UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

ESCA Project

Quality Assurance 344

D. Jacobsz

23547006

17 October 2022

saam vorentoe · masiye phambili · forward together

Department of Mechanical and Mechatronic Engineering
Departement Meganiese en Megatroniese Ingenieurswese
Privaat Sak X1, Private Bag X1, Matieland, 7602
Tel: +27 21 808 4204 | www.eng.sun.ac.za

ENGINEERING
EZOBUNJINELI
INGENIEURSWESE

Plagiarism declaration

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.

I declare that, 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.

I declare that have not allowed, and will not allow, anyone to use my work (in paper, graphics, electronic, verbal or any other format) with the intention of passing it off as his/her own work.

I know that a mark of zero may be awarded to assignments with plagiarism and also that no opportunity be given to submit an improved assignment.

Signature:  .....

Name: Dewan Jacobsz Student no: 23547006

Date: 17/10/2022

Table of contents

List of figures

Page

List of tables

Page

Introduction

1.1 Background

An online business has client data that needs to be analysed in order to better understand trend within buying products and concepts like delivery times etc. All of these concepts have a huge impact in the performance of the business.

1.2 Objectives

The data consist of instances that are not all valid, meaning that these instances should be removed in order to clean the data. This clean data can then be used to make insightful observation that the business can use to make useful decisions.

The data analysis process consists of data wrangling, descriptive statistics, statistical process control, optimising the delivery process, MANOVA testing and relevant calculations like reliability and mean delivery time calculations.

1.3 Motivation

Data analysis is very important in any business. It always businesses to understand trends within their performance and insight on how to better this performance.

Data Analysis

Part 1: Data Wrangling

Data wrangling is also known as data cleaning. Data cleaning refers to a variety of processes performed on raw data to transform it into useful or "clean data". This "clean data" can be used to make insightful observations or decisions. By this definition it is clear that data wrangling should be the first step in the data analysis process.

The raw data set

The raw data set was received as an excel file called "salesTable2022". The data set consisted of 180 000 instances and 10 descriptive features namely: X, ID, Age, Class, Price, Year, Month, Day, Delivery time, why bought.

X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
179987	34352	60	Household	10919.36	2026	3	26	51.0	Random
179988	46435	63	Clothing	263.12	2027	6	14	8.0	Recommended
179989	97774	89	Technology	38123.42	2028	1	22	22.0	Recommended
179990	95681	74	Food	457.93	2021	1	13	3.0	Random
179991	95080	36	Food	554.41	2029	4	14	2.5	Browsing
179992	35316	76	Gifts	2722.28	2027	6	1	15.0	Website
179993	45954	48	Technology	27232.11	2026	8	27	20.5	Browsing
179994	51524	63	Luxury	76538.02	2027	1	23	3.0	Recommended
179995	49178	82	Food	505.88	2024	2	20	2.5	Website
179996	65414	31	Gifts	3147.66	2026	2	1	13.0	Recommended
179997	57864	34	Gifts	1111.36	2023	6	4	10.0	Recommended
179998	48301	77	Gifts	3943.92	2028	4	29	17.0	Website
179999	96502	56	Sweets	243.00	2023	5	26	2.0	Website
180000	71587	53	Household	15362.39	2021	8	22	43.5	Website

Figure 1: Extract of the last 13 instances of the raw data

One big problem with this data set is that it has missing values. As part of the data wrangling process these missing values had to be removed from the set. This led to the data set being separated into valid data, a data set without instances containing missing values and incomplete or invalid data, a data set consisting only of instances containing missing values.

Invalid data

This is a data set created from the raw data that consists only of instances containing missing values, this means that this data set is incomplete and rendered invalid. This data set contains instances with NA and negative values.

The code to create this data set runs through each row in order to find missing values. When it finds a missing value, it moves that entire row to the invalid data set.

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

Figure 2: Invalid data set

From figure 2 it is clear that the Invalid data set consists of 17 instances all of which contains missing values.

Valid data

This is a data set created from the raw data set that does not consist of any missing values, this means that the data set can be consider as clean data and used to make insightful observations and decisions. This data set has no NA or negative values.

