ECSA GA4 GRADUATE ATTRIBUTES PROJECT

QA344

Pretorius, T, Miss [23540907@sun.ac.za]
Industrial Engineering, Stellenbosch University

Abstract

In this report the sales data from a company is presented to be analysed. The data is split into two datasets namely complete and incomplete data by removing the missing values. Descriptive statistics are used to visually represent the data and understand different features. The quality is assessed by Statical Process control with statistical techniques to identify the out-of-control processes. The accuracy of the data is assessed by find the probability of making a Type I and II error. A MANOVA test is performed to test stated hypothesis. The reliability of services and products are lastly assessed. The report concludes with final comments on the sales data.

Table of Contents

Table of Figures

List of Tables

Introduction

The importance of data analysis and data handling has increased over the past few years. The matter in which companies use accumulated data has a significant impact on the profitability and improvement of the company with an increase in the competitive ability in the market.

In this project a data set containing information regarding the client data and sales for an online business is given and used in analysis. The relevancy and importance of the different features will be analysed to ultimately make recommendations and conclusions that can assist the business. In the first part the data is cleaned and split into two namely a complete and incomplete dataset. The incomplete data consists of instances containing negative values and missing values which are identified and removed during data wrangling. This is ultimately to understand the dataset and to have accurate and complete data. By using descriptive statistics, it enables an understanding of the data. This part creates a visual representation of the data and identification of the different relationships between features.

The next part is Statistical Process control which is used to do quality control by means of statistical techniques. This helps identify the worst and best classes in the dataset and the season for their performance. The identification of problems contributes to the stability of the processes which can be analysed. The probability of having Type I and Type II is measured to know the accuracy of the results and conclusions from analysis can be. The MANOVA test is used to determine if different independent features have an impact on dependent features either on its own or in combination with other features. One business goal is to have reliable delivery to ensure customer satisfaction and fast delivery to customers. The investment into new projects is also investigated to analyse the effect and possible increase in advantages for the business in terms of profit and reliable delivery.

Part 1: Data Wrangling

The data wrangling part involves the cleaning, restructuring, and enriching of the data. It is a pre-processing step that involves the necessary transformation of data into a format that is functional for processing and decision making in terms of analysis (Peacer, 2020). As the data will be used in analyses it is important to have well prepared and cleaned data that can be analysed with ensured accuracy and correctness.

First a data quality report is produced to identify the quality of the data. This is also used to identify the missing values (Bright Data. (n.d.)). The dataset is divided into two separate datasets namely a valid sales and invalid sales data set with the invalid sales being all the instances in the data containing missing values which are the incomplete instances.

The following criteria was used to identify the quality of the data. The data should be relevant to the dataset. All instances should be complete meaning no missing values should be present and the instances should be up to date. It is also very important that the data is consistent (Shen, 2021).

There are 180 000 entries of data in the unprocessed dataset. The only feature containing missing values is Price. There are 17 instances that contain missing values. All the instances containing negative values should also be removed. Based on the criteria there are no other data quality issues identified by the data quality report. This leaves 179978 instances that are valid and 22 instances are removed. The complete and incomplete data is split into two separate datasets. This is mainly done to prepare and understand the data.

Feature identification:

It is important to classify and identify the different features. Features are classified as either categorical or continuous as follows:

Categorical features are ID, Age, Class, Year, Month, Day and Why bought. The continuous features are Price and delivery time as they are continuous over time and not discreet values.

Table 1 Main point from the data quality report for numerical features

Feature	Count	Miss	Card	Mean	Median	SD
ID	180000	0	15000	55235.08	55081	25739.67
Age	180000	0	91	54.57	53	20.39
Price	180000	17	78834	12293.74	2259.63	20888.97
Year	180000	0	9	2024.86	2023	2.78
Month	180000	0	12	6.52	7	3.45
Day	180000	0	30	15.54	16	8.65
Delivery time	180000	0	148	14.50	10	13.96

A feature with a high cardinality is the ID feature. This indicates that there are mostly unique values within this feature with no real relationship, the other features are relevant with lower cardinality giving an indication of possible correlation between instances that may arise during the analysis.

Part 2: Descriptive Statistics

Descriptive statistics are generally used to describe and analyse the basic features of the data that are relevant in the study. It gives a description of the data from various perspectives and by isolating different features to define the correlation. It forms a foundation for various other quantitative analysis of data. In this part the features will be visually represented individually and in relation to different features with different graphing techniques (Trochim, 2022)

By means of descriptive statistics a description of the nature of the business's data in terms of sales can be obtained. First the features were identified and described by visual representation to view the distributions. After that features were compared by means of different graphs to understand the distributions and possible rationale behind the tendency of certain features.

Identifying and defining the features

The features are looked at individually first and the range of values that define the specific features, in other words the distribution from minimum to maximum are defined that is needed when the feature will be compared to find relationships.

The features are analysed by identifying the following information:
Number of rows=179978
This relates to the instances of valid data that are complete in the dataset.
Number of columns=10

This gives an indication of the number of features present in the dataset:

Table 2 Mode characteristics for the categorical features

Feature	Count	Mode	Mode Freq	Mode%	2ndMode	2ndMode Freq	2ndMode%
ID	179978	41842	27	0.015	47570	26	0.014
Age	179978	38	3130	1.739	39	311	1.731
Class	179978	Gifts	39149	21.752	Technology	36347	20.195
Year	179978	2021	33443	18.582	2029	22475	12.488
Month	179978	12	15225	8.459	10	15221	8.457
Day	179978	17	6126	3.404	25	6122	3.402
Why Bought	179978	Recommended	106985	59.443	Website	29447	16.361

Feature 1: ID

There are 15000 unique values which means 15000 unique IDs for different customers who buy from the business. The customer that buys the most from the company is responsible for 27 of the 179978 sales, which is a small number of sales. This shows that there is not a specific customer that dominate the sales. As seen in the histogram in *figure 1* for the distribution of the different IDs of customers, it is uniformly distributed meaning that the number of sales over the range of customer IDs are not influenced by one or even a group of customers.

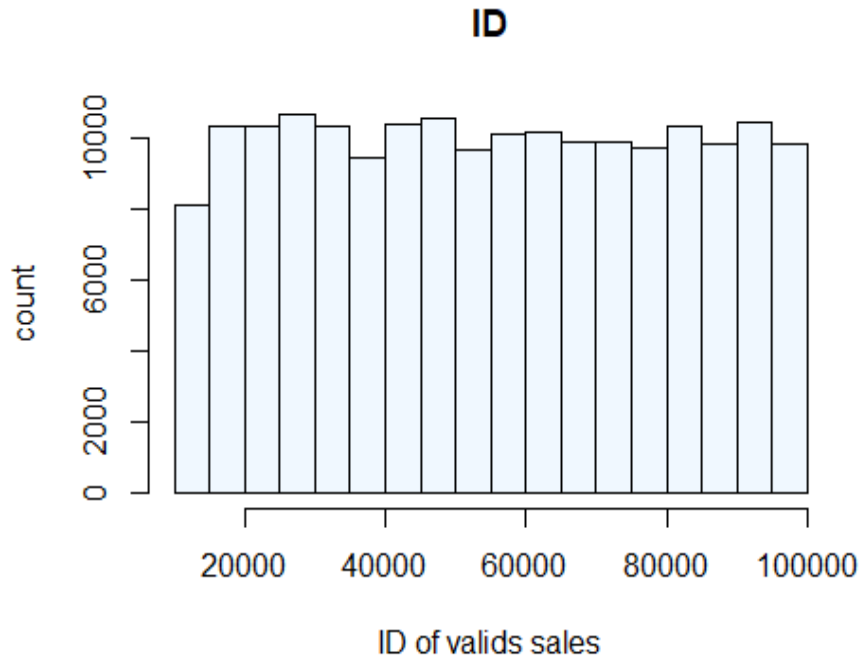


Figure 1 Histogram of the customer IDs

Feature 2: AGE

The age of the customers that buys the most products are 38 and the second most is customers of age 39. As both are relatively responsible for the same number of sales and from the graphs below it indicated that sales are not dominated by a specific age of customers.

Figure 2 Histogram of the customer Ages

The age of customers is distributed over the range of 18, which is the minimum, to 108 as the maximum age. The sales are relatively spread evenly between ages 30 and 55. The sales decrease from age 60 to 110. There are very few customers between age 15 and 20 and the sales increase from age 20 to 30.

Feature 3: Class

There are 7 different classes indicating 7 different products that can be bought by customers. These different classes are Technology, Clothing, Household, Luxury, Food, Gifts, Sweets.

The most bought classes are Gifts and Technology

Figure 3 Bar plot of the different product classes

The barplot above indicates the different classes and the number of customers who buy specific products. The least number of products sold are the Luxury products and the most are the Gifts products.

Feature 4: Year

The unique values indicated for the Years are from the year 2021 to 2029. The most sales happened in 2021 and the second most in 2029

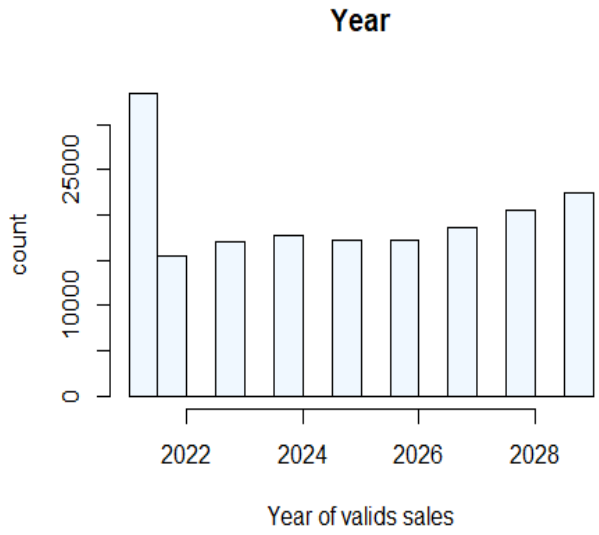


Figure 4 Histogram for sales over given years

Sales were high in 2021 and dropped in 2022. After this it followed a increasing pattern, increasing to 2029.

Feature 5: Month

All 12 months are represented in die data. The most sales occurred in month 12 and 10 and the sales are relatively close to each other.

Figure 5 Histogram for sales over different months

From the graphs it shows that the sales in different months are evenly distributed and do not differ, it however can differ in terms of the different times of the month. This indicates that there is not seasonality in sales within months.

Feature 6: Day

The days range from the 1st day of a month to the 31st of a month. From the table the first mode and second mode are very close to each other. From the graphs it indicates that the sales are evenly distributed within months and equal through the month. There is no seasonality that is shown within the data, it means that sales to customers stay relatively the same in the month.

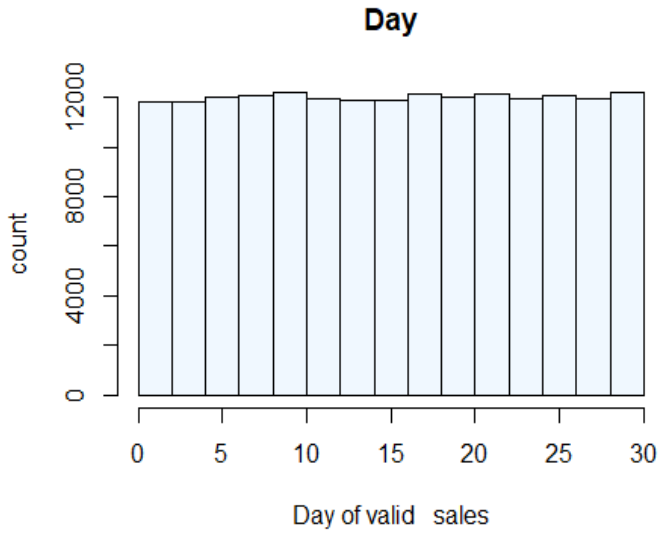


Figure 6 Histogram for sales over different days of the month

Feature 7: Delivery time

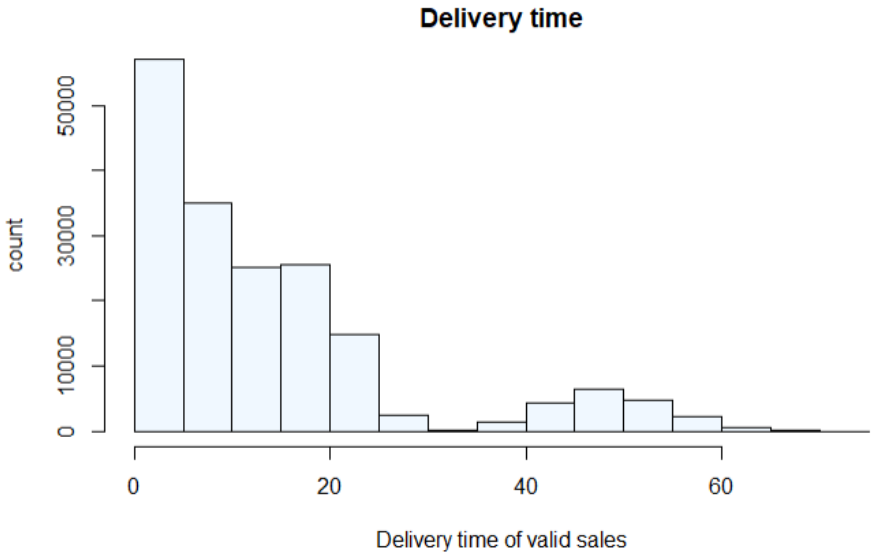


Figure 7 Histogram of the Delivery time for valid sales

The shortest delivery time is 0.5 days and the longest is 75 days. From the histogram it is clear that the delivery time peak is within the shorter range between 0 and 25 days. This is overall good for the business as it indicates reliable and quick delivery to customers.

