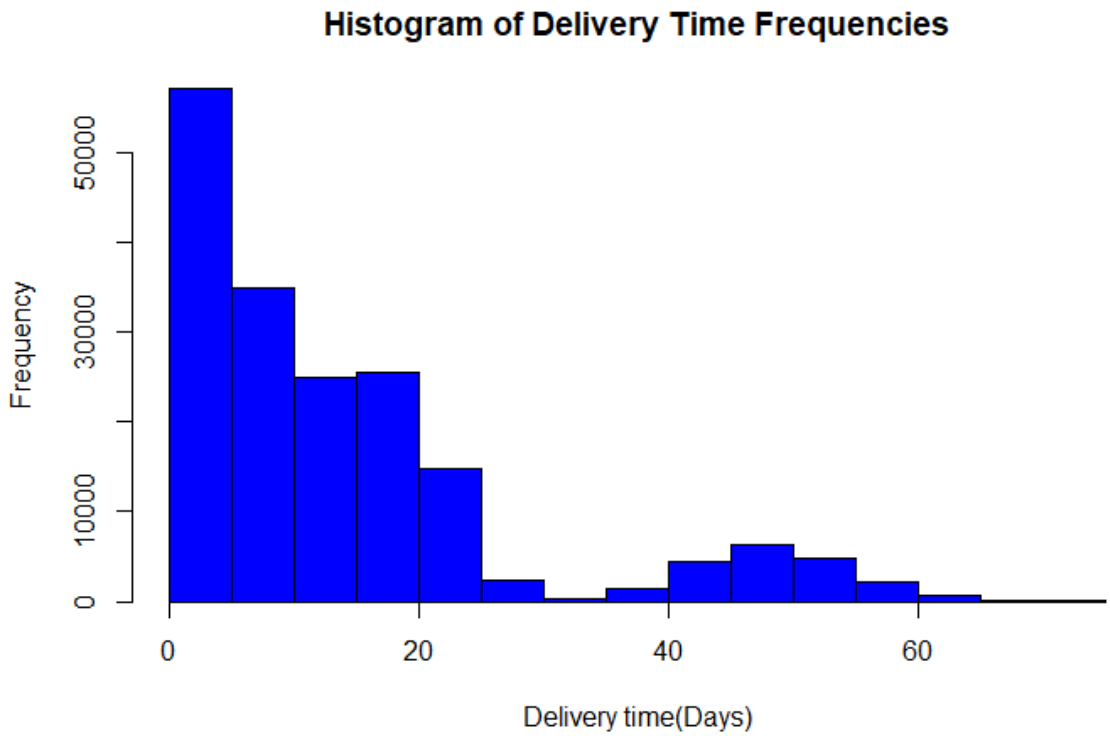


Figure 3: Histogram of the delivery time and the frequency of different delivery times.

***Analysis*:**

In the above figure it is evident that the data is skewed to the left as majority of deliveries are made within the first quartile. From the five-point summary generated above it is clear that there are outliers in the delivery time data has the mean is 14.5 days, maximum of 75 days and a minimum of half a day. These outliers are the values that greatly differ from the mean of the data. It is also observed that majority of deliveries are made within 20 days which would depend on the product ordered, the location of the customer and the availability of the product. In order to increase customer satisfaction the business should aim to minimize their delivery times which could be done by ordering larger lot sizes to ensure availability of products, investing in more delivery vehicles or taking a deeper look at the inbound and outbound warehouse in order to create a more efficient delivery process.

3.2.1.2 Price frequencies

The histogram of price frequencies that will be further analysed and interpreted is shown below:

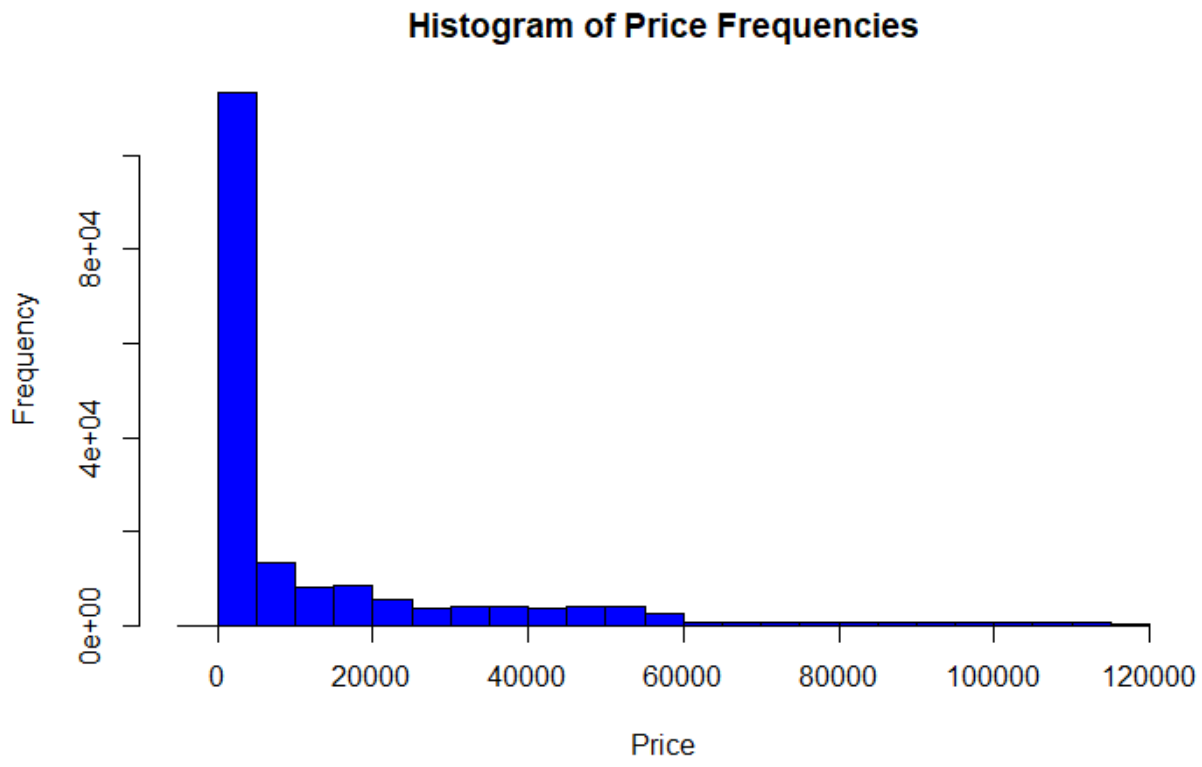


Figure 4: Histogram of the Price Frequencies

Interpretation:

It is clear that the above graph is skewed to the right as majority of product prices range between 0-5000 rand which makes sense as lower priced items generally sell faster and more frequently.

3.2.2 Categorical features

Categorical data is simply information aggregated into groups rather than being in numeric formats (Zuccarelli, 2020). Histograms of Customer exposure marketing methods and How frequent different product classes are bought have been generated below to aid the company in developing a strategic marketing plan by honing into fewer marketing channels in a more effective way instead of putting money, time and energy into ineffective marketing methods.

3.2.2.1 Customer exposure marketing methods

The reasons for why consumers bought products is given below in the form of a bar graph which is helpful in developing a good marketing strategy and will be further interpreted below:

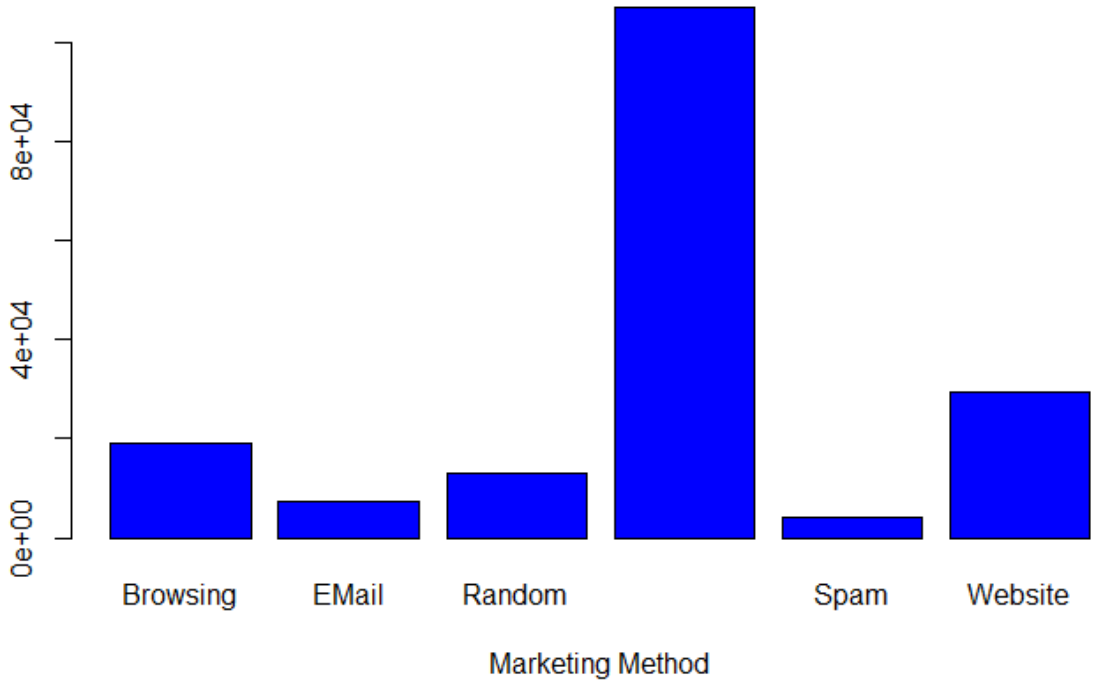


Figure 5: Figure showing a bar graph of the customer exposure marketing method

***Interpretation*:**

In the above figure it is evident that majority of customer exposure has been through recommendations which could be an indication of customer satisfaction and loyalty. Word of mouth advertising is free advertising triggered by customer experiences therefore in order to increase the customer exposure through this marketing strategy the business could implement promotions, ensure good customer service and increase customer loyalty by introducing loyalty cards.