The code to create this data set runs through each line looking for missing values. When it finds a missing values, it removes it from the data set.

n	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	1	19966	54	Sweets	246.21	2021	7	3	1.5	Recommended
2	2	34006	36	Household	1708.21	2026	4	1	58.5	Website
3	3	62566	41	Gifts	4050.53	2027	8	10	15.5	Recommended
4	4	70731	48	Technology	41843.21	2029	10	22	27.0	Recommended
5	5	92178	76	Household	19215.01	2027	11	26	61.5	Recommended
6	6	50586	78	Gifts	4929.82	2027	4	24	14.5	Random
7	7	73419	35	Luxury	108953.53	2029	11	13	4.0	Recommended
8	8	32624	58	Sweets	389.62	2025	7	2	2.0	Recommended
9	9	51401	82	Gifts	3312.11	2025	12	18	12.0	Recommended
10	10	96430	24	Sweets	176.52	2027	11	4	3.0	Recommended
11	11	87530	33	Technology	8515.63	2026	7	15	21.0	Browsing
12	12	14607	64	Gifts	3538.66	2026	5	13	13.5	Recommended
13	13	24299	52	Technology	27641.97	2024	5	29	17.0	Browsing
14	14	77795	92	Food	556.83	2025	6	3	3.0	Random
15	15	62567	73	Clothing	347.99	2024	3	29	8.5	Website
16	16	14839	47	Technology	54650.41	2027	12	30	18.5	Recommended
17	17	96208	44	Technology	14739.09	2028	3	17	13.0	Recommended
18	18	39674	69	Technology	22315.17	2026	8	20	20.5	Recommended
19	19	98694	74	Sweets	546.48	2025	5	9	2.0	Recommended
20	20	99187	54	Luxury	81620.21	2027	9	14	3.0	Recommended
21	21	59365	72	Gifts	3314.76	2028	4	30	13.0	Recommended

Figure 3: Valid data set

Figure 3 only shows the first 21 instances of the valid data set. This data set consists of 179983 instances.

The Index problem

The process of created two different data sets indirectly led to another problem. The number of instances no longer equal the Index of the instances. When removing the invalid data from the data the entire row gets removed therefor the instances are no longer consecutive. The solution to this problem was to add another feature namely PrimaryKey to fix this problem.

PrimaryKey	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bou
12344	12344	90260	34	Luxury	42891.66	2025	8	4	4.0	Recor
12345	12346	92286	32	Technology	38167.24	2028	7	6	19.5	Webs
12346	12347	89263	44	Clothing	891.71	2021	7	2	8.5	Recor
12347	12348	71191	49	Household	14936.31	2025	10	11	43.5	Recor

Figure 4: Index problem

From figure 4 it is clear the instances with index number 12345 consisted a missing value and is thus removed from the valid data set. This problem was fixed by adding the PrimaryKey feature.

Part 2: Descriptive Statistics

One of the most important things in the data analysis process is to understand the data set. To fully understand the data descriptive statistics is used. Descriptive statistics provides useful information in terms of the data's corelation. It also provides the minimum, maximum and mean values. Descriptive statistics can be used to compare different features to each other and make informative observations based on this comparison.

The data set consists of 2 Categorical features and 8 Continuous features. The following tables provide information on these two different types of features. The information includes the length of these features, the mode, mean, quartiles, maximum, minimum etc.

Categorical Features:

Feature	Length	#Missing values	Mode	Mode Frequency	Mode %	Cardinality
Class	179978	0	Gifts	39149	25.45	7
Why. Bought	179978	0	Recommended	106985	59.44	6

Table 1: Categorical features of the Data set Continuous Features:

Feature	Length	# Missing values	Minimum	1st Qrt.	Mean	Median	3rd Qrt.	Maximum	Cardinality
PrimKey	179978	0	1	44995	89990	89990	134984	179978	179978
ID	179978	0	11126	32700	55235	55081	77637	99992	15000
Age	179978	0	18	38	54.57	53	70	108	91
Price	179978	17	31	419.4	10690.5	1964.9	13279.1	101407.8	78832
Year	179978	0	2021	2022	2025	2025	2027	2029	9
Month	179978	0	1	4	6.521	7	10	12	12
Day	179978	0	1	8	15.54	16	23	30	30
Delivery time	179978	0	1	6	29	20	37	150	148

Table 2: Continuous features of the Data set

Visual representation and comparison of different features

As previously mentioned, different features can be compared to each other to make insightful observations. The following graphs are visual representations of different features and their statistical descriptions.