Feature 8: Why bought

There are 6 different reasons that customers buy the products of the business namely: Recommended, Website, Random, Browsing, Email and Spam. From the table it indicates that a significant number of customers buy from the business because it is recommended to them.

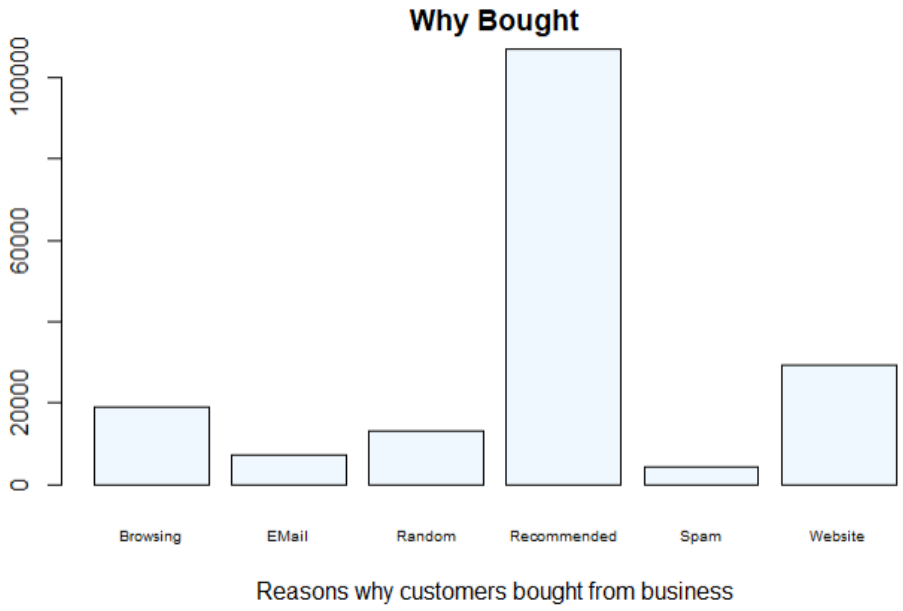


Figure 8 Bar plot to indicate the reasons that customers buy products

This plot gives a clear indication that one of the biggest drivers for customers to buy from the business is recommendations. This shows customer loyalty and satisfaction. The use of Email and Spam are not effective and can be improved.

Relationships between different features

Year vs delivery time

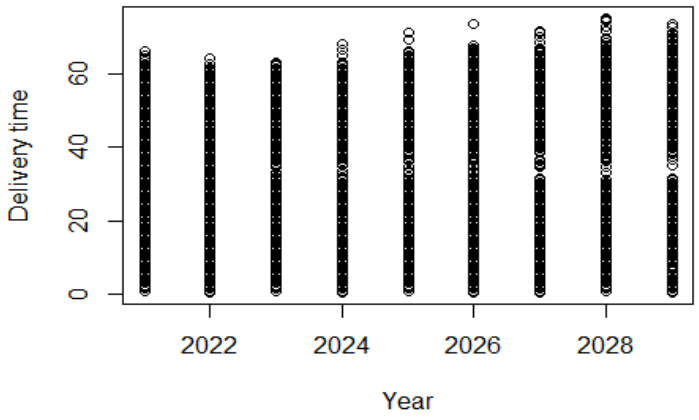


Figure 9 Scatterplot of delivery time compared over the years

There is a upward trend shown in the graph slightly increasing over the years. This can be an indication of more products that were sold over the years and that the workforce has not increased to ensure delivery time stays the relatively the same. With more products being sold the delivery time is increased and the company should improve this to ensure that future reliability in terms of delivery can be met.

Price vs class

The most expensive class is Luxury products. It is above $40000 and significantly more expensive than the other classes. The other two classes that are also more expensive is the technology and household. The other classes are less expensive.

Figure 10 Boxplots of different classes compared to price

Age vs class

Figure 11 Boxplot: Age of customers compared to class of products

The graph shows the distribution of ages for different classes. For certain ages different products are more popular. For Clothing the majority of customers are between 30 and 60 years of age. Different products have slightly different target markets.

Age vs why bought

The plots indicate that all the different classes are relatively the same. There is no conclusion that can be specifically made about the different ages and how it correlates to the reason for buying a product is bought by a customer. It is spread There isn't a specific marketing tactic that can be established for different age groups as the reason why the bought a certain product is an evenly spread for all ages.

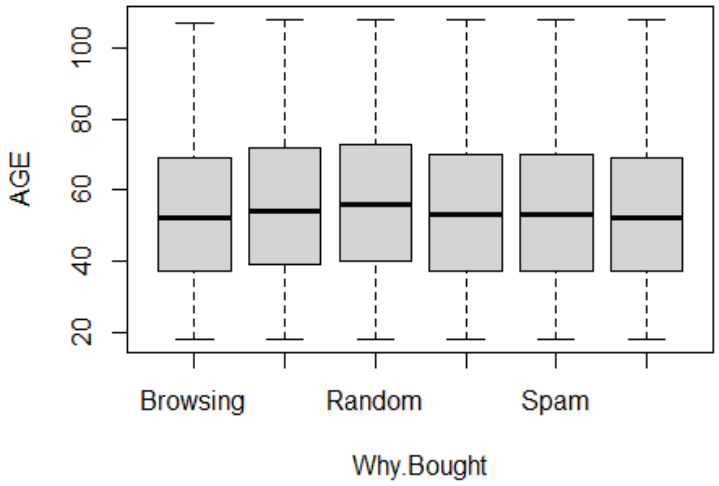


Figure 12 Boxplots: Why customers bought the products compared to age of customers

Age compared to price

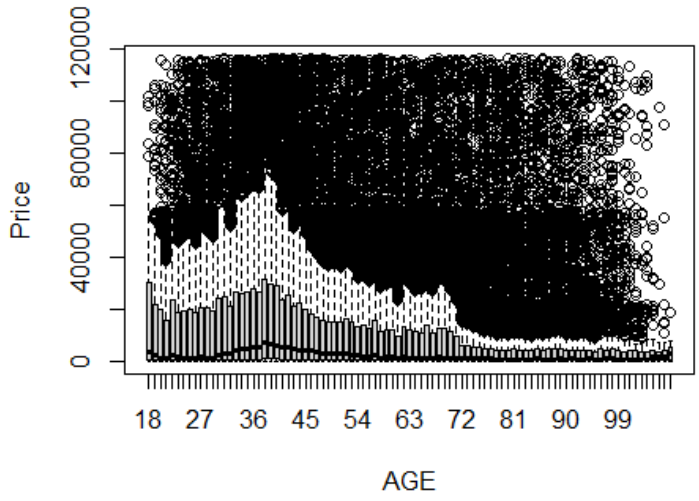


Figure 13 Scatterplot: Age of customers compared to price of products

This graph indicates that people buy products over all products. The age that tends to buy more expensive products are between 25 and 45. The business can use this information to sell more products that people in this age group wants to buy that will generate revenue, given that they are willing to pay more for products.

Range of prices

The graph indicates the range of prices of all the online sales of the company and the number of sales that fall in different price divisions. The graph shows that majority of sales fall into the less expensive range with a skew to the right distribution of the graph. The only information that is useful is that the more of the sales fall into the lower price class. The graph only indicates the distribution in larger groups by grouping the prices into $5000 intervals.

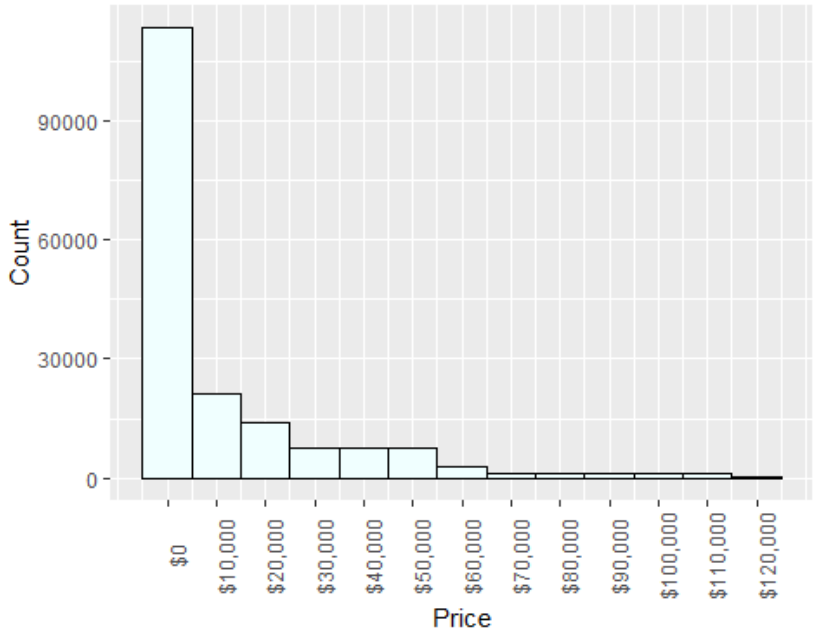


Figure 14 Histogram: Range of prices

Distribution of Prices in the peak are from previous graph

The graph below was taken over the range of prices between $0 and $2000 since the previous graph showed a peak in this area. It shows the distribution within this area and where the majority lies more specifically in the sales. The price range falls more within the less expensive products between $100 and $600.

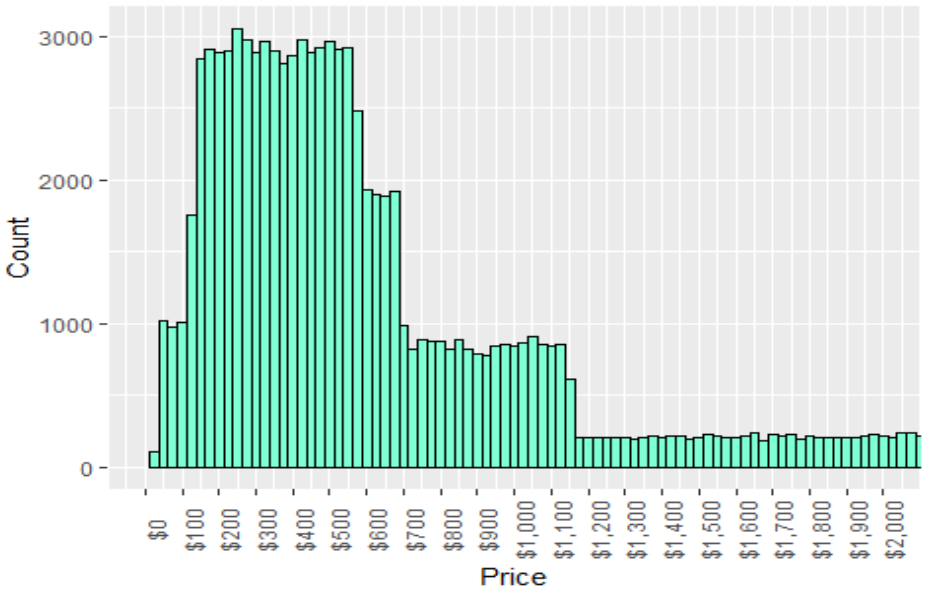


Figure 15 Histogram: Distribution of price at the peak

Features compared to price separately

The following graph shows the distribution of classes over the price range of $0 and $2000 to show what class causes the most sales that generate the income. The three classes that are responsible for the most sales are Clothing, Food and Sweets. This is because they are relatively the less expensive products being sold.