3.2.2.2 How frequent different types of items are bought

The bar graph of how frequently different product classes are bought is shown below:

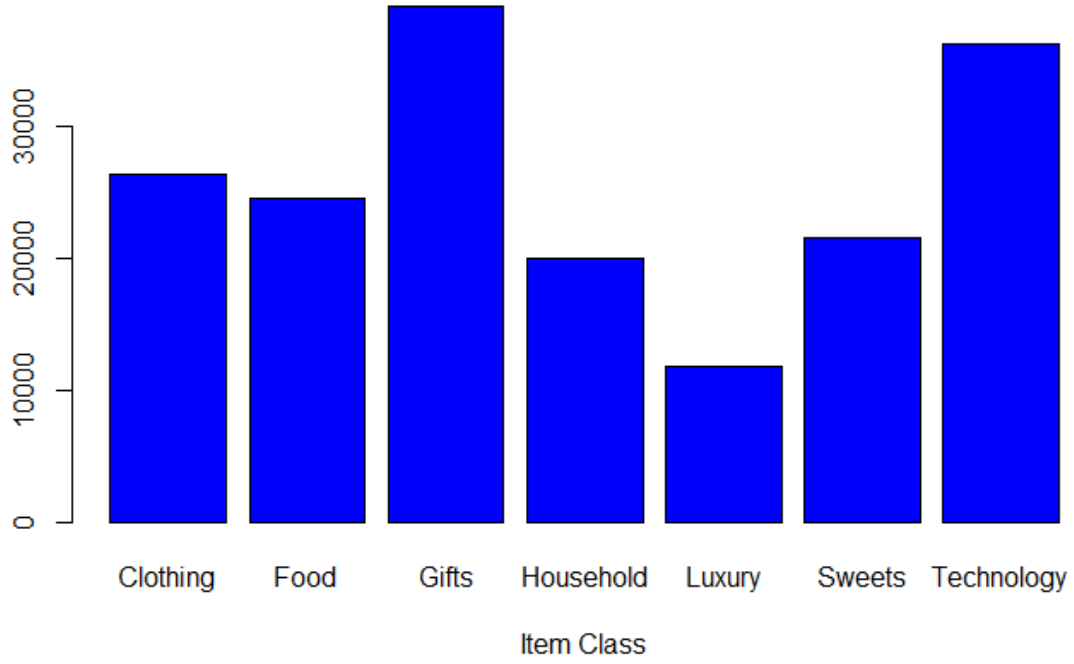


Figure 6: Figure showing the frequency that different items are bought

Interpretation:

In the above model it is clear that gifts are the most frequently bought product with technology products bought at a slightly lower frequency and luxury items being the least frequently bought product which makes sense due to the higher price of this product class. From this information the business can put marketing strategies in place to increase the number of sales of luxury items and keep the frequency of gist and technology sales at a high rate. The frequency of luxury product sales could be increased by targeting a higher income level market through advertising quality rather than price as high-income buyers are more attracted to the quality and design of the product rather than the price. In order to maintain the high frequency of the remaining product classes it is important for the business to stay consistent in the customer service they are currently providing.

3.3 Visualisation and Interpretation by Class

3.3.1 Price per Class

The box plot for the price distribution per class is shown below:

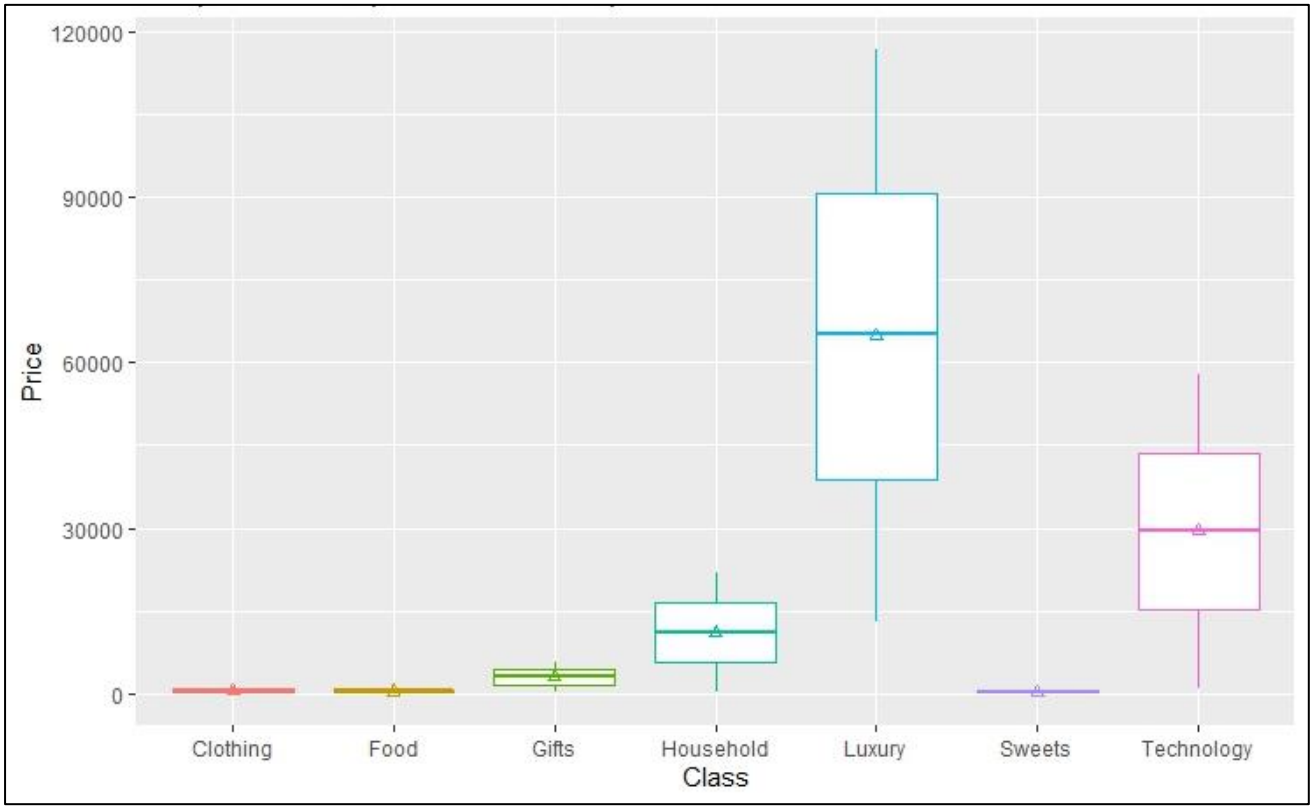


Figure 7: Box plots for the price distribution per class

<u>Interpretation</u>

In the above figure it can be seen that the luxury class follows a normal distribution and has the highest selling prices ranging from R37000-R90000 which is expected. The technology class also follows a normal distribution and prices range from R15000-R45000 which can also be expected as technology products are generally in a higher price range compared to the other product classes. It is evident that clothing, food, gifts and sweets all fall in the lower price range.

The figure below shows the different histograms of price distributions per class:

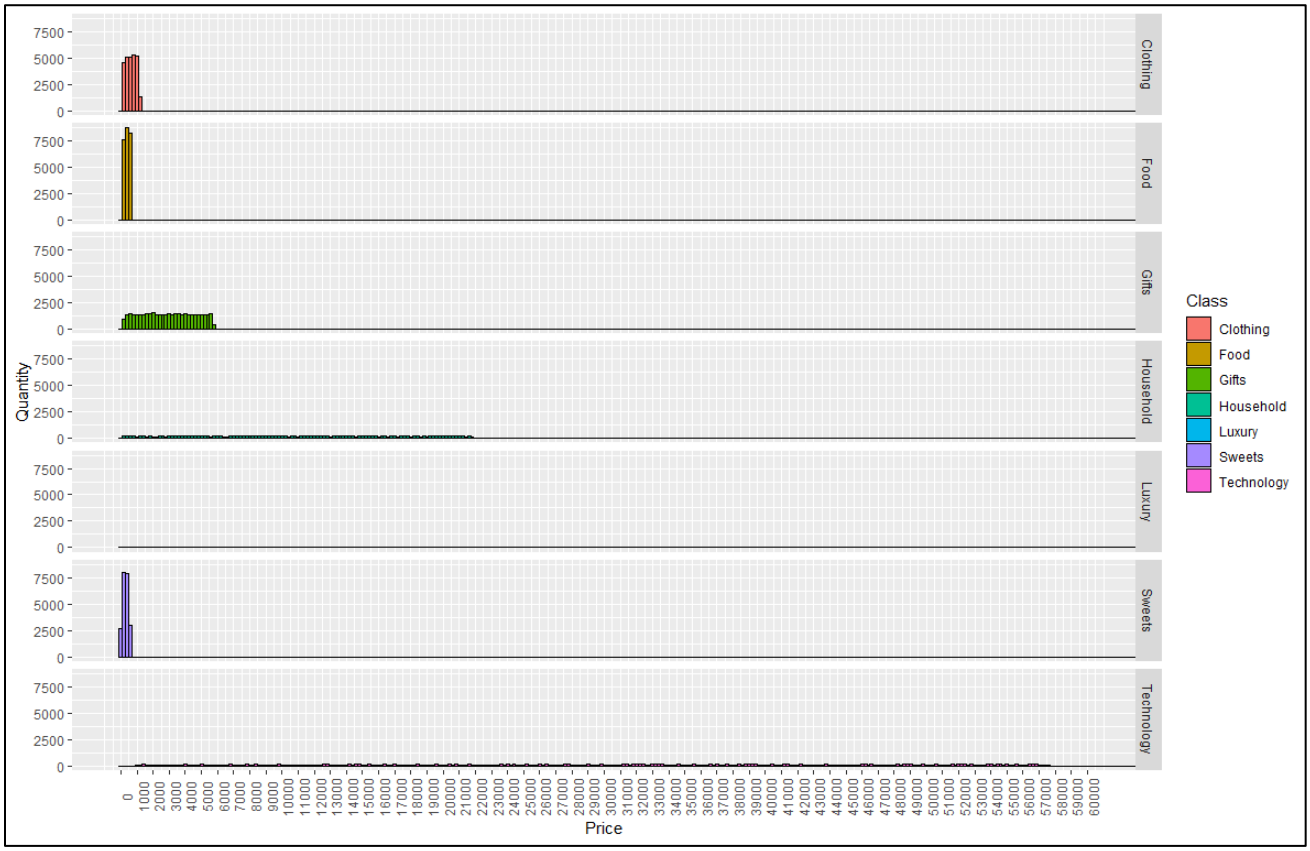


Figure 8: Histograms showing the price distribution per class

<u>Interpretation</u>:

The above figure is a box plot representing the different price distributions for the different product classes bought which is very insightful as it can help the business get an estimation of the optimal prices per class. By observing Figure five it is evident that the different product classes have different price ranges. Figure 8 allows the business to compare the price distribution per product class and the quantity of sales per class. It is observed that the price distribution for clothing is skewed to the left and has a high quantity of sales at fairly low prices. In the price distribution for food, the data is also skewed to the left with a large number of sales at even lower prices, which is expected as food is a necessity and frequently bought. Gifts follow a wide normal distribution as they vary in price and sales quantity stays fairly similar over different prices. Household, technology and luxury products are all in the higher price region and therefore it is evident in the figure that the price varies over a wide range with a low sales quantity over all prices which is expected. Lastly sweets follow a normal distribution in the lower price range. The business can use this information to find a niche target market for each product class to help define their market strategy.