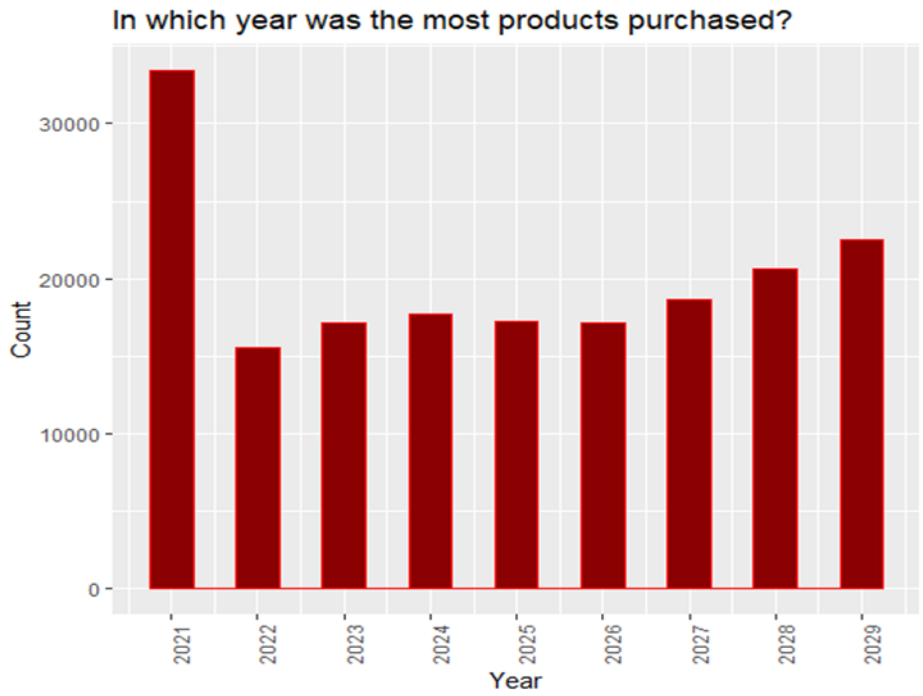


Figure 5: Sales for each year

Comment: From figure 5 it is clear that the most sales were made in 2021. From 2022 an increase in trend can be seen up until 2029. This can be because of the rising inflation rate that discourages people to spend money.

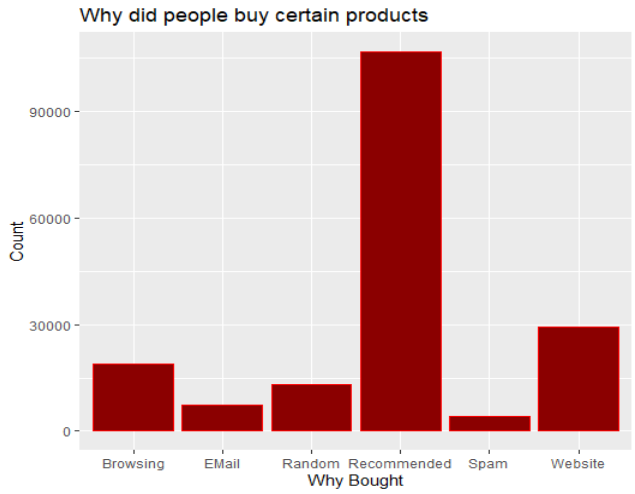


Figure 6: Why bought graph

Comment: From figure 6 it is clear that the main reason why people bought a product is because someone recommended that product. On the other hand, email was the least frequent reason for buying products. This can mean that most people see email from business as “spam”. People still value other peoples opinion which mean that if someone else recommends a product people are more inclined to buy that product.