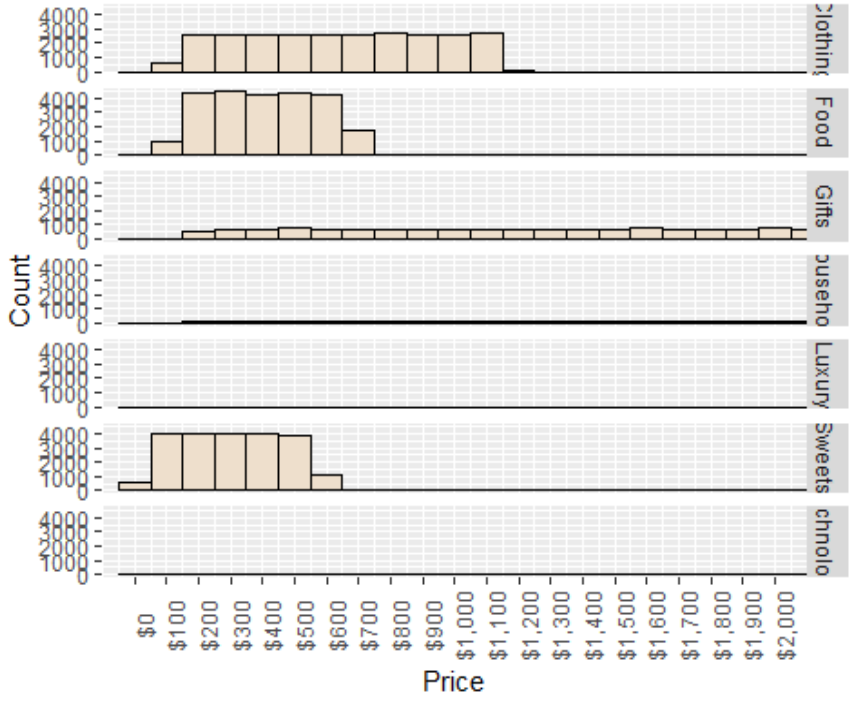


Figure 16 Faceted Histograms: Classes compared to price separately

Distribution of price for different classes

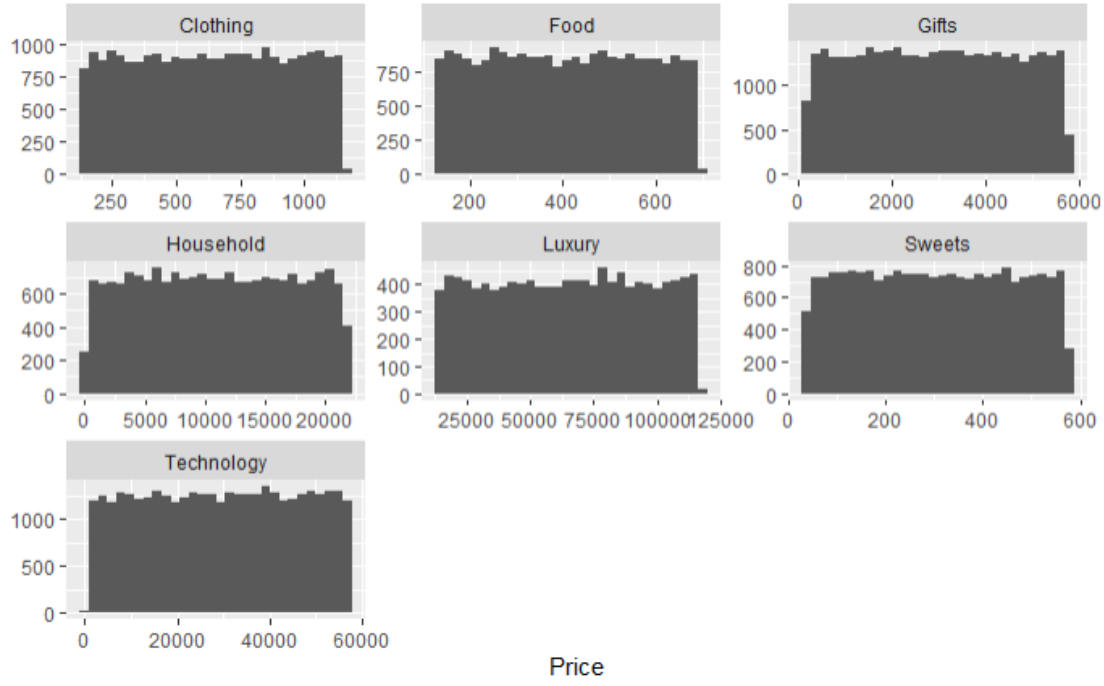


Figure 17 Faceted distribution of class compared to class

The different classes are shown separately in the graphs above and the range of prices The graphs in figure 17 is specifically where the different classes are averaged. This shows where most of the specific class lies, it shows whether on average the money spent on the specific class is very high or very low.

Table 3 Average prices of each class

Class	**Count**	**Average Price ($)**
Clothing	26403	640.53
Food	24582	407.82
Gifts	39149	2961.84
Household	20065	11009.27
Luxury	11868	64862.64
Sweets	21564	304.07
Technology	36347	29508.06

Prices per class

The classes with lowers cost such as the clothing, food and sweets have smaller price and higher quantity sold while products with higher price have lower quantity sold such as household items and gifts the most expensive items have a large price range and sold in much lower quantities. This figure specifically indicates where each class has the most instances.

.

Figure 18 Different classes compared to price

Boxplots of different years in terms of price

The price range distributed over the years are relatively constant. The average price from 2021 to 2029 stays in the same range and from the box plots it is clear that there is variation in the years but only slightly. Relatively the same revenue is generated in the business by the products, and this can be increased to ensure benefits and profitability of the company.

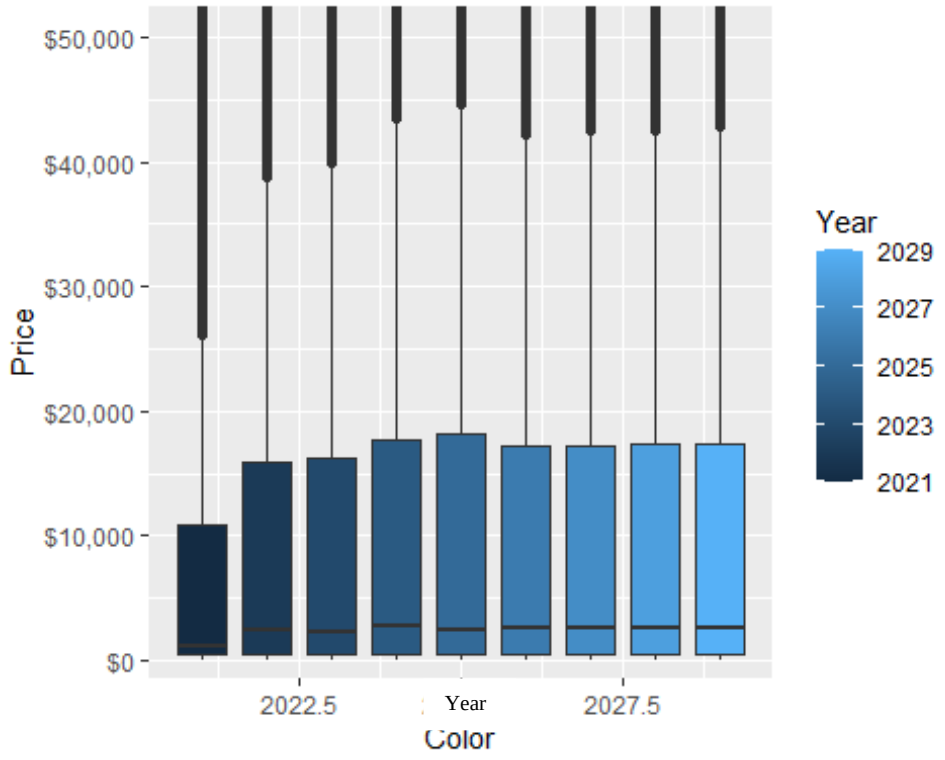


Figure 19 Boxplot: Different years compared to price

Comparison of price and delivery time

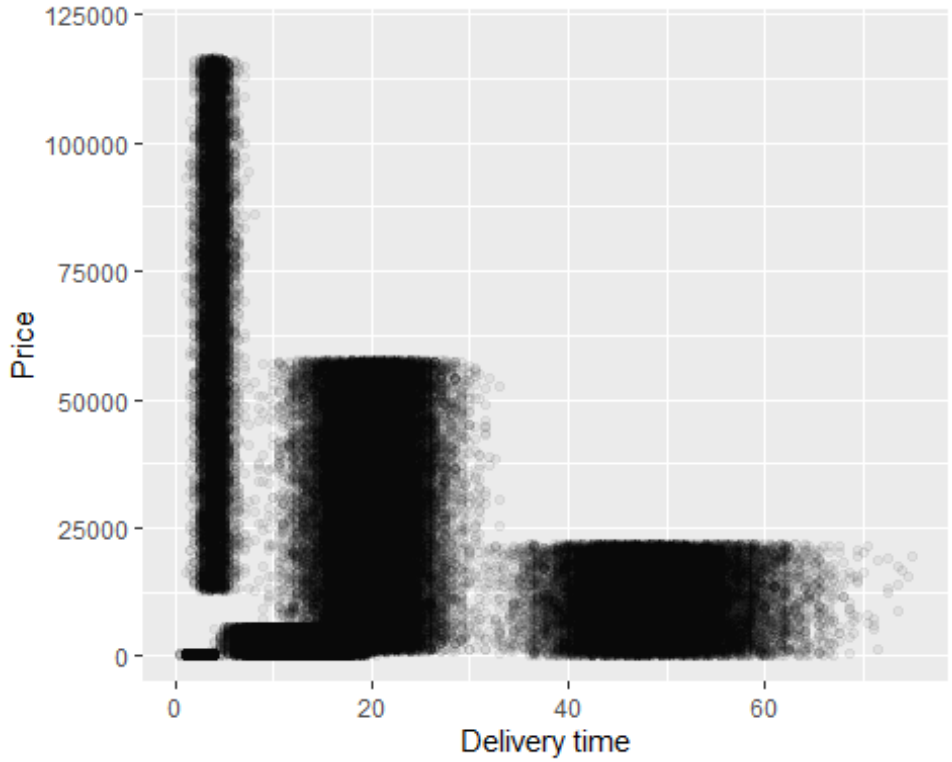


Figure 20 Comparison of price and delivery time in a scatterplot

From *figure 20* it indicates that the products in the range of $0-$20000 has delivery times ranging between 0 and 75 days. Between $20000 and $60000 the delivery time ranges approximately 10 and 20 days. Further for greater than $60000, the more expensive products, have faster delivery times.

This comparison of the two features is to indicate if the price of a product if it is more expensive will have a faster delivery time.

From the graph it can be concluded that the relationship mostly between the price of a product and the delivery time of that product is that the higher the price of the product or more expensive products the shorter the delivery time meaning faster delivery.

Different features compared based on reasons for buying the different classes

The graphs in *figure 21* indicates the exponential distribution between the price of different class of products. This is to indicate the distribution in terms of the number of products in each class and the reason for buying the product. The exponential graphs indicate the upward trends in the different classes as the price increases. From this graph it is clear that the recommendations are big influencers on customers to buy the products. The logarithmic scale of price shows a less severe increase or decrease and allows to visualize the distribution (Investopedia, n.d.).

Figure 21 The log of price compared for all the classes

Delivery time per class in comparison with price

The graph shows the distribution per class in terms of the different delivery times and how this correlate to the price. Luxury has a short delivery time but is more expensive as indicated in the blue. Sweets also have fast delivery and is again less expensive items, this might be because the demand is so high and that it is easily accessible. Household items have longer delivery times and is generally not that expensive. Overall if you look at the price it correlates with the delivery time as follows: More expensive products have somewhat faster delivery times.