3.3.2 Age per Class

The box plot for the age distributions per class is shown below:

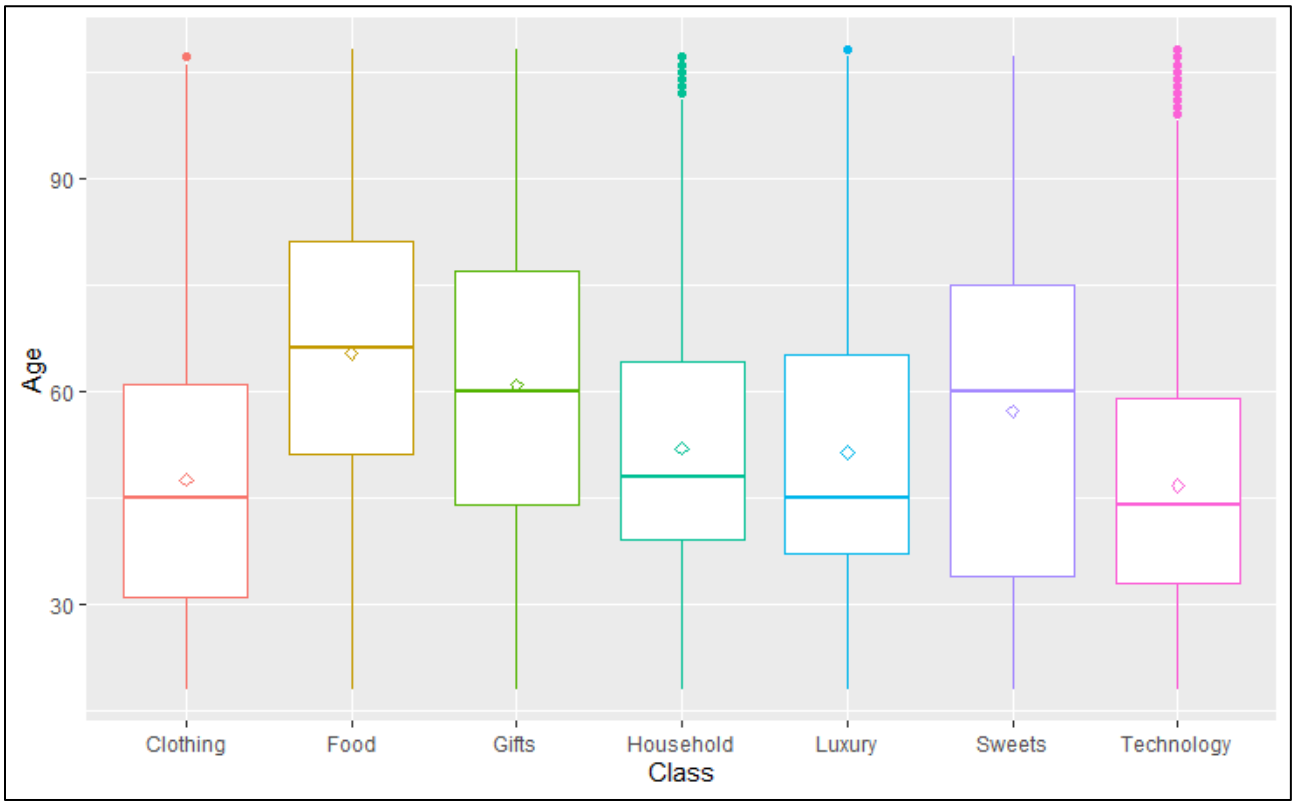


Figure 9: Figure showing age distribution per class

The figure below shows the histograms of different age distributions per class and will be interpreted below:

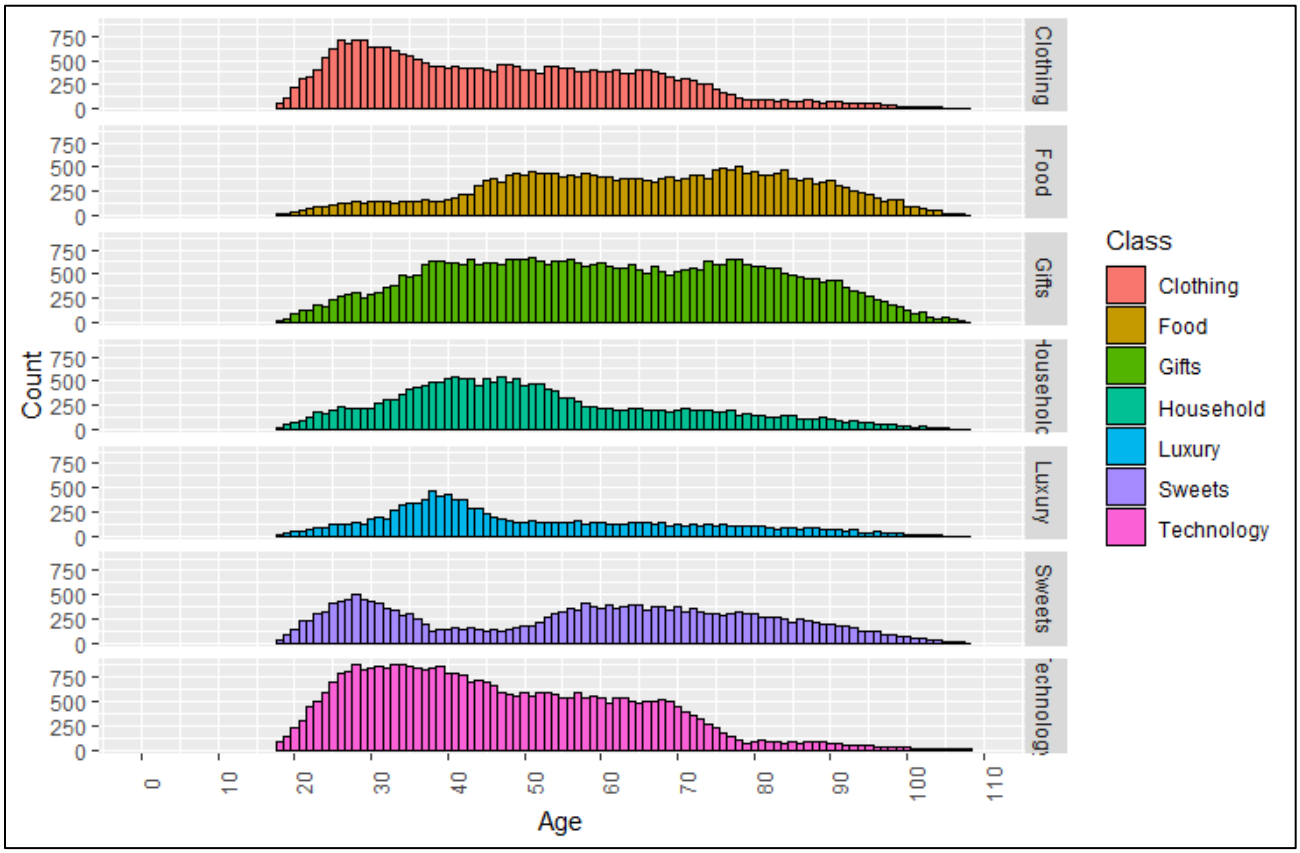


Figure 10: Figure showing the histograms of age distribution per class

<u>Interpretation</u>:

The above figure is a box plot representing the different age distribution for the different class of products bought which is very insightful as it can help the business develop their market strategy by tailoring specific offers to different customer groups.

The technology class follows a distribution that is skewed to the right which indicates that younger customers buy more technology products than the older age group which is expected as young people are more technologically inclined.

Customers who purchase food online have an age distribution that is slightly skewed to the left but generally approximates a normal distribution. A relatively small percentage of the clientele is under the age of 40. Given that food is a necessity, this is unexpected. One may assume that since the clients in this age group are still youthful and active, they prefer to purchase meals offline rather than online. However, it is still important to look at the causes of the tiny number of food purchasers under the age of 40 in order to find ways to increase the number of online sales in this product class.

The age breakdown of clothing shoppers is biased to the right. This suggests that younger customers are using the website more frequently and purchasing clothing. Investigating the causes of the skewed distribution is necessary as it could suggest that there is a gap in variety of clothes aimed at the older age groups. By investigating this issue, the business could expand its customer base in the clothing class and increase the number of sales which intern will increase their profit margin A by simply stocking a larger range of clothing.

The age distribution of customers buying gifts is a uniform distribution. This indicates that clients of all ages use the online site to buy gifts which is expected as buying gifts is not bias to any age group.

The age distribution of customers buying household items is skewed to the right, indicating that household items are bought more frequently by younger customers rather than older customers. Household items are bought mainly by customers between the ages of 30 and 60. This could be due to the fact that younger people in their thirties are generally at the stage in their life where they purchase their first house and start gaining financial stability therefore are in search of household products to accompany their house purchase.

The age distribution of customers buying luxury items is also skewed to the right, with high frequencies between the age of 30 and 50. This is expected as this is the age that most people are at the peak of their career and financially healthier than ever. Therefore, this gives the business insights into how to market their luxury products in the most effective way to reach their target market.

The age distribution of customers buying sweets is approximating a multi modal distribution, with two peeks. The first peek in ages buying sweets is between 25 and 35, and the second peak between 60 and 85. There is a large decrease in the sales of sweets occur 20 between the ages of 35 and 60 which is expected as younger people are generally more weight conscious and older people are generally more focussed on their health. The sweets age distribution also drastically differs from the food age distribution. This is not expected, and the company should investigate why these two distributions differ so drastically.

3.3.3 Delivery per class

The box plots of delivery time distributions per class are shown below:

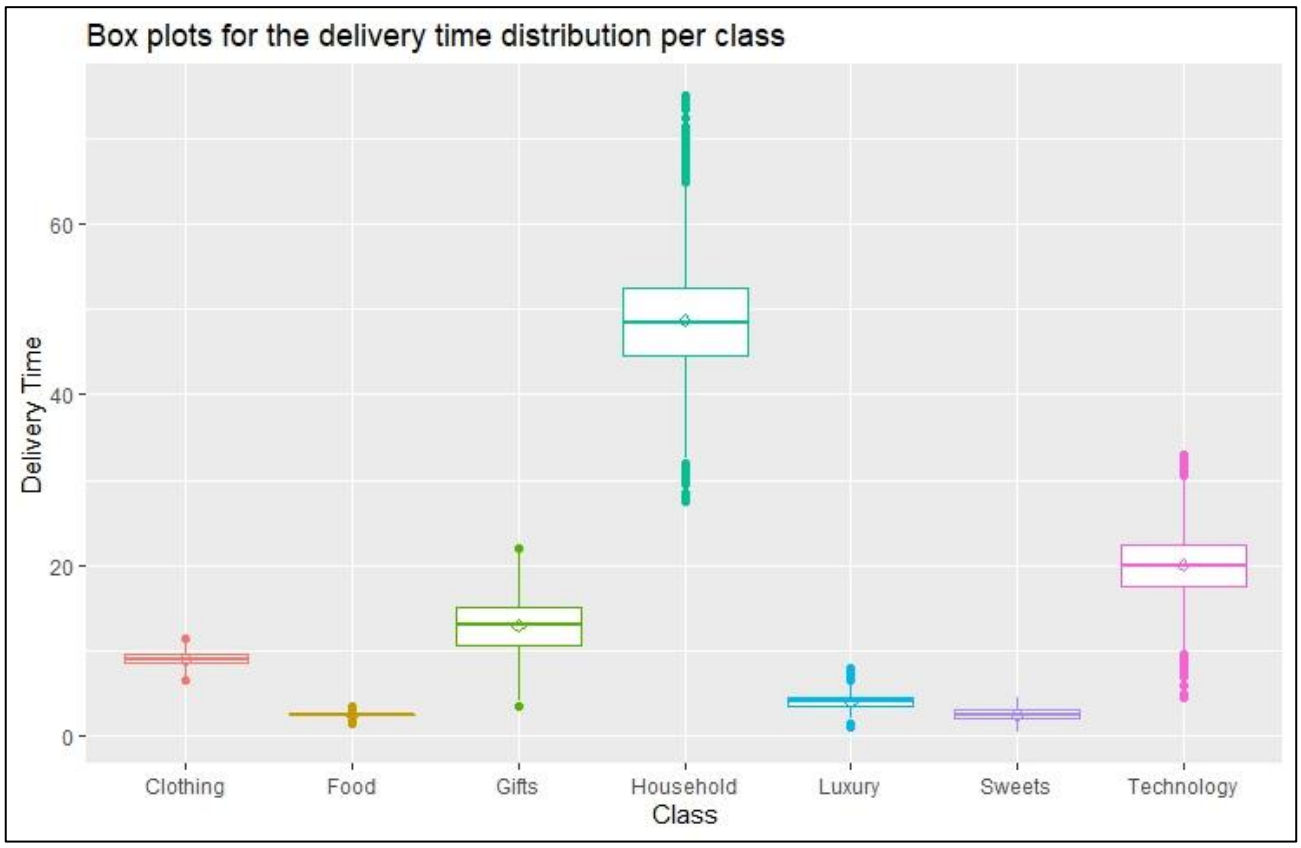


Figure 11: Figure showing the box plots for the delivery time per class

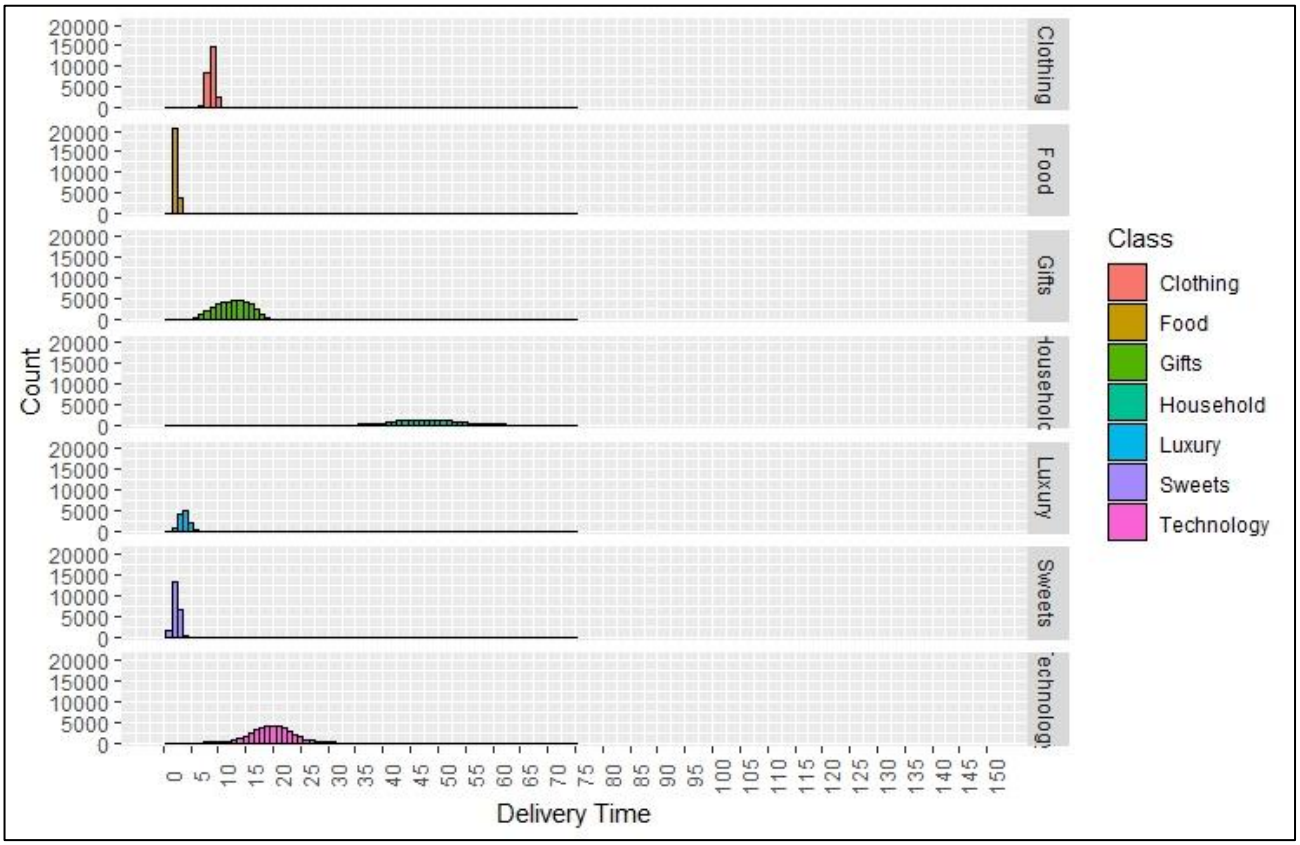


Figure 12: Figure showing the histograms of the delivery time per class

<u>Interpretation:</u>

From analysing the above box plot and histograms of the delivery time distributions per class that food products distribution is skewed to the left and ranges between 0-3 delivery days which is expected as food is a necessity and therefore almost always in stock. The sweets class follows a similar distribution to food but follows a normal wider distribution ranging from 0-5 days. Gifts and technology both follow a normal distribution with gifts ranging from 5-25 days delivery time which makes sense as gifts come in a wide range of products and therefore could lower the availability causing the delivery time to increase. Technology delivery time ranges from 7-32 days which is also expensive as the online store is not able to have a large amount of stock due to the high price of technology products therefore will need to be ordered first and then delivered if not in stock. The household class follows a normal distribution ranging from a deliver time of 35-60 days which is expected as household products are often customized by colour or fabric and therefore need to be ordered. The delivery time distribution of gifts is slightly skewed to the left and ranges between 5-15 days. The delivery time distribution of luxury products follows a normal distribution ranging from 0-7 days. This short delivery time could be due to the high price of the item that could be accompanied with higher priority.

4 Process Capability Measurements

4.1 Definition

Process capability analysis is a tool or a method to quantify the capability of the process to meet the design specifications so the business can build a product to meet customer expectations. Capability analysis helps the business understand and quantify what portion of their product is not going to meet the customer's needs. The business can then begin to imagine or visualize the amount of non-conforming products which aids in minimizing defects and ultimately saving on unnecessary costs.
Control deviation to avoid defects

4.2 Process Capability Indexes

The figure below gives a visual representation of the process capability indices to give a better understanding of the non-conforming parts and where they lie

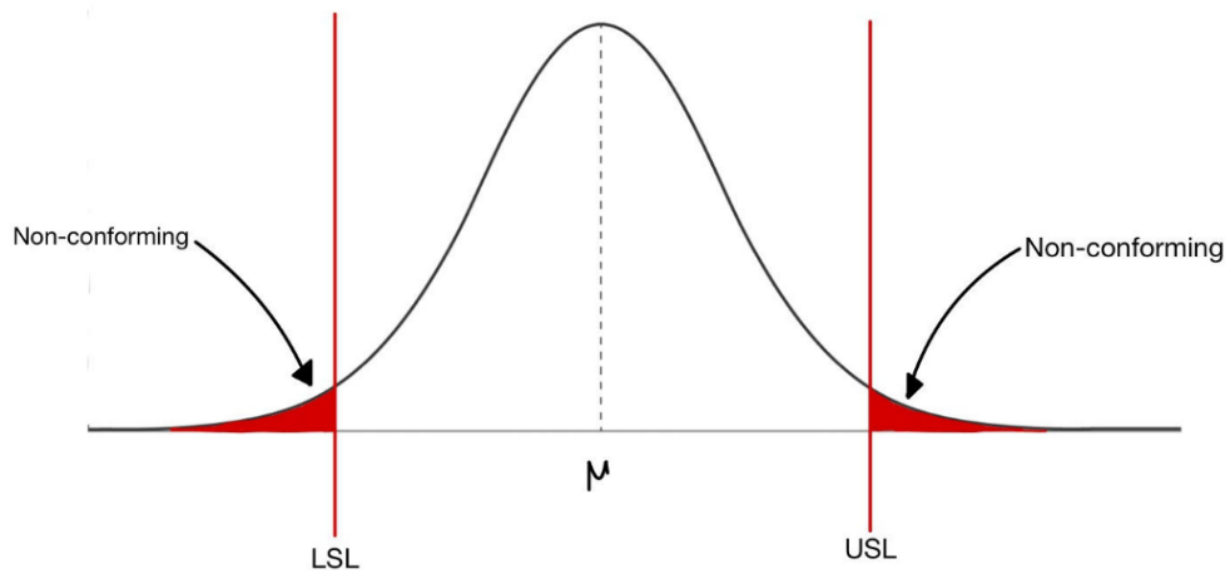


Figure 13: Figure showing the Process Capability Indexes

4.2.1 Process Capability

$$C_p = \frac{USL - LSL}{6\sigma} = 1.142207 \qquad [1]$$

C_p is the room available on specifications divided by the room needed for the process to operate. It measures the spread of the process variation

The upper specification limit was given as 24 days, where the lower specification limit was given as 0 days. A lower specification limit of zero days is logical since the online business cannot deliver an item before it is ordered online. The quickest possible delivery time is thus if an item were to be delivered on the same day the order was placed and paid for online, resulting in a 0-day delivery. The standard deviation of the technology class delivery times was calculated to be 7. 987665 days.

4.2.2 Upper-One-Sided Index

$$C_{p_u} = \frac{USL - \mu}{3\sigma} = 1.90472 \qquad [2]$$

4.2.3 Lower-One-Sided Index

$$C_{p_l} = \frac{\mu - LSL}{3\sigma} = 0.3796933 \qquad [3]$$

Both the upper-one-sided index and lower-one-sided index are used to calculate the process capability index which gives an indication of capability of the process relative to specification limits.

4.2.4 Process Capability Index

$$C_{p_k} = \min\left[C_{p_u};\, C_{p_l}\right] = C_p(1-k) = 0.3796933 \qquad [4]$$

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

$$T = \frac{USL - LSL}{2}$$

This ratio is defined as the process capability index and is used to summarize how a process is running relative to specification limits. This ratio only works well with normal distributed data and for other distributions it is merely an approximation. A ratio of less than 1 indicates that the process is not capable of meeting its requirements, therefore this is an alarming ratio and steps need to put in place to improve the businesses process.

5 Statistical Process Control

Statistical Process Control (SPC) is a methodology for monitoring a process to identify special causesof variation within a process and signal the need to take corrective action (Evans & Lindsay, 2018). SPC thus offers a way for the online business to show off its capacity for excellence.

SPC relies on Control Charts. A control Chart is simply a run chart to which three horizontal lines, called

control limits are added; the upper control limit (UCL), the central line (CL), and the lower control limit (LCL) (Evans & Lindsay, 2018). Plotting the control charts can help the online business pinpoint the exact moment that a process starts to vary, allowing them to concentrate their resources and efforts on preventing these variations (and subsequent errors) from happening. The control charts utilize data (in the form of in-flight measurements of the product and process) that is presented on the appropriate graphs with pre-established control limits. The capability of the process determines the control limitations, whereas the needs of the client dictate the specification limits.

It was noted in earlier sections that there were discrepancies in the business's delivery schedules for the various classifications of goods. Given that the company sells seven different product categories, a more thorough investigation of the delivery timeframes for each category of product would be necessary to pinpoint the precise location of the variation.

5.1 X-Chart Generation for Process Control Analysis

No matter how well a business process is designed there will always be some variation within the process. If the variation keeps you from meeting deadlines, it negatively impacts the business process. In such a scenario you will have to take necessary measures and this is exactly where control charts are beneficial for the organisation. The control charts are used to identify the causes of the variations in the process and prevents manufacturing a defective product.

A control chart is defined as the graphical representation that depicts whether the organizations products or processes meet the required specification.

For all product categories sold on the online store, control charts of delivery times were created. Every class category's data was organized from the earliest valid data to the most recent valid data. The descriptive features YEAR, MONTH, and DATE were used to carry this out. After the data had been sorted, samples of the data, each having 15 sales instances, were created, starting with the oldest valid data and moving through the ordered data.