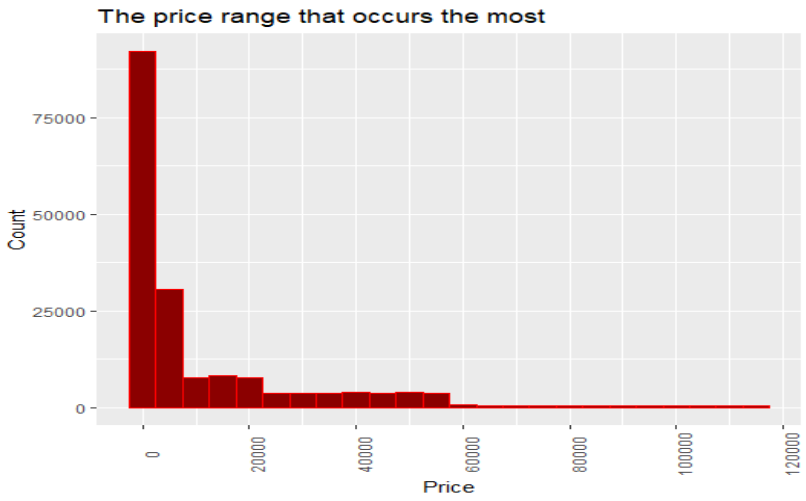


Figure 7: Price graph

Comment: Figure 7 clearly indicates that products within a lower price range was bought more frequently than products within a higher price range. This can be because of the fact that customers are more likely to pay for cheaper products than more expensive ones.

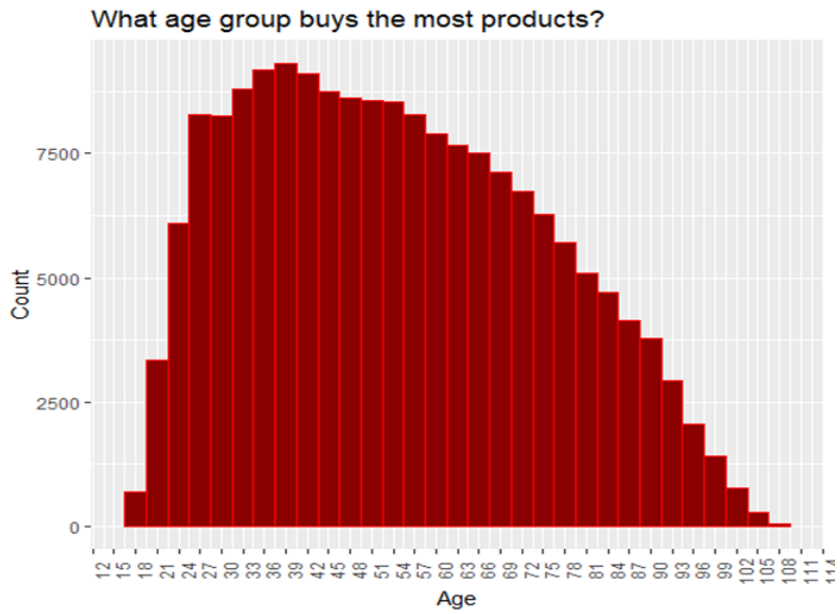


Figure 8: Age graph

Comment: This distribution is slightly skewed to the right. Meaning that it tends to lower values, with the most frequent values, or in this cause age group being between 33 and 40. This makes sense because most people between 33 and 40 would have young kids that often need a lot of products for school, sport etc.

Figure 9: Delivery time graph

Comment: Figure 9 clearly shows a normal distribution between 35 and 65 days, this distribution indicates the most frequent delivery time to be between 2 and 5 days. The reason for this delivery time being the most frequent might because of the fact that people value good service and low delivery times.

Calculation of Process capability indices

The measure of "how much" variation a process experience is know as the process capability index.

To interpret these indices the focus will be on the delivery times of the class "Technology". The following figure represents the relative calculations relating to this specific process capability index.

Assuming LSL=0 and USL is 24. The reason for USL being 0 is logical because of the fact that the delivery time can not be less than 0, no product can be delivery in a time less than zero.

```
Cp= (USL-LSL)/(6*sd_tech)          #Cp = 1.142207
Cpu= (USL- mean_tech)/(3*sd_tech)  #Cpu = 0.3796933
Cpl= (mean_tech - LSL)/(3*sd_tech) #Cpl = 1.90472
Cpk= min(Cpl,Cpu)                  #Cpk = 0.3796933
```

Figure 8: Process Capability Indices

This process is not very capable. The reason for this is because the CP is slightly larger than 1. This means that the process is not fully capable of delivering products within the specified time limits without being perfectly centred. This problem can be improved by making the standard deviation of delivery times smaller.
This process is not centred. The reason for this is because CPK of 0.3797 is far from the CP value of 1.142. This can be fixed by shifting the process to fit the target better.