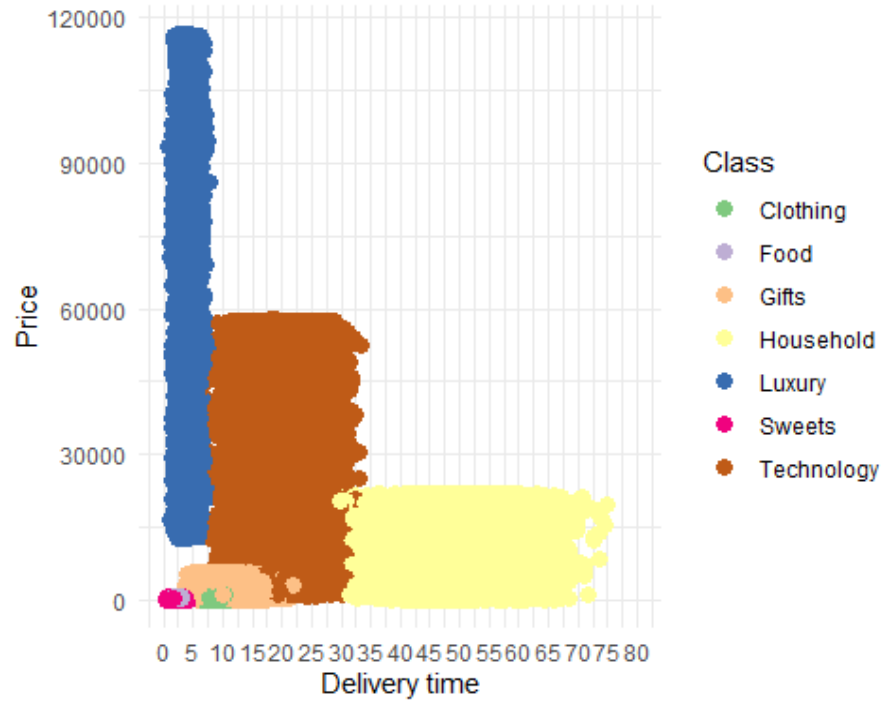


Figure 22 Scatterplot of different classes compared to delivery time

Class and price compared to age in scatter plot

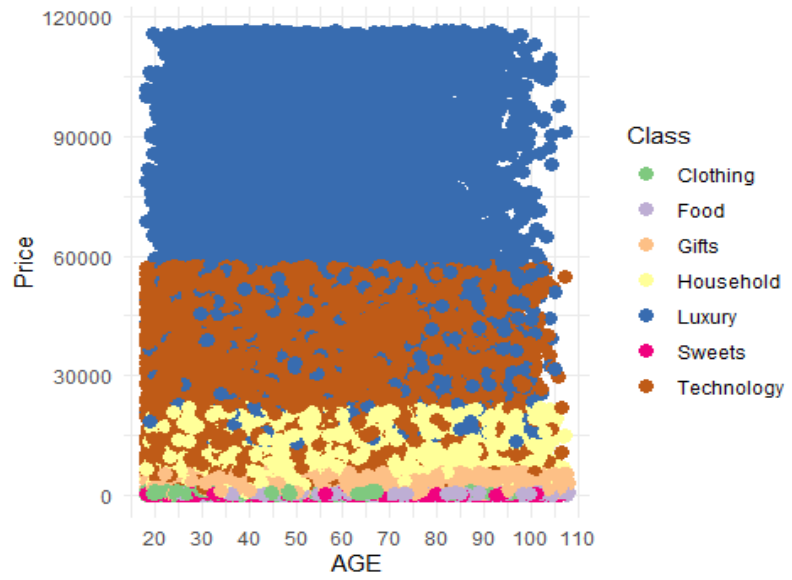


Figure 23 Scatterplot of age compared to price for seperate classes

From the above graph it is clear that most items as generally spread out across the age groups, some classes are more popular under certain age groups than others.

Graph to show the relationship between the different reasons for buying different classes of products

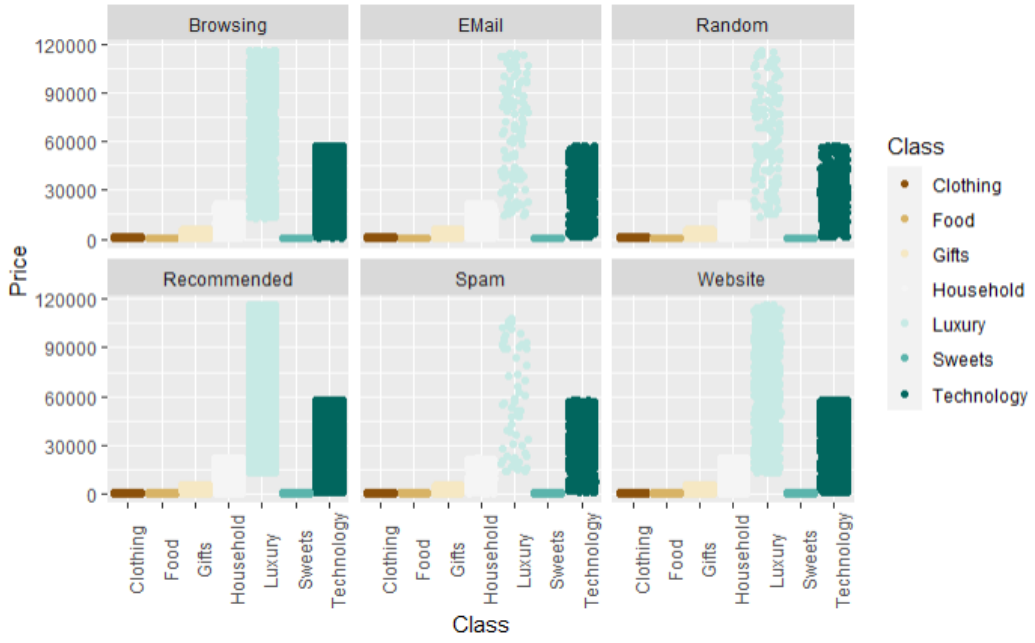


Figure 24 Different reasons why products are bought compared to prices of different classes

The graph shows the different classes and reasons for them being bought along with the price accompanying the different classes. It again shows that the classes are all bought for

different reasons, some reasons such as email and spam were not that influential in convincing someone to buy the product.

Process Capabilities

Table 4 Process capabilities indexes

Index	Equation	Definition
Cp	(USL - LSL)/6σ	Process capability for two-sided specification limits; does not take into account where the process is centered (i.e., what the process average ($\overline{X}$) is).
Cpu	$Cpu = \frac{USL - \overline{X}}{3\sigma}$	Process capability based on the upper specification limit.
Cpl	$Cpl = \frac{\overline{X} - LSL}{3\sigma}$	Process capability based on the lower specification limit.
Cpk	Minimum of Cpu, Cpl	Process capability for two-sided specification limits taking into account where the process is centered.

The formulas in Table were used to calculate the process capabilities indexes Cp, Cpu, Cpl and Cpk respectively. The Cp value is greater than 1 indicates that the spread of the process or width is slightly less than the specification. Since the Cp value is near one it can be seen that the process is almost near the specification in terms of the spread. It means that the process will be able to fulfil the specification. The Cpl of greater than 1 gives an indication that the lower specification limits of the process is capable of being met. The Cpu of less than 1 indicates that the capability of meeting the upper limit specification will possibly not be achieved. The mean is closer to the USL and far from the LSL that means it is not a centred process, which is also a reason why the Cpk value is probably 0.

Table 5 Results for process capabilities indexes

Mean	Standard deviation	Cp	Cpl	Cpu	Cpk
20.012	3.502	1.142	1.905	0.380	0.380

The distribution over the different delivery times of technology indicates that it is relatively centred around the mean, but also very close to the USL.

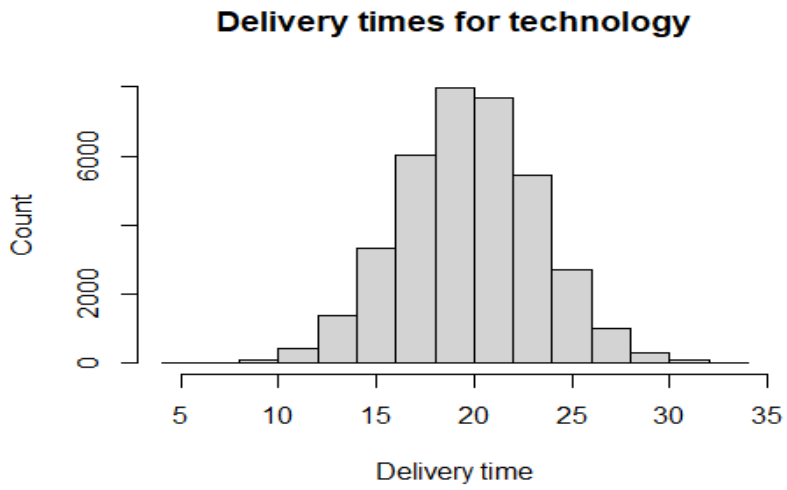


Figure 25 Histogram of the distribution of delivery time for technology

Part 3: Statistical Process Control

Statistical process control or SPC is defined as the using of statistical methods and techniques to control a process or a production method *(ASQ,2022*). The tools and procedures can help to monitor the behaviour or discover issues in the systems internally and find solutions to the production issues.

The Valid sales data should first be ordered in chronological order to represent present data. The data is grouped into groups of 15 The first 30 samples of 15 each was used to develop control charts in 3.1. The SPC values serve as the critical values that will be used in 3.2 and in part 4.1 to control the process and to also identify whether there are processes that are not in control.

3.1 First 30 samples

The S-chart is a control chart that is used to monitor the process variability or standard deviation when the samples are larger than 5. Each point on the chart ultimately represents one of the samples

Table 6 S-Chart for 30 samples

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.181	4.552	3.924	3.296	2.667	2.039	1.410
Clothing	0.867	0.761	0.656	0.551	0.446	0.341	0.236
Household	7.344	6.453	5.563	4.672	3.781	2.890	2.000
Luxury	1.511	1.328	1.145	0.961	0.778	0.595	0.411
Food	0.437	0.384	0.331	0.278	0.225	0.172	0.119
Gifts	2.246	1.974	1.701	1.429	1.157	0.884	0.612
Sweets	0.835	0.734	0.633	0.531	0.430	0.329	0.227

The X-chart indicates the average change over time. It is used to process the sample means in set intervals from the process. The main purpose is to monitor the process variables.

Table 7 X-Chart for 30 samples

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22.975	22.108	21.241	20.374	19.508	18.641	17.774
Clothing	9.405	9.260	9.115	8.970	8.825	8.680	8.535
Household	50.248	49.020	47.791	46.562	45.334	44.105	42.876
Luxury	5.494	5.241	4.988	4.736	4.483	4.230	3.977
Food	2.709	2.636	2.563	2.490	2.417	2.344	2.271
Gifts	9.489	9.113	8.737	8.361	7.985	7.609	7.234
Sweets	2.897	2.757	2.618	2.478	2.338	2.198	2.059

The samples within the first 30 that was out of control was not used within the plotting the SPC values. This could create the wrong SPC values if it is included.

Technology

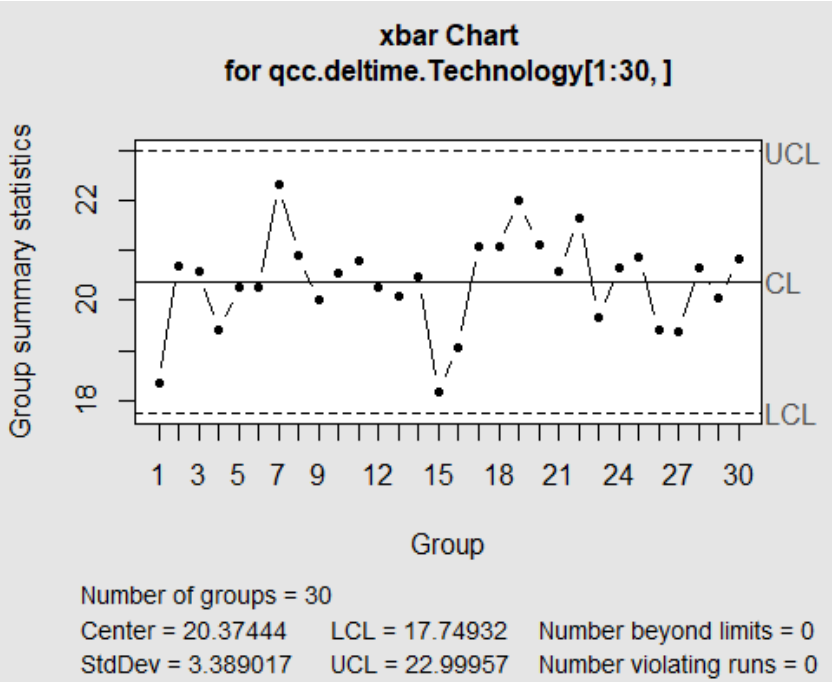


Figure 26 X-chart of Technology

There is a lot of variation in the technology xbar plot. This can be because there is a high standard deviation seen in the s chart. The process is within limits and in control even though the plot has variation.

Clothing

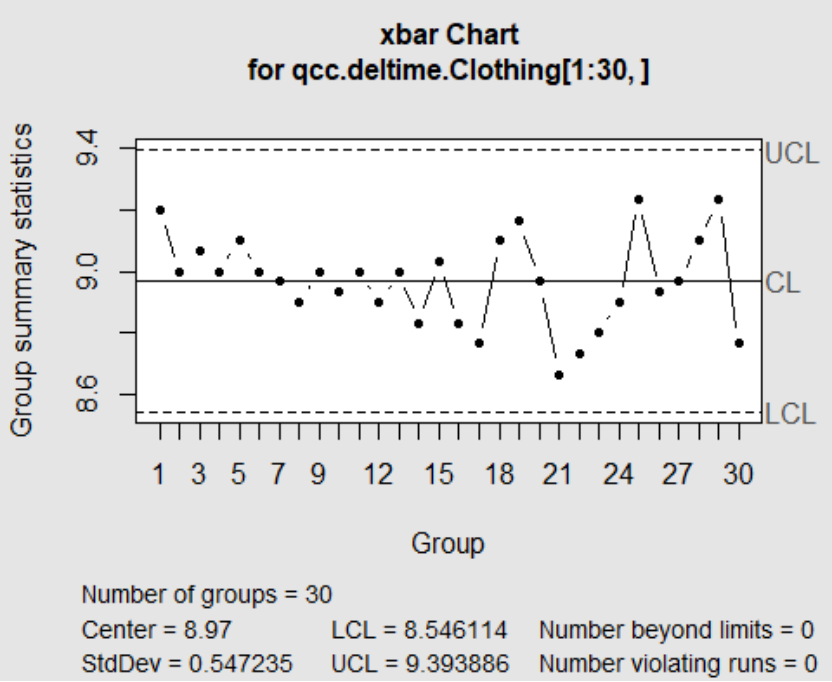


Figure 27 X-chart of Clothing

The s chart for clothing has a low standard deviation and the point are relatively close to the centre line. In the x chart some of the samples have high variation from the centreline but all of them are close to the centreline with a few outliers which means it is relatively stable.