A control chart for that particular class was created using the first 30 samples from each class. The first 30 samples of each class were used to create the control charts of type X-Bar. The rest of the samples of each class were used to create the s-charts. For each class, the 30 samples were used to determine the central limit, upper control limit, lower control limit, one sigma control limit, and two sigma control limit. These restrictions were also noted on the corresponding control charts. If the evaluated delivery times fell within the control range, they would have appeared between the upper and lower control limits. Appendix A contains the control charts created using the first 30 samples. Refer to the table below for an overview of the outcomes for each class from the x-charts:

Table 4: Table showing an overview of the outcomes for each class from the x-charts

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22.97462	22.10789	21.24117	20.37444	19.50772	18.641	17.77427
Clothing	9.404934	9.259956	9.114978	8.97	8.825022	8.680044	8.535066

Food	2.709458	2.636305	2.563153	2.49	2.416847	2.343695	2.270542
Gifts	9.488565	9.112747	8.736929	8.361111	7.985293	7.609475	7.233658
Household	50.24833	49.01963	47.79092	46.56222	45.33352	44.10482	42.87612
Luxury	5.493965	5.241162	4.988359	4.735556	4.482752	4.229949	3.977146
Sweets	2.897042	2.757287	2.617532	2.477778	2.338023	2.198269	2.058514

The table below shows an overview of the outcomes for each class from the s-charts:

Table 5: Table showing an overview of the outcomes for each class from the s-charts

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.18057	4.552222	3.923875	3.295528	2.667181	2.038833	1.410486
Clothing	0.86656	0.761455	0.656351	0.551247	0.446142	0.341038	0.235934
Food	0.437247	0.384213	0.33118	0.278147	0.225113	0.17208	0.119047
Gifts	2.246333	-2.64755	-0.60929	1.428965	3.467224	5.505484	0.611597
Household	7.34418	6.45341	5.56264	4.67187	3.7811	2.89033	1.99956
Luxury	1.511052	1.327777	1.144503	0.961229	0.777955	0.59468	0.411406
Sweets	0.835339	0.734022	0.632704	0.531386	0.430069	0.328751	0.227433

The headings from Table 4 are as follow:

- **UCL:** Upper control limit. This is also the three-sigma control limit, meaning that it is three standard deviations above the mean.
- **U2Sigma:** Upper two-sigma control limit, meaning that it is two standard deviations above the mean.
- **U1Sigma:** Upper one-sigma control limit, meaning that it is one standard deviation above the mean.
- **CL:** This is the center line of the delivery times, which indicates the mean of the delivery times across thedifferent samples for a specific class.
- **L1Sigma:** Lower one-sigma control limit, meaning that it is one standard deviation below the mean.
- **L2Sigma:** Lower two-sigma control limit, meaning that it is two standard deviations below the mean.
- **LCL:** Lower control limit. This is also the three-sigma control limit, meaning that it is three standard deviations below the mean

5.2 Continuous Charts for Control Purposes

To accurately ascertain whether the delivery time process is under control, all valid data after the first 30 samples must now be assessed. Using the control limits derived from the first 30 samples, s-charts were plotted for all remaining samples for this evaluation. The areas of concern will be identified in the prior sections using the s-charts that were created. The continuous s-charts created for every class category are displayed in figures below.