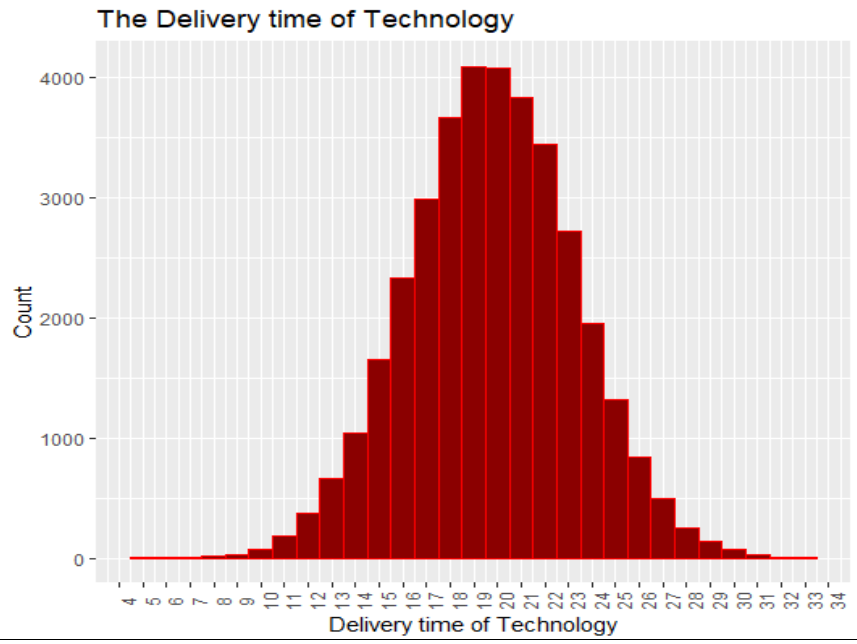


Figure 9: Delivery time of technology graph

Visual representations of the correlation between different features

In order to further understand the data, correlations between the different features need to be analysed and interpreted. For example, how would we know what class of product usually takes the longest to get delivered. To answer this, we would have to look at the correlation between delivery times and class. The following figure is a visual representation of this correlation.

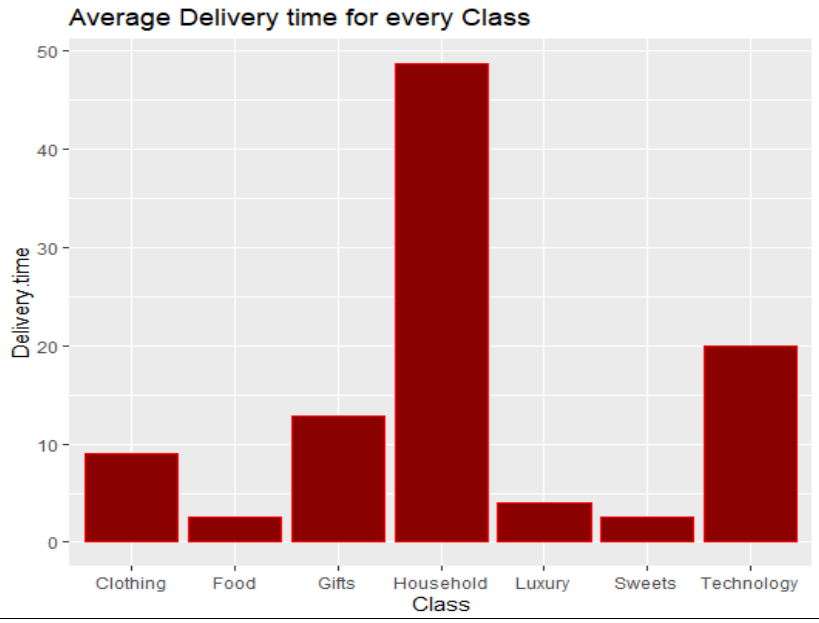


Figure 12: The average delivery time for every class

Comment: From figure 12 it is clear that the delivery time for household item is the longest, this can be because of the fact that household item tends to be larger items which take more time to deliver to customers. On the other hand, items like sweets or food have much less delivery time because these items tend to be much smaller in size and therefor quicker to deliver.