Household

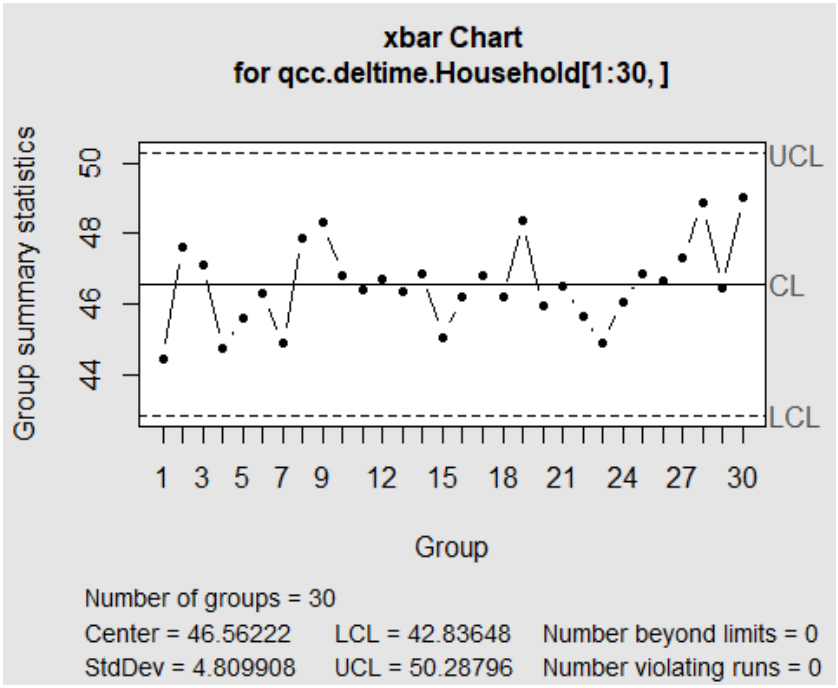


Figure 28 X-chart of Household

There is again high standard deviation from the s plot which indicates a lot of variation. The s chart also shows a sample that is close to the upper limit. The x bar plot has variation as a result. 86

Luxury – check if it got replotted

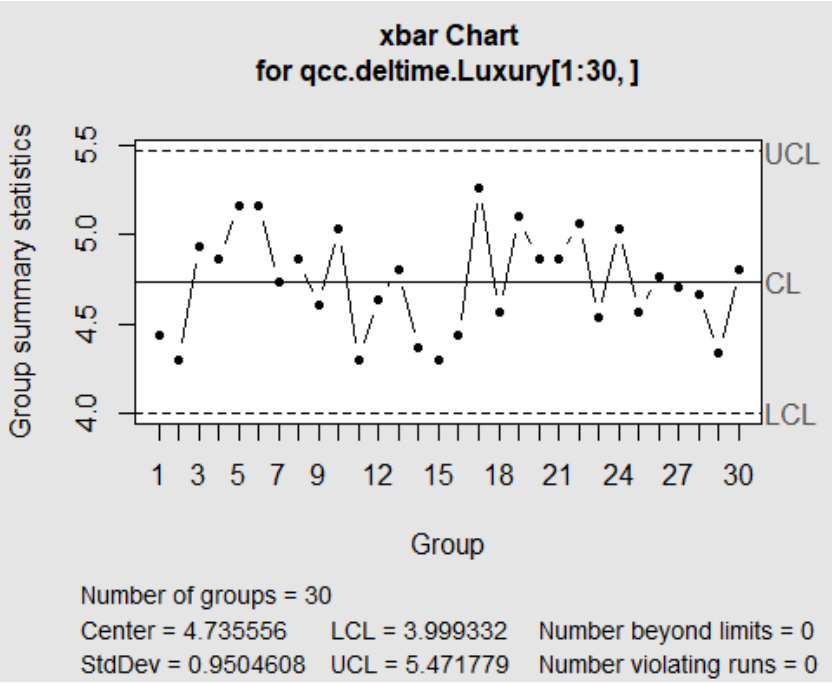


Figure 29 X-chart of Luxury

Luxury has a relatively low standard deviation on the s plot but has some variation on both plots. The x plot however has variation but is relatively stable as it does not go too far from the centreline.

Food

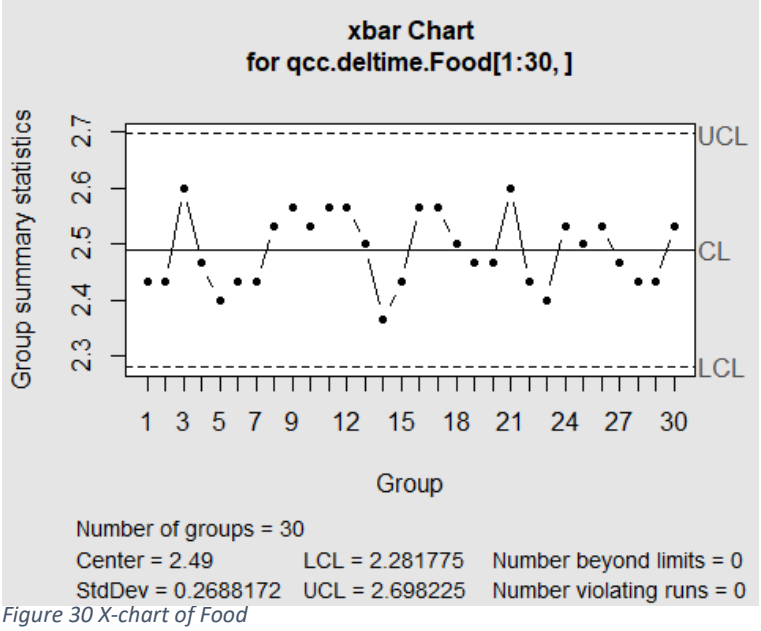


Figure 30 X-chart of Food

Food has a low standard deviation according to the s plot but has an outlier that is not between the limits. The x plot however is stable and close to the centreline.

Gifts

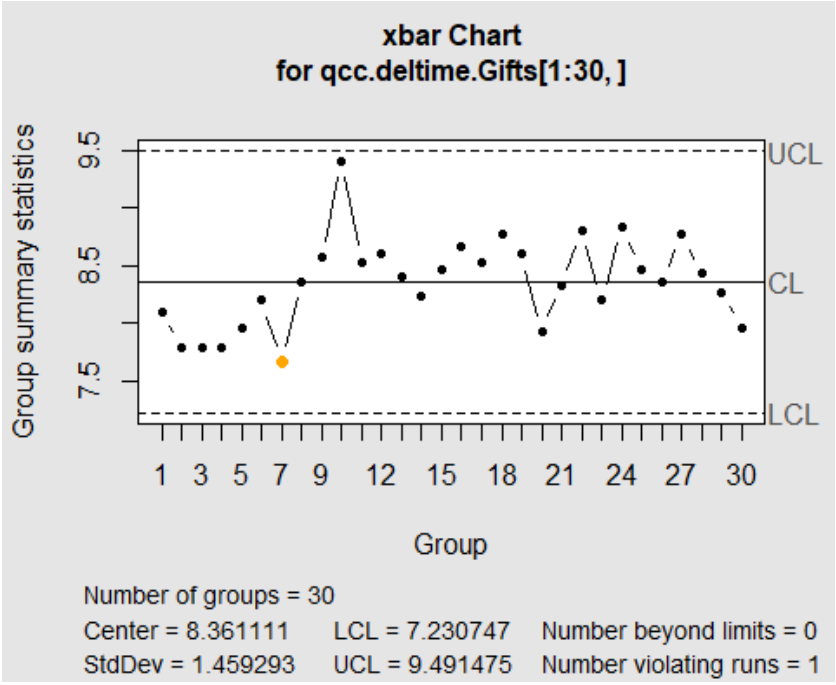


Figure 31 X-chart of Food

The standard deviation for gifts is not that high but the x plot shows that relatively all the points are close to the centreline with a few outliers. All the points are near the centreline, but the few outliers can cause the stability of the process to deviate.

Sweets

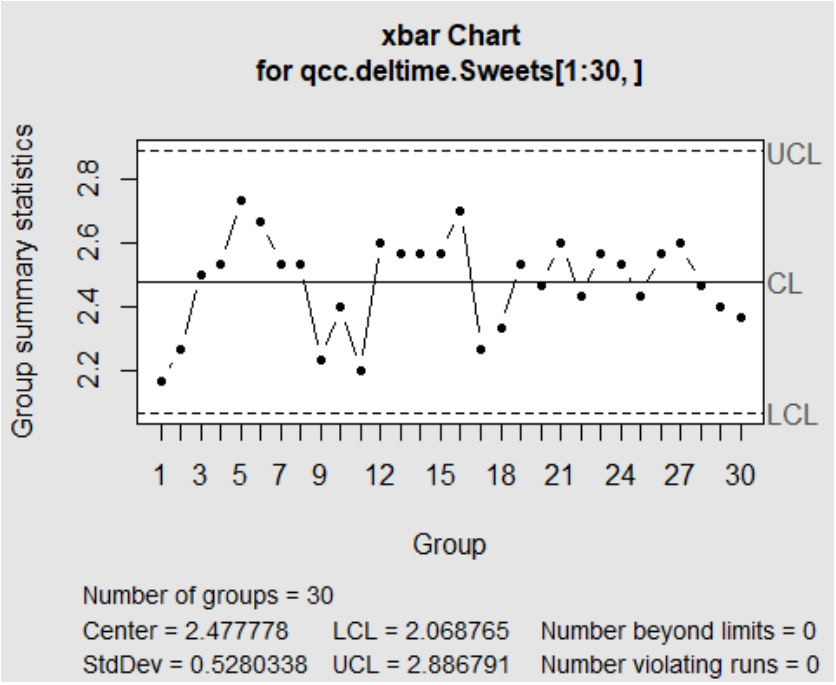


Figure 32 X-chart of Sweets

There is a low standard deviation in the s plot with a few points that are near or outside the limits. This is maybe the reason for the variability in the x chart but since it is still within the limits and mostly near the centreline, it is stable.

3.2 Control of the process for samples 31 onwards

Technology

The process average increased with more samples over time. There are a few samples that are outside of the UCL and LCL which makes this process out of control

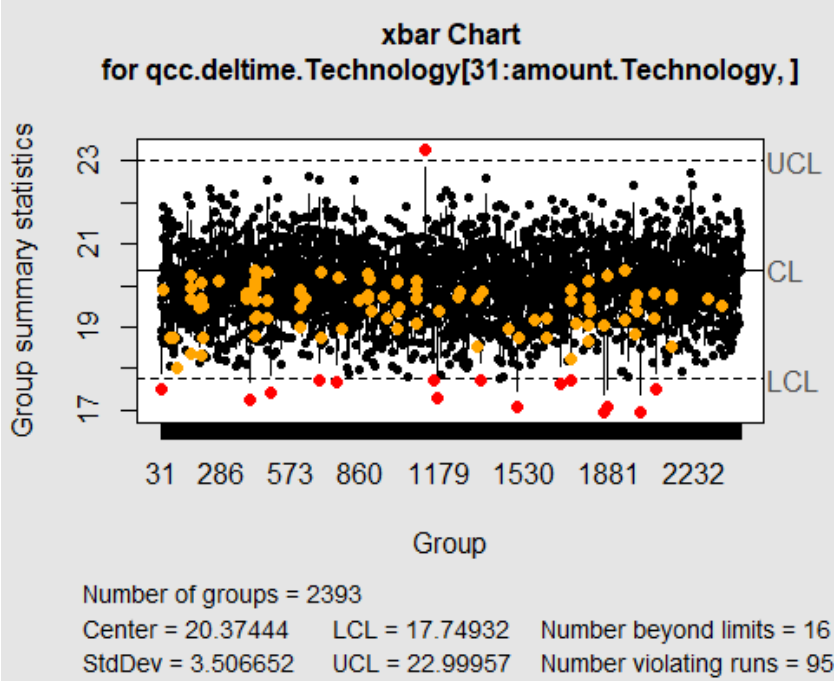


Figure 33 X-chart for Technology sample 31 onwards

Clothing

Clothing started stable with a few outliers but later it started to deviate further with more outliers outside the limits. It can be that there are problems that arises later like unreliable delivery to customers, and this makes the system unstable.