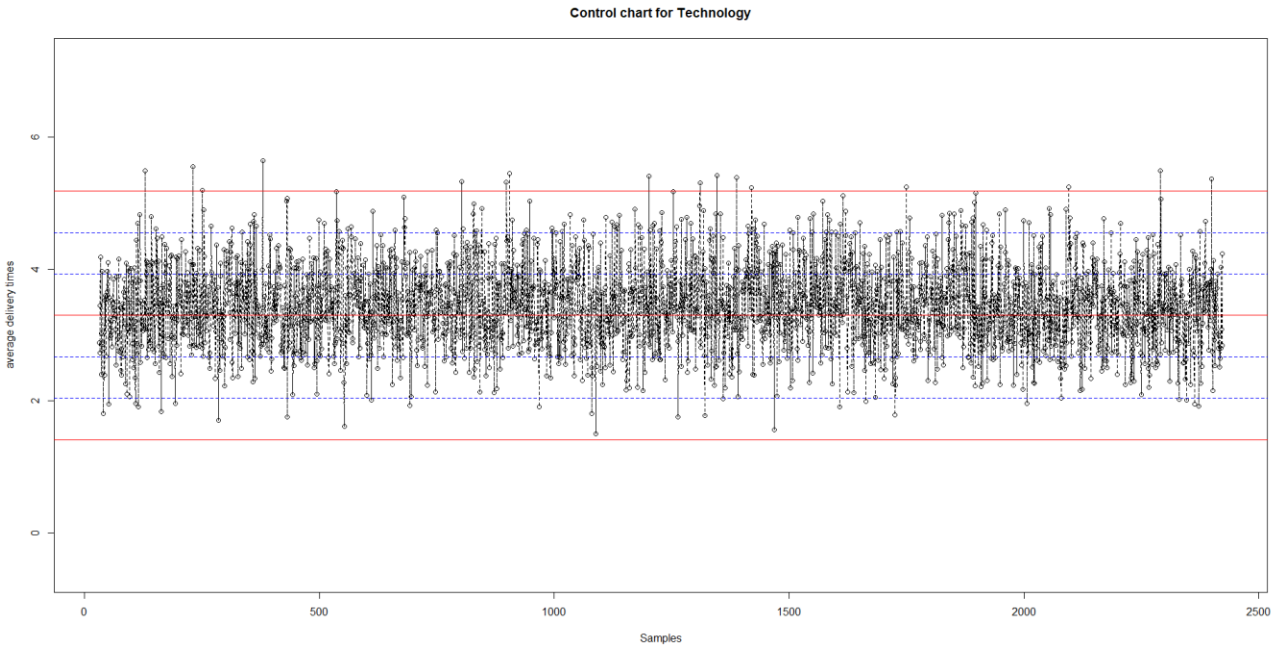


Figure 14: Continuous control chart for technology items

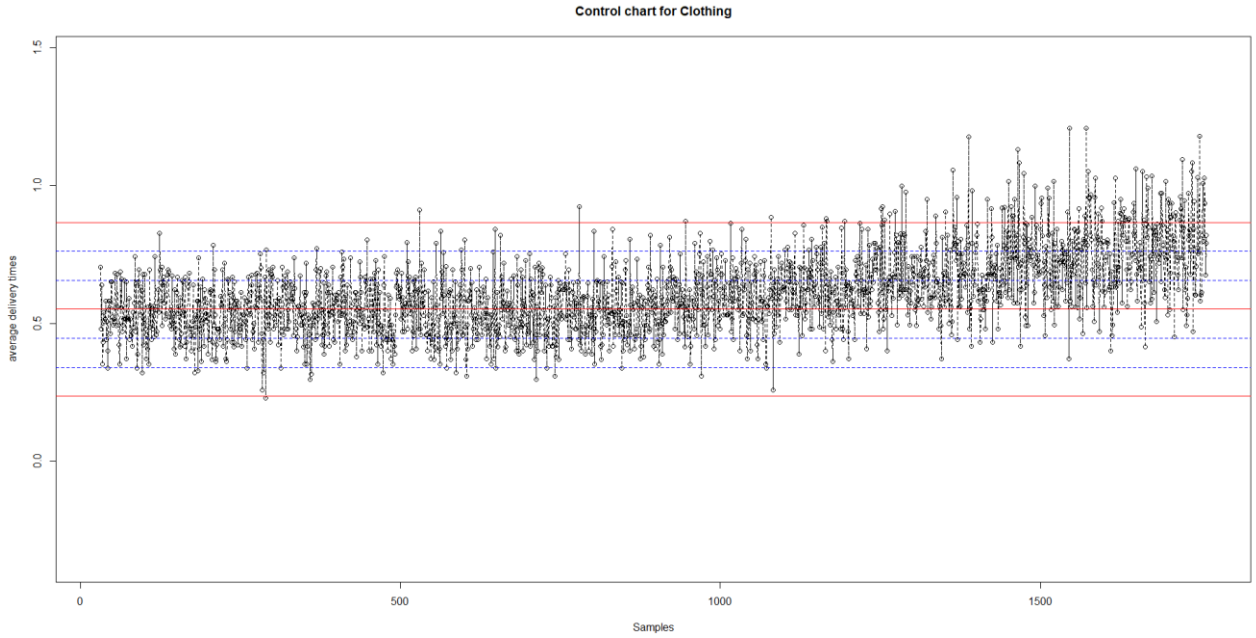


Figure 15: Continuous control chart for clothing items

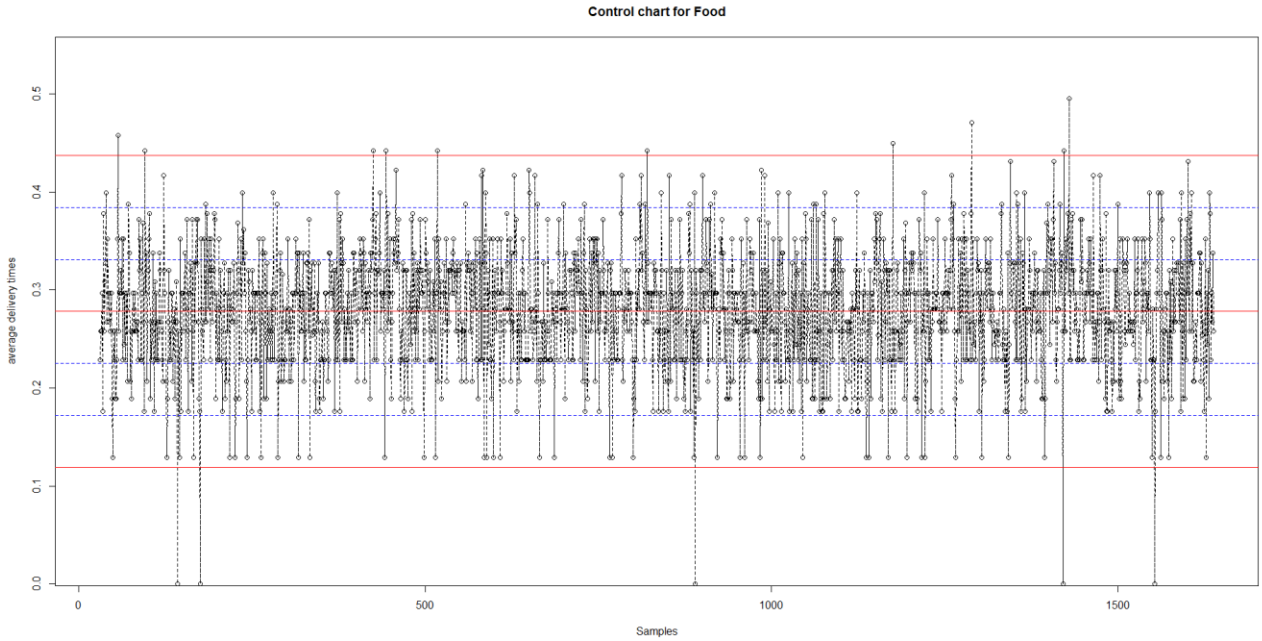


Figure 16: Continuous control chart for food items

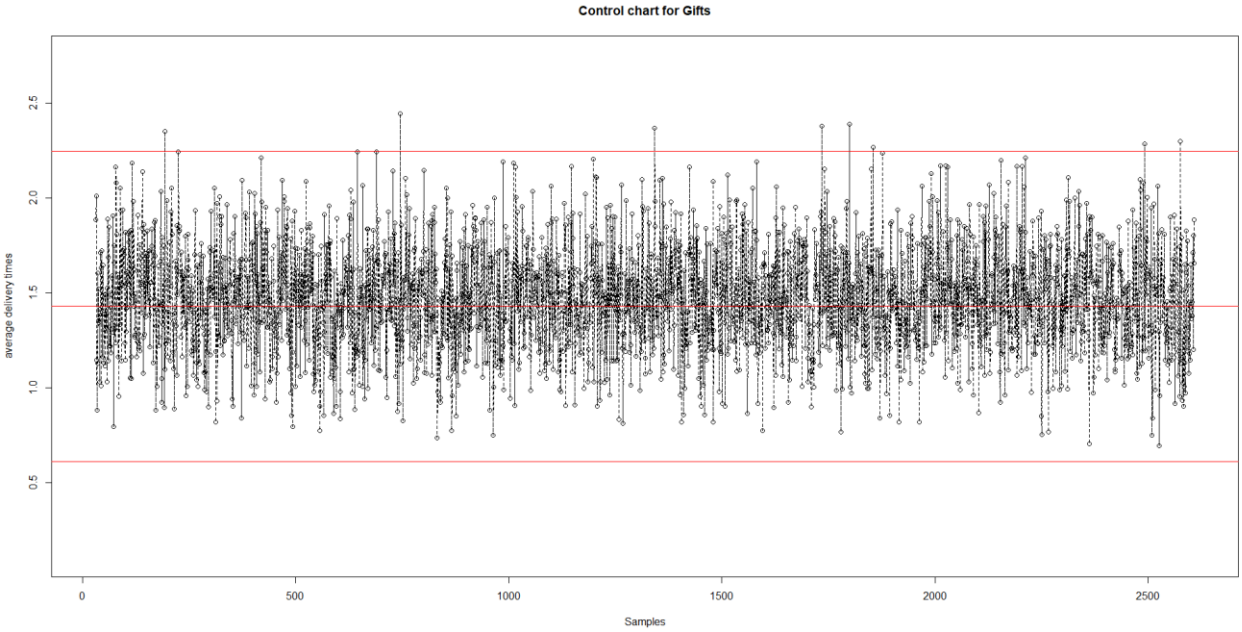


Figure 17: Continuous control chart for gift items

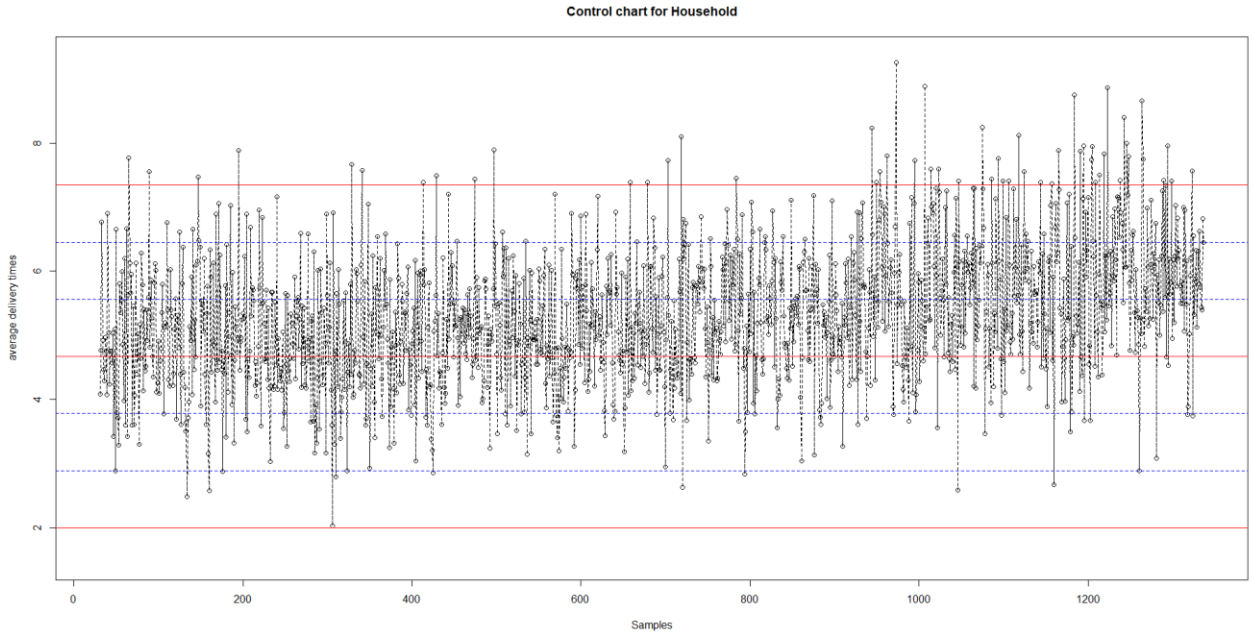


Figure 18: Continuous control chart for household items

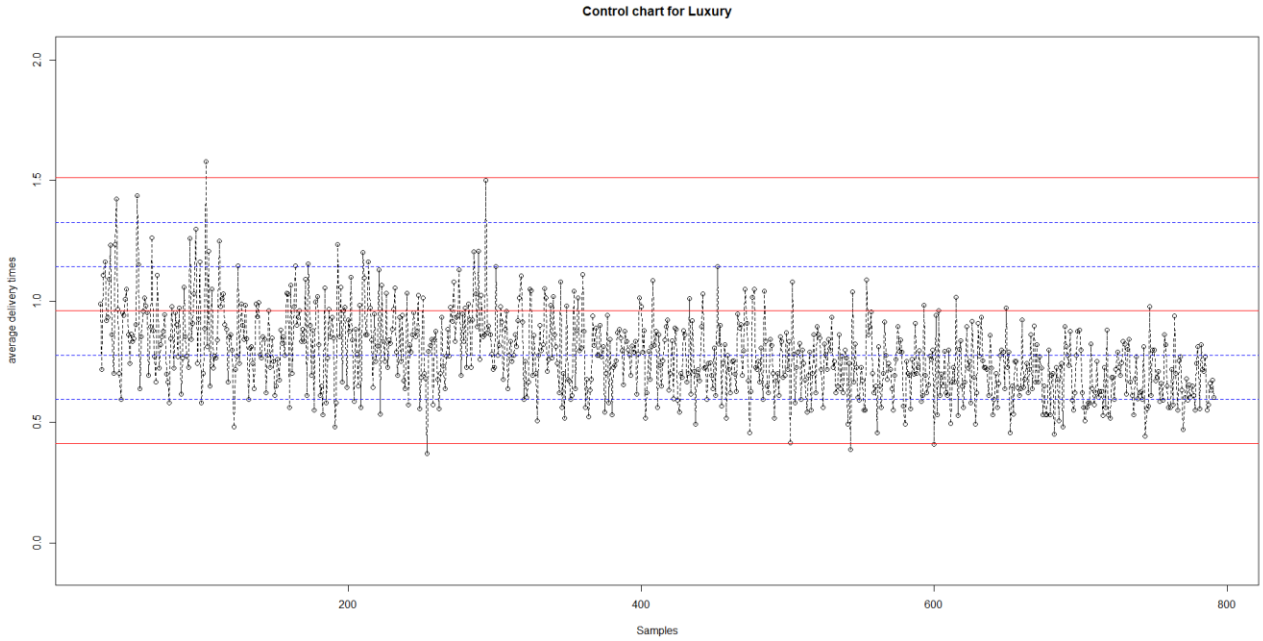


Figure 19: Continuous control chart for luxury items

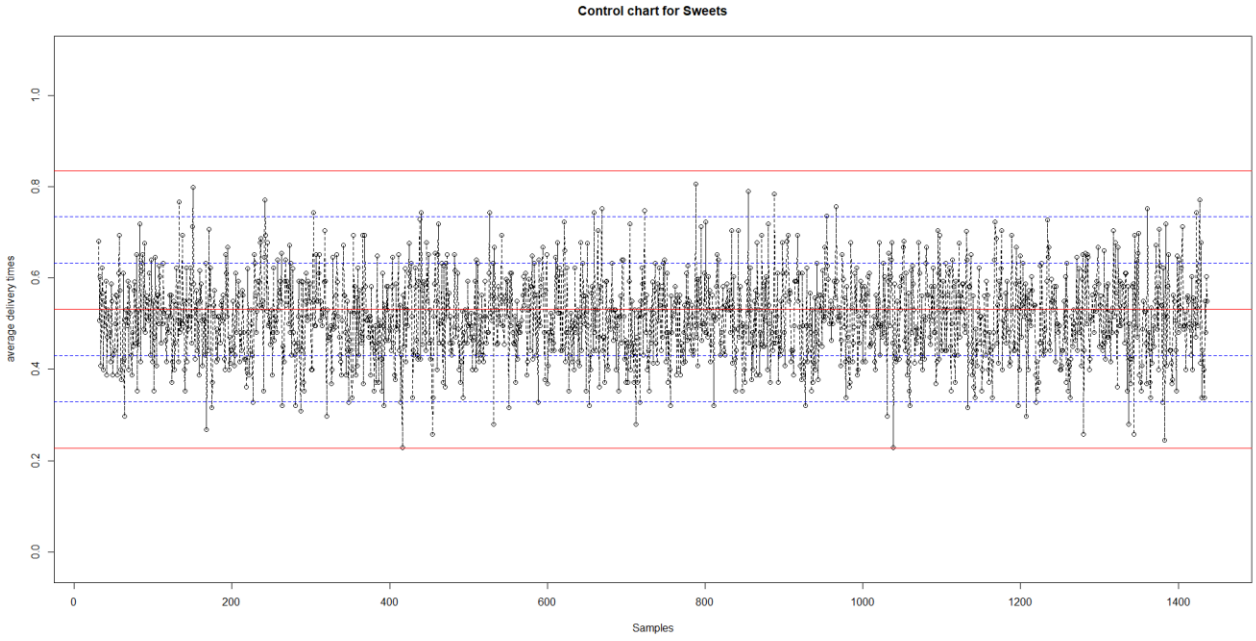


Figure 20: Continuous control chart for sweets

6 Optimizing Delivery Times

To determine whether a process is statistically under control, control charts are utilized. A process can be considered to be in control when examining the sample points on a control chart and meeting the criteria listed below:

- All points lie within the control limits. Thus, no points are above the upper control limit and nopoints are below the lower control limit.
- Points should be more or less equally distributed around the center line. Thus, the number of points above the center line should be very close to the number of points below the center line.
- The number of points should be more densely packed towards the center line, and less densearound the upper and lower control limits.

Only if the initial data set is regularly distributed are the aforementioned requirements true. Given that the delivery times for all class categories roughly correspond to normal distributions, as shown in the Quantity vs class graphs, the three aforementioned criteria can be utilized to determine whether the delivery times of the various class categories are under control.