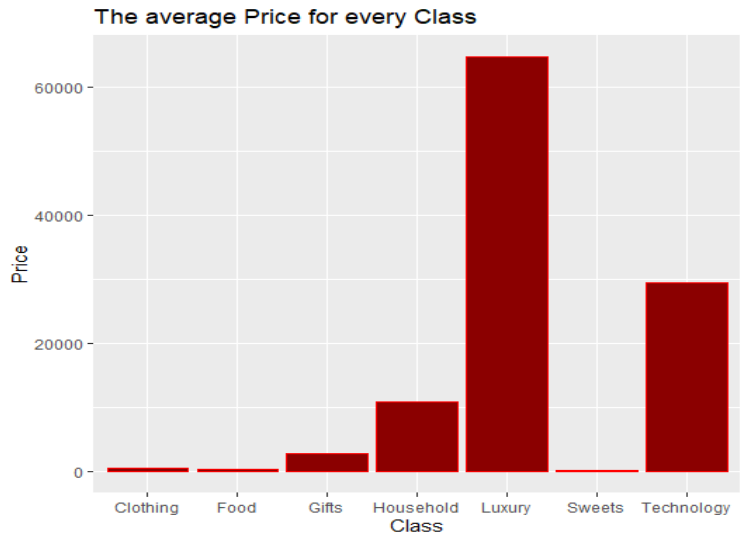


Figure 10: The average price for class

Comment: From figure 13 it can easily be noted that luxury items have the highest price, technology is the second most expensive product. Items like clothing, food and sweets are relatively cheap in comparison to luxury and technology items.

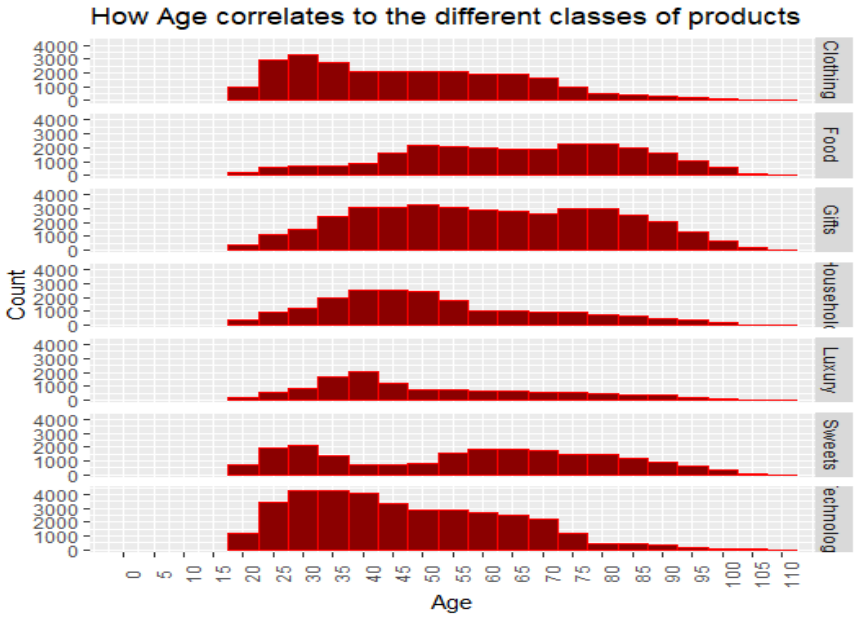


Figure 11: Age vs class

Comment: All of the classes are spread over a wide range of ages, from the age of 15 up to the age of 110. Its clear that clothing are purchased more frequently between the ages of 20 and 35. Food has a more consistent trend ranging from the ages 40 to 80. Gifts are widely distributed over all ages. Household items are more frequent at the age of 30 to 50, and then a strong decline in frequency after the age of 50. Luxury is exponentially distributed with most frequent value at the age 37. Sweets are distributed in such away that it peaks at two different age groups, first peak is at the age of 27 and the second peak is in the rage of ages 50 to 70. Technology is most frequent at the ages 25 to 35 with a slight decline as the value of age becomes larger.