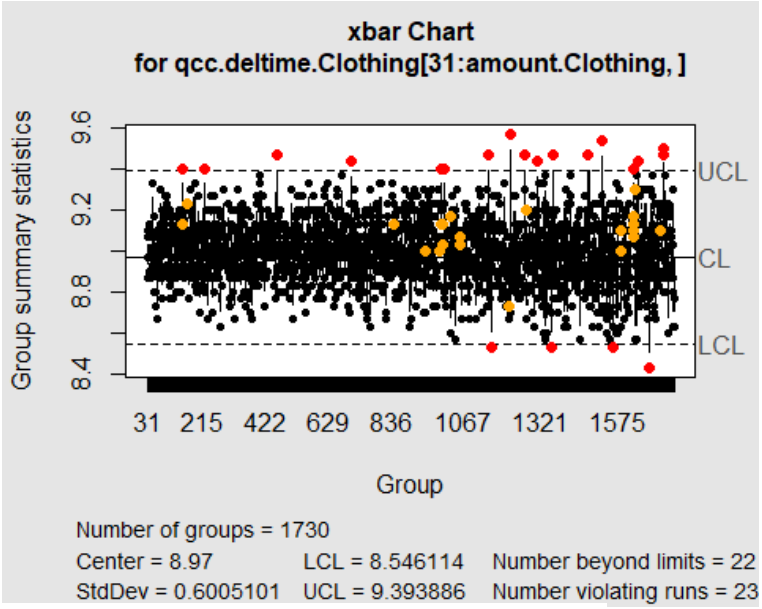


Figure 34 X-chart for Clothing sample 31 onwards

Household

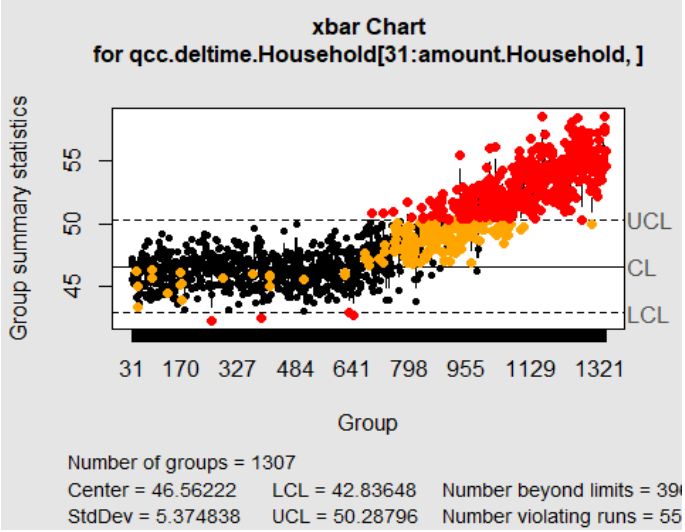


Figure 35 X-chart for Household sample 31 onwards

The Household chart is stable at first and increases rapidly far past the upper limit making this system very unstable and out of control. The delivery time increases over time for household products

Luxury

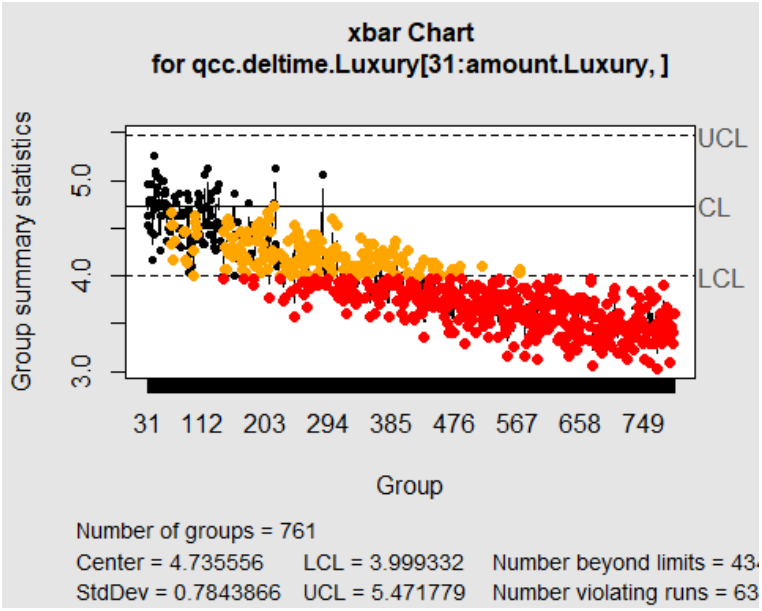


Figure 36 X-chart for Luxury sample 31 onwards

There is a slight downward that is noticed at first and then it starts decreasing significantly. The system goes out of control as it moves past the lower limit. This means faster delivery times which is not necessarily bad for the business.

Food

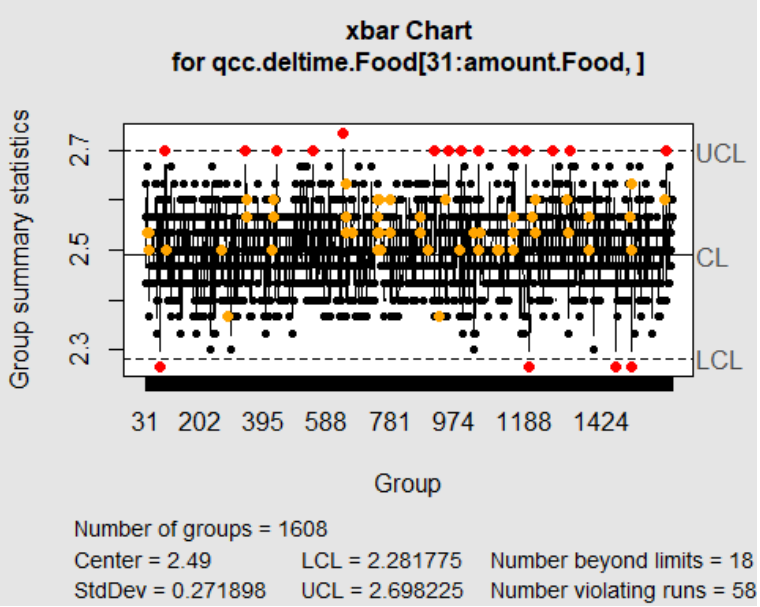


Figure 37 X-chart for Food sample 31 onwards

The food data is relatively stable with a few outliers, but this does not make the system entirely out of control since more of the point are within the control limits.

Gifts

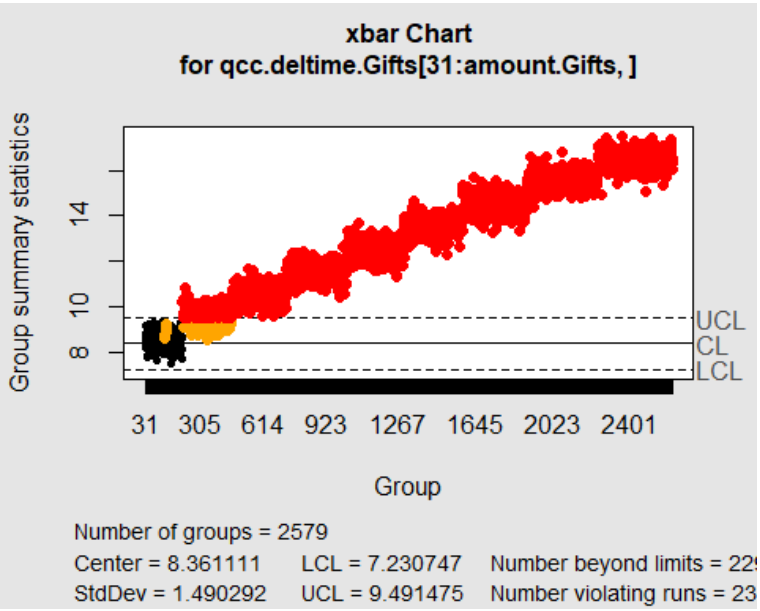


Figure 38 X-chart for Gifts sample 31 onwards

This chart is stable at first and rapidly increases with time. This indicates that the delivery time rapidly increases over time and makes the process unstable. The delivery time over the samples have a growing trend which can be an indication that the Type I error cannot be calculated.

Sweets

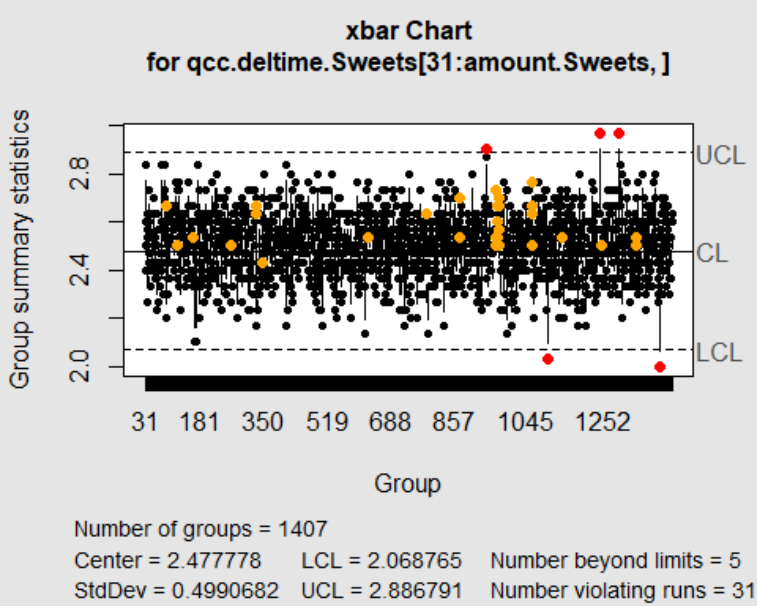


Figure 39 X-chart for Sweets sample 31 onwards

The sweets are stable throughout with a few outliers in the last part, this does not cause the system to be unstable or out of control.

Part 4: Optimizing the delivery process

4.1 Processing and inspecting of X Charts

Table 8 X bar samples outside of the outer control limits

Class	Total Found	1st	2nd	3rd	3rd Last	2nd Last	Last
Technology	16	37	398	483	1872	2009	2071
Clothing	22	148	217	455	1677	1723	1724
Household	396	252	387	629	1335	1336	1337
Luxury	434	142	171	184	789	790	791
Food	18	75	93	338	1467	1515	1621
Gifts	2290	213	216	216	2607	2608	2609
Sweets	5	942	1104	1243	1294	1403	-

Larger number of outliers can be found in Household, Luxury and especially the Gift class of products. There is large variation in the process. The areas of improvement should be identified so that the problems of outliers can be decreased.

Table 9 Samples between -0.3 and 0.4 sigma

Class	Maximum between sigma length	Last sample position of first	Last sample of Last
Technology	6	1523	1888
Clothing	191	701	701
Household	0	0	0
Luxury	11	28	28
Food	1638	1638	1638
Gifts	10	75	75
Sweets	163	164	164

This inspection of the charts can show where the processes are in control. Taking the maximum in each class within the sigma range is not fully enough information to conclude to whether this makes the system in control. This does not show an accurate representation of

the whole process as it can show a stable interval over the sigma range but leaves out other unstable and out of control instances that can cause the process to be out of control.

4.2 Probability of making a Type I error

Type I error is when the SPC indicated the process is not fine when it in reality is fine. It can cause that products be rejected and a loss of profit because the products are unnecessarily rejected.