6.1 Analysis of Delivery Time X-Charts per Class Category

Using the above criteria and control charts presented above the samples that appears to be out of control processes can now be identified by: (A) identifying X-Bar samples and S-Bar samples that are outside of the outer control limits, and (B) Identifying the most consecutive X-Bar samples within the -0.3 and +0.4 sigma control limits. The X-Bar samples that are outside of the outer control limits are shown below:

Table 6: The X-Bar samples that are outside of the outer control limits

Class	1st out	2nd out	3rd out	3rd last out	2nd last out	1st out	No. Samples out
Technology	37	398	483	1872	2009	2071	17
Clothing	455	702	1152	1677	1723	1724	17
Food	75	633	1203	N/A	1467	1515	5
Gifts	213	216	218	1317	1318	1319	2290
Household	252	387	629	331	1332	1333	400
Luxury	142	171	184	775	776	777	434
Sweets	942	1104	1243	N/A	1294	1403	5

The S-Bar samples that are outside of the outer control limits are shown below:

Table 7: S-Bar samples that are outside of the outer control limits

Class	1st out	2nd out	3rd out	3rd last out	2nd last out	1st out	No. Samples out
Technology	129	230	251	2095	2290	2400	16
Clothing	289	530	780	1754	1756	1757	98
Food	19	57	96	1422	1429	1553	16
Gifts	193	746	1342	1855	2493	2576	8
Household	65	89	147	1294	1299	1323	54
Luxury	103	254	543	N/A	N/A	600	4
Sweets	18	N/A	N/A	N/A	N/A	N/A	1

The number of consecutive samples in each class are listed below:

Table 8: Consecutive samples

Class	Number of Consecutive Samples
Technology	6
Clothing	346
Food	4
Gifts	3
Household	3
Luxury	2
Sweets	5

Interpretation:

Using the information in the tables above and the s-charts above, the classes that gave indication of out-of-control delivery times were identified.

6.1.1.1 Clothing items:

When looking at the figure below and the results obtained from the analysis documented in Table 4, it is evident that clothing delivery time process is out of control. From the 1760 clothing delivery time samples, 98 arefalling outside of the control limits. This means that 6% of the luxury delivery times are not in control. Looking at Figure 19, this is mostly because of a sustained increase in the deviations from the mean, as circled in blue. This indicates that the delivery times for the clothing items started deviating a lot from the average delivery times from approximately sample 1250, which means that the process had not been adjusted accordingly. To improve the clothing delivery times, it is important for the business to redesign their processes to decrease delivery times.

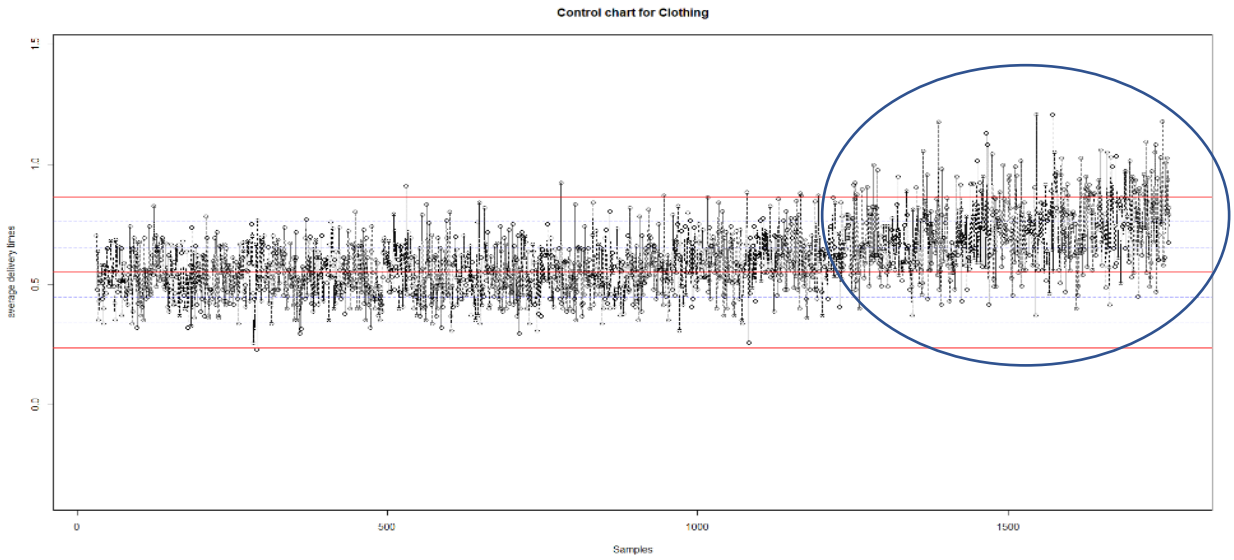


Figure 21: Figure showing the out of control delivery times for clothing items

6.1.1.2 Household Items:

When looking at Household s-chart and the results obtained from the analysis documented in Table 4, it is evident that the household delivery time process is out of control. From the 1337 household delivery time samples, 54 are falling outside of the control limits. This means that 4% of the household delivery times are not in control. This percentage can mostly be due to the sudden and sustained increase in the deviation from the mean of sample 950 and onwards, as circled in orange on below. This resulted in a large upwards shift of the delivery time. Therefore, the company must halt the process and conduct thorough investigations to determine what changes were made around this period that may have

contributed to the significant variances in the delivery schedules for the household goods. Household delivery times has a standard deviation of 6.22.

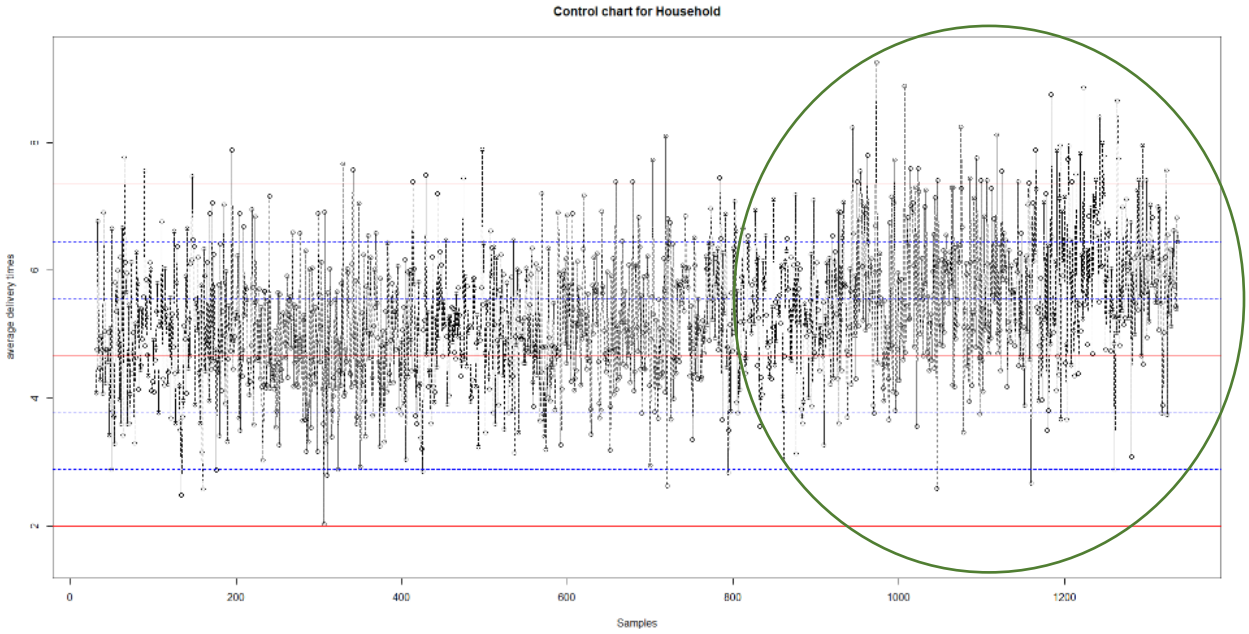


Figure 22: Figure showing the out of control delivery times for household items

6.1.1.3 Important Note on Technology Delivery Times:

It was once thought that the high standard deviation of technology lead times might be a sign that the method for controlling technology delivery times was out of control. However, after examining the Technology s-chart and the findings of the study listed in Table 3, it is obvious that the process of technology delivery time is actually under control. This indicates that even though the standard deviation and mean are alarmingly large, the procedure is functioning as intended. Technology items have a reasonable average delivery time of 20 hours

6.2 Error Analysis

A Type 1 error is the very first kind of error that can happen. A Type 1 error occurs when a process is terminated even though it was functioning as intended because of some sign that it is not behaving as expected. Due to this, a company may waste time and money looking for answers to an issue that doesn't actually exist, as well as unnecessarily pausing a process. Type 2 errors are the second type of error that might happen. When a process continues to run even when there are signs that it is not working as planned but there are signs that it is, this is known as a Type 2 error.

6.2.1 Type 1 Error Analysis

The probability of making a type 1 error when (A) identifying X-Bar samples that are outside of the outer control limits, and (B) Identifying the most consecutive X-Bar samples within the -0.3 and +10.4 sigmacontrol limits, was calculated. The results of this calculation are documented in the table below.

Table 9: Table showing the probability of making a Type 1 error for all classes

Type 1 Error	Probability of Type 1 Error
A	0.27 %
B	27.33 %

Interpretation:

In order to evaluate Type 1 error probabilities correctly, it is necessary to first comprehend the null hypothesis. For the Type 1 error calculations, the null hypothesis is that the process is under control and centered on the center line. This means that while the business presumes the process is centered and under control, the probabilities show a higher likelihood that this is not the case.

From It can be inferred from the probabilities shown in the table above that utilizing method A to identify out of control processes is appropriate because there is only a 0.27% probability that technique A will result in a type 1 error. It will be illogical to use technique B to establish whether a process is under control. This is because when identifying out-of-control processes using approach B, there is a 27.33% risk that you will make a Type 1 error. According to the outcomes of technique B, a process is stopped 27.33% of the time even though it wasn't required.

6.2.2 Type 2 Error Analysis

The likelihood of making a type 2 error will now be illustrated by using the delivery times of Technology items. The current mean delivery time for technology items is 20 hours. If the delivery time mean wereto shift to 23 hours, the probability of making a type II error is 3.4%. This indicates that if the process were to shift to a mean of 23 hours, 3.4% of the time the process needs to be stopped, it won't be stopped.

6.2.3 Re-Engineering Technology Delivery Times

Finding the best delivery times for different classes of items is crucial for cost savings, but it should also be viewed by the company as a significant lever for increasing customer satisfaction.

The business will benefit from recalculating and reexamining its delivery process for technology items because the technology class has the biggest variation in lead times.

If it costs R0.25/item/day to reduce the average delivery lead time by one day, and the organization loses R329/item-late/day in lost sales for delivering later than 26 hours, the cost function for increasing/decreasing delivery times is as illustrated the figure below.