Table 10 X bar samples outside of the outer control limits

Class	Total Found	Probability of Type I error
Technology	16	0
Clothing	22	3.09e-57
Household	396	0
Luxury	434	0
Food	18	5.81e-47
Gifts	2290	0
Sweets	5	1.43e-13

Probability of sample being outside UCL and LCL = 1.43e-13

Table 11 Samples between -0.3 and 0.4 sigma

Class	Maximum above centreline length	Probability of Type I error
Technology	2	4.17e-04
Clothing	191	2.56e-108
Household	5	0
Luxury	11	6.36e-07
Food	0	0
Gifts	10	2.33e-06
Sweets	163	1.52e-92

Probability of sample being between -0.3 and +0.4 sigma limits = 0.0417

4.3 Optimize delivery process

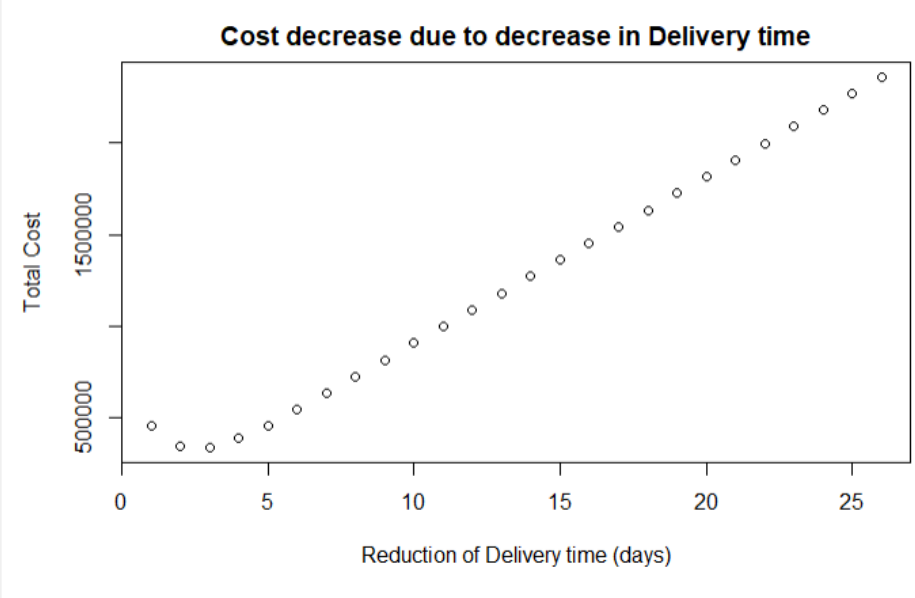


Figure 40 Cost decreases due to decrease In Delivery time

The overall cost can be reduced when the delivery time is decreased. As shown in *figure 40 as* the delivery time is decrease by more days it has a significant impact on the price of the products. To keep the prices of more expensive products as efficient as possible it should be aimed to reduce the amount of time for delivery. This means getting the products to the customers quicker will reduce the costs of the products. Decreasing the delivery time too much can have a negative impact on the costs meaning it can have additional costs that are not to the advantage of the business. The optimal number of days should thus be found to ensure that the costs are minimised for the business.

The decrease in delivery time as indicated by *figure 40* has a positive effect on the profit of the business. As the days are reduced more profit can be made. When the reduction of days passes 3 days the profit starts to decrease again.

4.4 Likelihood of making a type II error

The probability of making a Type II error is when H0 is not rejected and H0 is indeed not true and should have been rejected. This is very important because it means that the process might be seen as being in control when it is in fact out of control. In the instance A it would mean that the sample means are withing the specified control limits when it is not. For B it would mean that there are no large consecutive delivery times that fall within the bounds of the -0.3 and 0.4 sigma limits, but there are a large consecutive group that falls within this limit. This can only be done when the process is not in control, or the variation is very large.

UCL= 22.974616
LCL=17.774

Sd=(UCL-LCL)/6

Normal probability of UCL-Normal probability of LCL

=0.4883183-0.0000000008234293

P(Type II error)= 0.4883183 =48.83%

The probability of making this error is 48.83%

Part 5: DOE and MANOVA

The common goal of the multivariate analysis of variance or MANOVA is to determine whether multiple levels of independent features on their own or in combination with one another have an effect on the dependent features *(Statistics Solutions,2021)*. The dependent variables need to meet the parametric requirements. Two hypothesis tests were done using the MANOVA that was set up and will now be discussed.

Hypothesis 1: Test whether the delivery time and price differ for the different product classes.

A hypothesis is formed that the average delivery time and the price for each class are alike or similar. The impact on the company can be that the reliability of the delivery can be impacted if it is assumed that the time of delivery is the same. The priorities of the company will then be equal for all classes of products which also decreases the loyalty of customers.

Ho: The class of the product does not have an influence on the delivery time and product.

Ha: The class of the product does have an influence on the delivery time and product.

Table 12 Average delivery times and prices for different classes

Class	Number of customers who bought this class	**Average delivery time**	**Average price ($)**
Technology	36347	20.011	29508.063
Clothing	26403	9.000	640.525
Household	20065	48.720	11009.274
Luxury	11868	3.972	64862.639
Food	24582	2.502	407.815
Gifts	39149	12.891	2961.841
Sweets	21564	2.501	304.063

By inspecting the MANOVA table the p value is exceedingly small. This gives an indication that the average price and delivery time for the different classes are not similar. The H0

hypothesis is rejected as the impact of the average price and mean are not the same for the different classes given that they are not the same values for each product.

The conclusion is that the company cannot promise the same delivery time for each product that they sell to the product and should base this of the price of the class but also state it to the customer that the delivery time differs for the different price products.

The following graph is to validate whether the hypothesis is true and plots the different boxplots to show distribution of the delivery time and price per class next to each other. From this we can see that the hypothesis is true that the class of the product has an influence on the delivery time and price. The average for each class is different as shown in the graph.

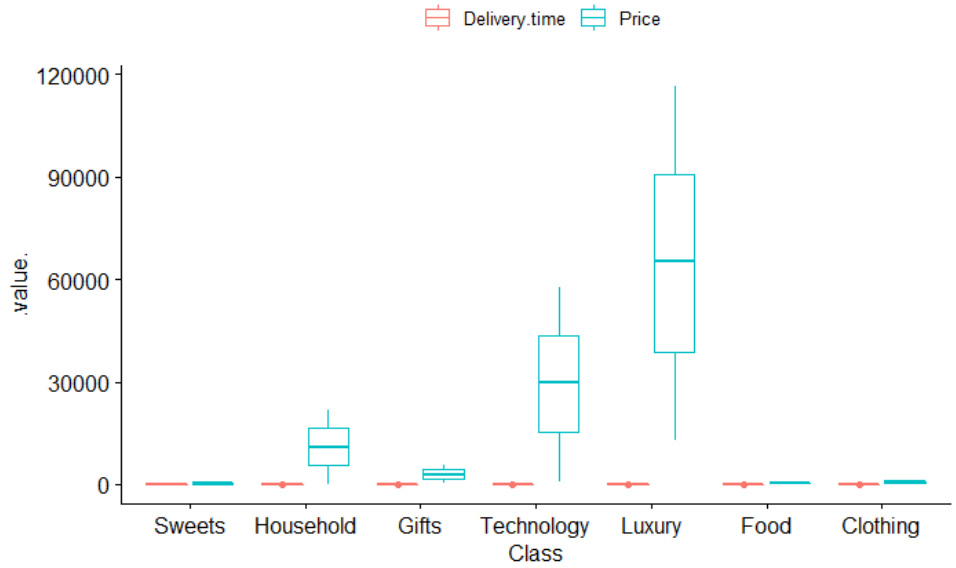


Figure 42 Boxplots: Comparing averages of different classes

Hypothesis 2: Certain ages of people prefer some class of products over other products.

The hypothesis is stated to see whether the company should focus on a certain age group for a specific class of products because it is dominantly bought by this age group of people. The average age of customers will be taken to investigate the hypothesis

Ho: The age of the customers does not have an influence on the class of products they buy.

Ha: The age of the customers has an influence on the class of products they buy.

Table 13 Average age of customers who buy certain class of products

Class	Number of customers who bought this class	Average age of customers
Technology	36347	46.644
Clothing	26403	47.470
Household	20065	51.927
Luxury	11868	51.339
Food	24582	65.371
Gifts	39149	60.826
Sweets	21564	57.153

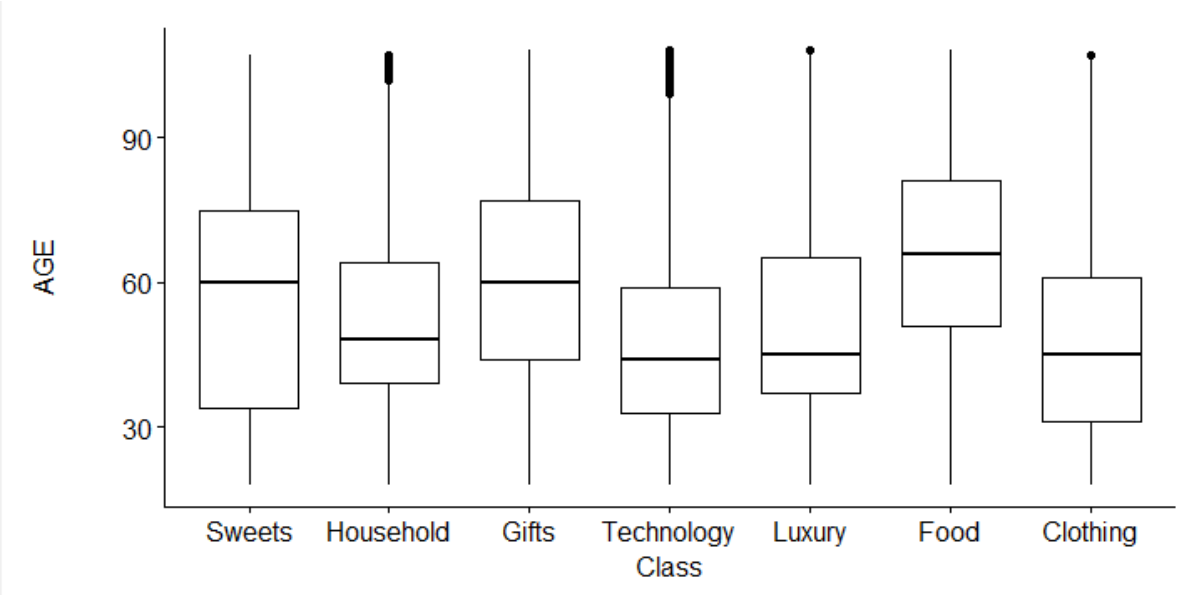


Figure 43 Boxplots: Different classes compared by average age

Again, p value is small which means in this context that the age of the customer does not have an influence on the class of products they buy and the H0 is rejected.

From the graph all the products are distributed over most of the age groups and around the averages. It is not clear that the age of the customer influences their choice of the class of product they buy or whether an age group is responsible for a dominant part of the sales of a product group.

Part 6: Reliability of the service and products.

6.1 Reliability of service and products

Problem 6 p359 11th edition

Specification: 0.06 ± 0.04 cm

Cost to scrap: \$45

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

$45 \;\; = \;\; k\,(0.04)^2$

$k \;\; = \$28\,125$

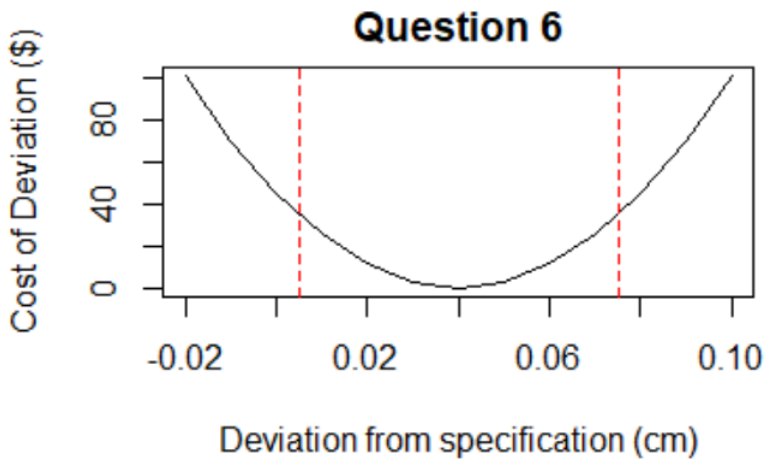


Figure 44 Question 6 loss function plotted

Problem 7 p359 11th edition

Scrap cost reduced to \$35 per part

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

$35 \;\; = \;\; k\,(0.04)^2$

$k \;\; = \$21\,875$

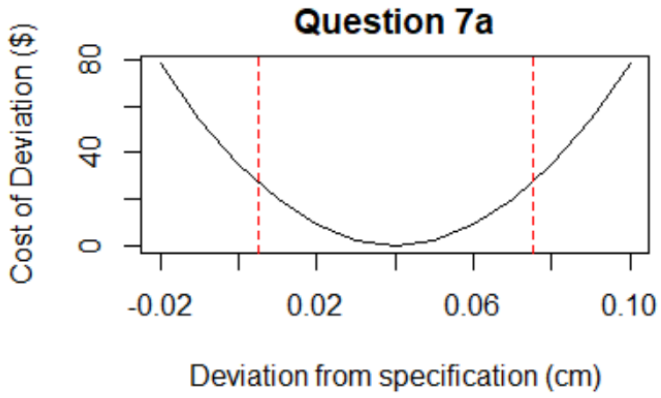


Figure 45 Question 7a loss function plotted

b) Process deviation reduced to 0.027 cm

$L(x) = 21\,875\,(0.027 - 0.04)^2$

$= \$3.70$

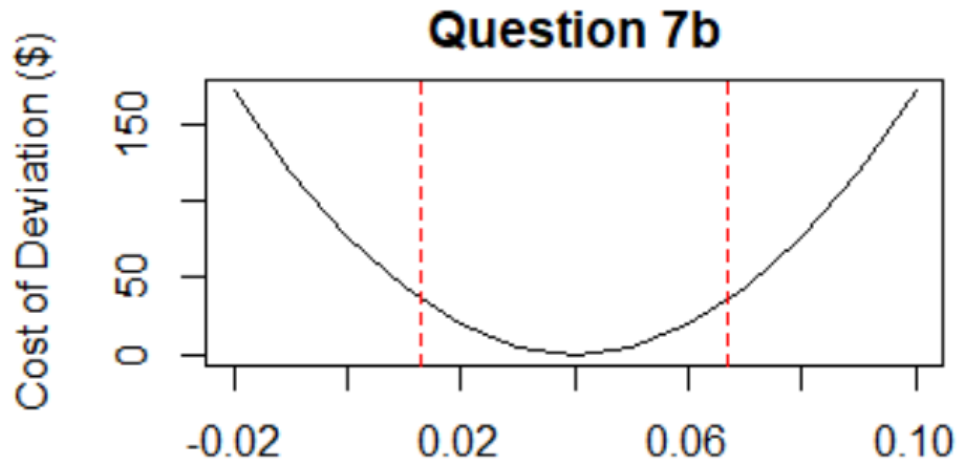
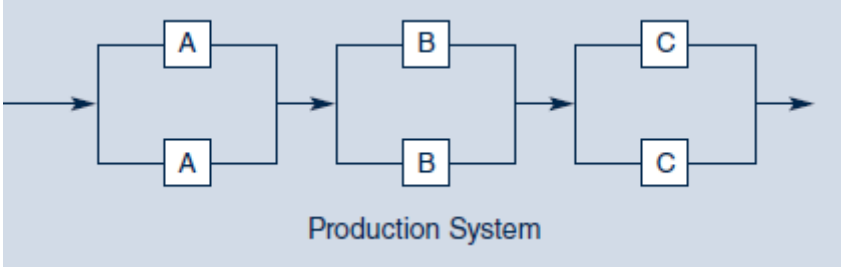


Figure 46 Question 7b loss function plotted

6.2 System reliability



Machine	Reliability
A	0.85
B	0.92
C	0.90

a.) Only one works at each stage thus the machines are in series

$R_s = R_A \times R_B \times R_C$

$= 0.85 \times 0.92 \times 0.90$

$= 0.7038$

The reliability is 70.38%

b.) Two machines in parallel at each stage. The two machines together

$R_p = (1-(1-R_A)\,x(1-R_A)) \times (1-(1-R_B)\,x(1-R_B)) \times (1-(1-R_C)\,x(1-R_C))$

$R_p = (1-(1-0.85)\,x(1-0.85)) \times (1-(1-0.92)x(1-0.92)) \times (1-(1-0.90)x(1-0.90))$

$=0.9615$

The reliability increases to 96.15% which is a major increase in the reliability when there are two of the same parts in parallel at each of the three stages.

6.3 Binomial properties

Table 14 Available days of vehicles

Vehicles (/20)	Days available (/1560)
20	190
19	22
18	3
17	1

Table 15 Available days of drivers

Drivers (/21)	Days available (/1560)
20	95
19	6
18	1

Assume the following for the rest of the 1560 days:

-21 Vehicles will be available

-21 Drivers will be available

For the process to be reliable we need at least 19 vehicles available and for the 19 vehicles to be feasible 19 drivers need to be available.

The binomial distribution will be used to calculate the probability of a reliable process in order to calculate the number of days:

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

Part 1: For 21 vehicles and 21 drivers

Vehicles:

P(21 vehicles)=$\frac{1560-190-22-3-1}{1560}=\binom{21}{0} \times p^0 \times (1-p)^{21-0}$

p= 0.007071612

P(20 vehicles)= $\frac{190}{1560} = \binom{21}{1} \times p^1 \times (1-p)^{21-1}$

p= 0.006643289

P(19 vehicles)= $\frac{22}{1560} = \binom{21}{2} \times p^2 \times (1-p)^{21-2}$

p= 0.008927524

P(18 vehicles)= $\frac{3}{1560} = \binom{21}{3} \times p^3 \times (1-p)^{21-3}$

p= 0.01217067

P(17 vehicles)= $\frac{1}{1560} = \binom{21}{4} \times p^4 \times (1-p)^{21-4}$

p= 0.01967464

Weighted p = $\frac{1344(0.007071612\ )+190(0.006643289)+22(0.008927524)+3(0.01217067)+1(0.01967464)}{1560}$

p = 0.007063502

P(0 fail) =$\binom{21}{0} \times 0.007063502\ ^{0} \times (1 - 0.007063502\ )^{21-0}$=0.8616898
Expected number of days=0.8616898*1560 =1344.236 days

P(1 fail)= $\binom{21}{1} \times 0.007063502\ ^{1} \times (1 - 0.007063502\ )^{21-1}$=0.1287268
Expected number of days=0.1287268*1560 =200.838 days

P(2 fail)= $\binom{21}{2} \times 0.007063502\ ^{2} \times (1 - 0.007063502\ )^{21-2}$ =0.009157301
Expected number of days=0.009157301x 1560 =14.28539 days

P(3 fail)= $\binom{21}{3} \times 0.007063502\ ^{3} \times (1 - 0.007063502\ )^{21-3}$=0.0004125708
Expected number of days=0.0004125708x 1560 =0.6436104 days

P(4 fail)= $\binom{21}{4} \times 0.007063502\ ^{4} \times (1 - 0.007063502\ )^{21-4}$=0.00001320717
Expected number of days=0.00001320717x 1560 =0.02060318 days

Theoretical number of days calculated:
Expected percentage of reliable days= $\frac{1268.76+263.48+26.05}{1560}$ = 0.9995739x 100
=99.95% reliability
Expected number of days=365 x 0.9995=364.59 $\approx$ 364 *days at least reliable*
364.9998814

Drivers:
P(21 drivers)= $\frac{1560-95-6-1}{1560} = \binom{21}{0} \times p^{0} \times (1-p)^{21-0}$
p= 0.003229808
P(20 drivers)= $\frac{95}{1560} = \binom{21}{1} \times p^{1} \times (1-p)^{21-1}$
p=0.003088413
P(19 drivers)= $\frac{6}{1560} = \binom{21}{2} \times p^{2} \times (1-p)^{21-2}$
p=0.004473441
P(18 drivers)= $\frac{1}{1560} = \binom{21}{3} \times p^{3} \times (1-p)^{21-3}$
p=0.008258873
Weighted p =$\frac{1458(0.003229808\ )+95(0.003088413\ \ )+6(0.004473441\ )+1(0.008258873\ )}{1560}$ = 0.003229204

P(0 fail)= $\binom{21}{0} \times 0.8772^{0} \times (1 - 0.8772)^{21-0}$= 0.9343324
Expected number of days= 0.9343324*1560 =1457.559days

P(1 fail)= $\binom{21}{1} \times 0.8772^{1} \times (1 - 0.8772)^{21-1}$= 0.06356541
Expected number of days= 0.06356541*1560 =99.16205 days

P(2 fail)= $\binom{21}{2} \times 0.8772^{2} \times (1 - 0.8772)^{21-2}$= 0.002059307
Expected number of days= 0.002059307*1560 =3.212519 days

P(3 fail)= $\binom{21}{3} \times 0.8772^{3} \times (1 - 0.8772)^{21-3}$= 0.00004225261
Expected number of days= 0.0000417226*1560 =0.06591408days

Expected percentage of reliable days=$\frac{1457.98+98.76+3.19}{1560}$= 0.9999571x 100
=99.99% reliability.

Expected number of days=365 x 0.9999571 =364.994≈ 364 *days at least reliable*

Vehicles and drivers:

P(Reliable) = P(Vehicles reliable) x P(Drivers reliable)
= 0.9995739 x 0.9999571
= 0.999531 = 99.95 % reliability
Days: 0.999531 x 365 = 364.8288=364 days

Part 2: For 22 vehicles and 21 drivers

Vehicles:

Additionally, 1 more vehicles can fail and still be a reliable process

P(0 fail) =$\binom{22}{0} \times 0.007063502^{\,0} \times (1 - 0.007063502)^{22-0}$= 0.8556033

P(1 fail)= $\binom{22}{1} \times 0.007063502^{\,1} \times (1 - 0.007063502)^{22-1}$= 0.1339041

P(2 fail)= $\binom{22}{2} \times 0.007063502^{\,2} \times (1 - 0.007063502)^{22-2}$= 0.01000188

P(3 fail)= $\binom{22}{3} \times 0.007063502^{\,3} \times (1 - 0.007063502)^{22-3}$= 0.0004743392

P(3 fail)= $\binom{22}{4} \times 0.007063502^{\,4} \times (1 - 0.007063502)^{22-4}$= 0.00001602807

P(Reliable) =0.8556033 + 0.1339041 +0.01000188+0.0004743392
= 0.9999836

Drivers:

Stays the same as in previous part thus a reliability of 0.999957666 or 99.99%. No addition or changes made to the drivers of the company.

Vehicles and drivers:

P(Reliable) = P(Vehicles reliable) x P(Drivers reliable)
= 0.9999836x 0.999957666
= 0.9999407= 99.99% reliability
Days: 0.999899271 x 365 = 364.9783=364 days

The Number of reliable days is still 364 days with an increased reliability from 99.95% to 99.99%. The addition of another vehicle increases the reliability because of the increase in the probability of having more than 19 drivers available on a given day.

Conclusion

The company has a variety of customers that have interest in different products and range from a wide variety of ages. The sales for certain products, namely Technology and Luxury, are in the higher frequency than others, these classes of products are bought more frequently by customers and are responsible for the biggest part of the ultimate profit of the business. The identification and analysis of these products are very important to optimize and maximize profits. It gives an indication of where the focus should be shifted and how business goals like an optimal service level can be achieved.

The loyalty of customers is reflected in in the fact that most of the sales are centred around recommendations. This shows that customer satisfaction is achieved and should be aimed to fulfil in the future. Some of the classes are more stable than other classes of products. These classes are Technology, Clothing, Food and Sweets even though there is some variation in the processes as indicated by the control charts. The out-of-control processes should be managed and controlled.

The relationship between the class and delivery time also different for different products. This can be seen when investigating products in terms of price. In general products with higher prices tend to have shorter delivery times. This can contribute to an increase in profits as the delivery time is decreased for more reliable delivery to customers.

The Type I errors made are very small and this can conclude that the outcomes are indeed accurate. The calculations of the Type II error are significantly larger and should be monitored as the process goes along. As the assets that are used in delivery are increase the reliability of the reliable days are also increased. This can also lead to higher profits in the long term and seen as good investments.

The data analysis of the business sales has insightful information that can be used to improve the business. It can also be used to identify problem areas and increase the overall functioning of a company.

References

ASQ (n.d.). *What is Statistical Process Control? SPC Quality Tools | ASQ*. [online] Available at: https://asq.org/quality-resources/statistical-process-control#:~:text=Statistical%20process%20control%20(SPC)%20is.

BPI Consulting. (2004). *Process Capability Part 3*. [online] Available at: https://www.spcforexcel.com/knowledge/process-capability/process-capability-part-3.

Bright Data. (n.d.). *Guide to Data Wrangling*. [online] Available at: https://www.google.com/aclk?sa=l&ai=DChcSEwjjmayu8d36AhVY4O0KHZ3aA9UYABAAGgJkZw&sig=AOD64_1V2lRa1YeLvVsGuqTFFKCt1mDa5Q&q&adurl&ved=2ahUKEwj5uKau8d36AhXObMAKHcUgApAQ0Qx6BAgHEAE [Accessed 20 Oct. 2022].

Investopedia. (n.d.). *Logarithmic vs. Linear Price Scales: What's the Difference?* [online] Available at: https://www.investopedia.com/ask/answers/05/logvslinear.asp#:~:text=Logarithmic%20price%20scales%20are%20better [Accessed 20 Oct. 2022].

Manova (2021) *Statistics Solutions*. Available at: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/manova/ (Accessed: October 18, 2022).

Peacer, H. (2020) *Guide to data wrangling*, *Google*. Google. Available at: https://www.google.com/aclk?sa=l&ai=DChcSEwjjmayu8d36AhVY4O0KHZ3aA9UYABAAGgJkZw&sig=AOD64_1V2lRa1YeLvVsGuqTFFKCt1mDa5Q&q&adurl&ved=2ahUKEwj5uKau8d36AhXObMAKHcUgApAQ0Qx6BAgHEAE (Accessed: October 17, 2022).

Shen, S. (2021) *7 steps to ensure and sustain data quality*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/7-steps-to-ensure-and-sustain-data-quality-3c0040591366 (Accessed: October 17, 2022).

Trochim, P.W.M.K. (2022) *Descriptive statistics*, *Research Methods Knowledge. Base*. Conjointly. Available at: https://conjointly.com/kb/descriptive-statistics/ (Accessed: October 17, 2022).

What is Statistical Process Control? (2022) *ASQ*. Available at: https://asq.org/quality-resources/statistical-process-control#:~:text=Statistical%20process%20control%20(SPC)%20is,find%20solutions%20for%20production%20issues. (Accessed: October 17, 2022).