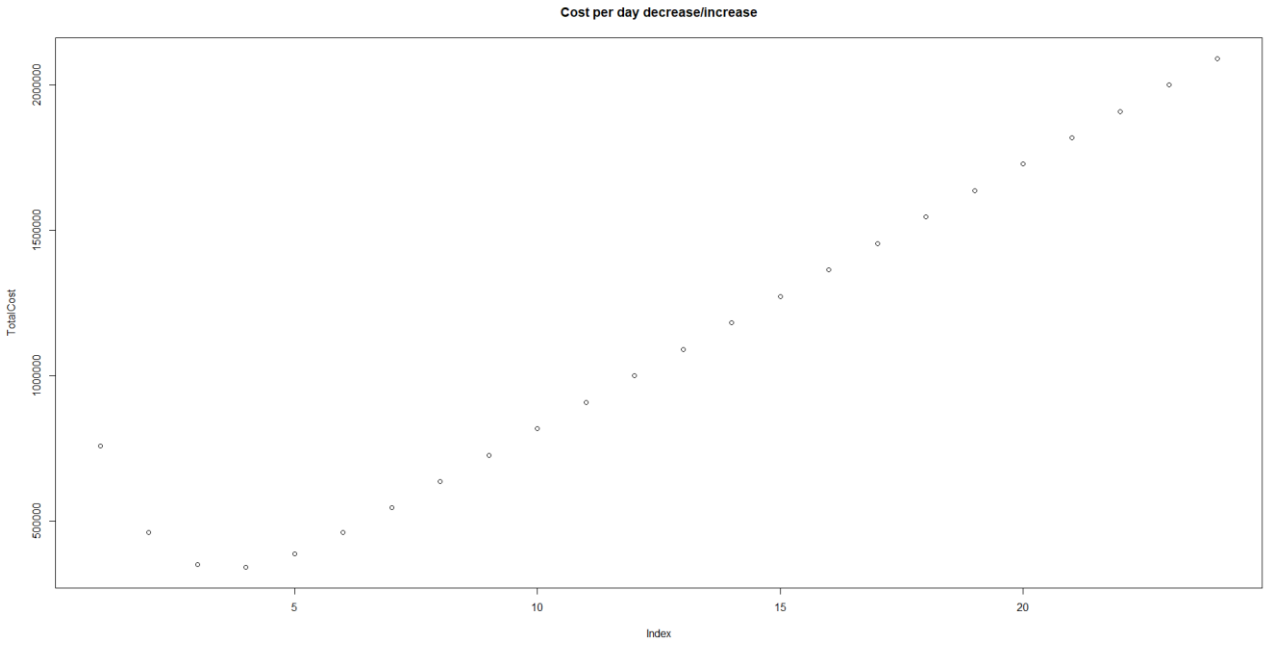


Figure 23: Figure showing the cost function per day increase/ decrease in delivery times

The business should cut the mean delivery time of technological goods by 3 hours, according to statistical analysis, to keep the cost of delivery to R 340 870. The decided ideal delivery mean time will be 17 hours because the first mean delivery time for technological items was 20 hours. This further supports the conclusion that the company should redesign its method for delivering technologies.

7 MANOVA

MANOVA is about comparing one or more dependent values across two or more groups. MANOVA stands for multivariate analysis of variance. MANOVA can have two or more groups and when you do it we can look at one or more dependent variable. Does a curve fit of the dependent variables and does a comparison of the different curves it is important that they have normality?

The association between the earlier-discussed descriptive characteristics and the goal feature, "delivery time," is examined using a MANOVA. The MANOVA findings will show how substantial the influence of various characteristics on a process' delivery time is. The MANOVA is run with an alpha level of 0.05 and the following hypotheses:

The assumption is that an item's class and price have a substantial impact on how quickly it is delivered.

The alternative is that the class and price of an item purchased do not significantly affect how quickly it will be delivered.

Basic assumptions of MANOVA include:

- Independent random sampling
- Level measurements of variables
- Linearity of the dependent variables
- Multivariate normality
- Multivariate homogeneity of variance within groups and between groups.

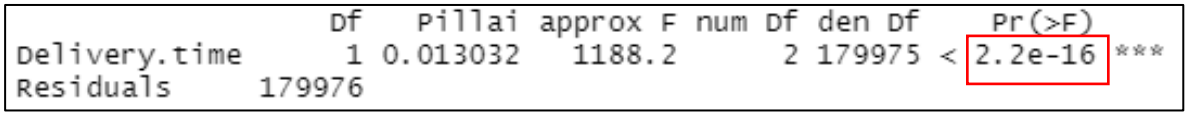


Figure 24: Figure showing the P-Value from the MANOVA results

The P-Value, shown by the red box in Figure 21, is lower than the previously selected significant level of 0.05. This suggests that it is appropriate to accept the null hypothesis. Thus, it can be inferred from the aforementioned MANOVA results that an item's price and class have a considerable impact on the item's estimated delivery date.

The association between the descriptive parameters Age and Price is also examined using a MANOVA and the Class of the target feature, which is the focus of this MANOVA. The MANOVA findings will show whether or not the class of item sold may be determined using various customer demographics and item pricing.

With an alpha level of 0.05, the following hypotheses are tested using MANOVA:

The assumption is that there is no substantial correlation between the type of product offered and the age of customers and the cost of the goods sold. The opposing argument is that there is no substantial correlation between the class of product offered and the age of customers and pricing of goods purchased.

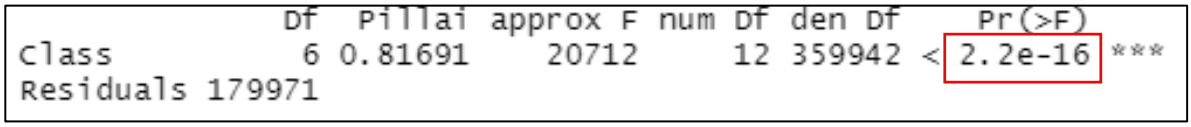


Figure 25: Figure showing the P-Value from the MANOVA results

The P-Value, shown by the red box in Figure 22, is less than the previously selected significant level of 0.05. This suggests that it is appropriate to accept the null hypothesis. Thus, it can be inferred from the aforementioned MANOVA results that there is a substantial correlation between the kind of item sold, the average customer age of 41, and the pricing of the things sold. This means that the class of an item sold may be determined using a variety of customer demographics and item pricing.

8 Reliability of the Service and Products

8.1 Lafrigeradora Analysis

The ecommerce store receives parts for food refrigerators from Lafrigeradora. The online retailer wants to know if the refrigerator components meet the requirements, it gave Lafrigeradora and how much the present level of quality is costing them. The Taguchi Loss Function is applied to find this.

Genichi Taguchi, a Japanese engineer, defined quality as the financial benefit of decreasing variance and manufacturing in accordance with the nominal standards (Evans & Lindsay, 2018). Taguchi defined quality as the deviation from a design specification's target value and converted that deviation into an economic "loss function" that expresses the monetary cost of variation (Evans & Lindsay, 2018).

The Taguchi Loss Function is as follow:

$$L(x) = k(x - T)^2 \qquad [5]$$

Where:

- $L(x) = monetary\ loss\ (assumed\ to\ increase\ quadratically)$
- $T = target\ value\ so\ as\ to\ optimise\ performance\ x{=}any\ actual\ value\ of\ the\ quality\ characteristic$
- $(x{-}T) = deviation\ from\ the\ target\ value\ T\ k{=}contant, associated\ with\ the\ cost\ of\ a\ cetain\ deviation\ from\ T$

Function 8.1 can now be used and applied to problem 6 and problem 7 of Chapter 7 (page 363) of Evans & Lindsay (2018) to obtain the following:

Working with a scrap value of $45 per unit and a tolerance of 0.035cm, we can obtain the constant associated with this cost and tolerance as follow:

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

Given the target value of 0.04 and substituting these values into equation 8.1, we obtain the loss function below to describe the situation:

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

When the scrap value is then changed to $35 per unit, a new constant associated with this cost can be computed as follow:

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

Given the changed scrap cost and resulting change in the constant value, the new Taguchi Loss Function to describe the situation is:

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

If the process deviation from target can be reduced to 0.027cm, the Taguchi Loss is as follow:

$$x = 0.027 + 0.04 = 0.067$$

$$L(0.067) = 21875(0.027)^2 = 15.94688$$

8.2 Magnaplex Analysis

Magnaplex manufactures some of the technology items sold on the website. At this current moment Magnaplex, Inc. has a complex manufacturing process, with three operations that are performed in series. Because of the nature of the process, machines frequently fall out of adjustment and must be repaire. To keep the system going, two identical machines are used at each stage; thus, if one fails, the other can be used while the first id repaired.

Reliability using single machines:

$Reliability = Machine\ A\ reliability \times Machine\ B\ reliability \times Machine\ C\ reliability$

$Reliability$ = 0.85 × 0.92 × 0.9 = 0.7038

Reliability using parallel machines:

$Reliability$ = 1 − (1 − $Machine\ A\ reliability$) × (1 − $Machine\ B\ reliability$) × (1 − $Machine\ C\ reliability$)

$Reliability$ = 1 − (1 − 0.85) × (1 − 0.92) × (1 − 0.9) = 0.9988

It is evident from the previous calculations that the reliability percentage significantly increases when using machines in parallel, from 70.38% to 99.88% reliability. Therefore, it would be more efficient for Magnaplex to use parallel machines.

8.3 Delivery Process Analysis

This analysis was performed using the past 1560 days of data. The result of this analysis was that there is a 73.47% probability to have 20 trucks and 21 drivers available, therefore failing to meet delivery process demands. This means that the delivery process is reliable for 268 days of the year, with 97 days of unreliable delivery. This is not optimal as it effects customer satisfaction and loyalty therefore the decision was made to investigate the delivery process reliability if an additional truck were added, while keeping the existing number of drivers. By increasing the number of trucks from 20 to 21, the delivery process reliability goes up to 83.16%. This means that the delivery process is reliable for 303 days of the year, with 62 days of unreliable delivery. This is still not an acceptable reliability if the business aims to be responsive and provide good customer service therefore more trucks will need to be purchased and an investigation of the delivery process will need to be done in order to identify key problem areas and improve the overall reliability of the delivery process.

9 Conclusions and Recommendations

The original raw sales data set was pre-processed and cleaned by getting rid of missing values and unwanted data. This cleaned data was then turned into meaningful business insights through the generation of statistical models that were analysed in depth where useful recommendations were made to aid the business in solving multiple problem areas identified throughout the report and further improve their marketing strategy. Once the statistical models were interpreted, process capability measures were calculated in order to quantify the capability of the process to help the business meet design specifications. During the generation of statistical models, control charts were developed to monitor a process in order to identify special causesof variation within a process and signal the need to take corrective action which offers a way for the online business to show off its capacity for excellence. Following this, MANOVAs were carried out to determine which descriptive variables in the dataset were reliant on one another as well as how closely they correlated with delivery timings. Finally, the dependability of the internet business's suppliers and internal delivery procedures were investigated.

10 References

What are numeric features in data science? (no date) *Educative*. Available at: https://www.educative.io/answers/what-are-numeric-features-in-data-science (Accessed: October 23, 2022).

Saha, S. (2022). *Data Science vs Business Analytics – Top 5 Differences*. [online] www.knowledgehut.com. Available at: https://www.knowledgehut.com/blog/data-science/data-science-vs-business-analytics [Accessed 15 Oct. 2022].

Mahoney, M. (n.d.). *4 Data Wrangling | Introduction to Data Exploration and Analysis with R*. [online] *bookdown.org*. Available at: https://bookdown.org/mikemahoney218/IDEAR/data-wrangling.html [Accessed 5 Oct. 2022].

Calzon, B. (2021). *See Top Analytical Report Examples & Business Templates*. [online] BI Blog | Data Visualization & Analytics Blog | datapine. Available at: https://www.datapine.com/blog/analytical-report-example-and-template/ [Accessed 20 Oct. 2022].

Hessing, T. (2014). *X Bar R Control Charts | What you need to know for Six Sigma certification.* [online] Six Sigma Study Guide. Available at: https://sixsigmastudyguide.com/x-bar-r-control-charts/.

Calzon, B. (2021). *See Top Analytical Report Examples & Business Templates*. [online] BI Blog | Data Visualization & Analytics Blog | datapine. Available at: https://www.datapine.com/blog/analytical-report-example-and-template/ [Accessed 20 Oct. 2022