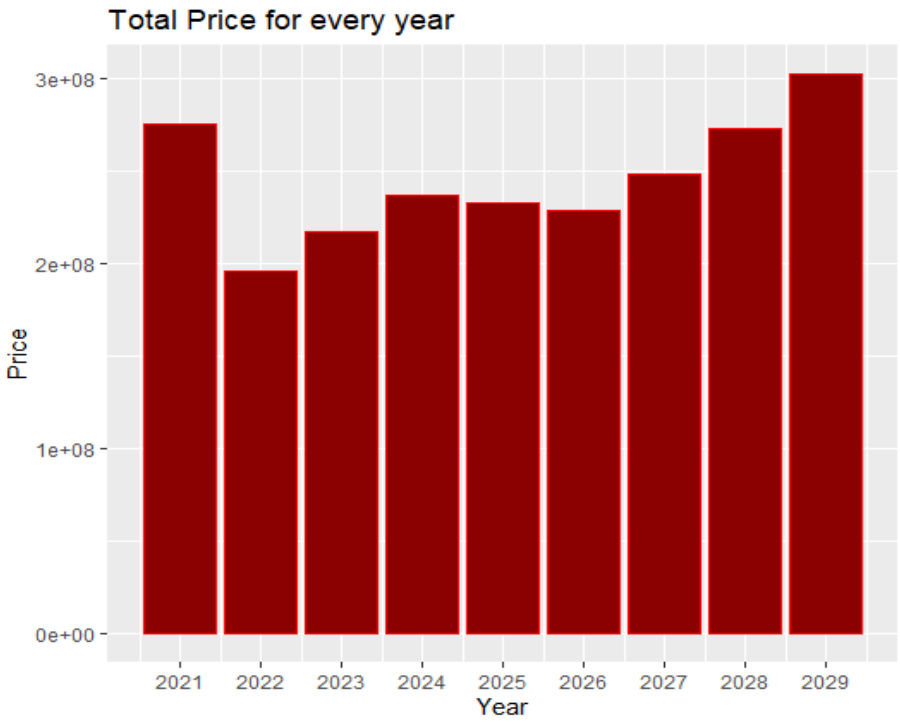


Figure 12: Total price for every year

Comment: Figure 15 shows that prices in 2022 are considerably less than in 2021. After 2022 the prices gradually increase year after year until it surpasses 2021's prices in the year 2029.

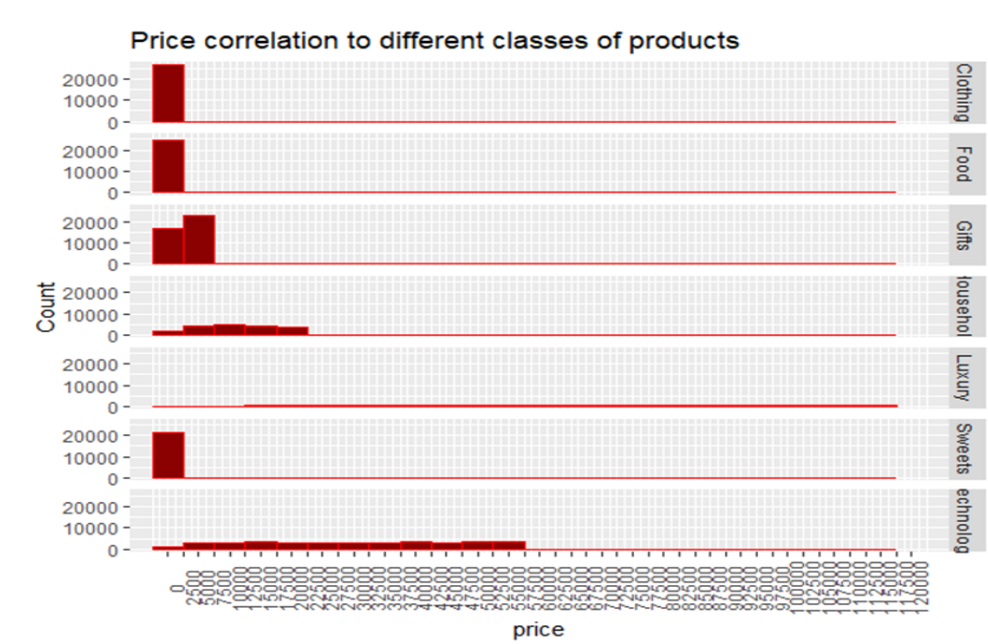


Figure 16: Price vs class

Comment: How price correlates to the different classes is showed in figured 16. Clothing, food, gifts and sweets are more frequently low priced. Where as technology-and-luxury's prices range from low to medium-high.

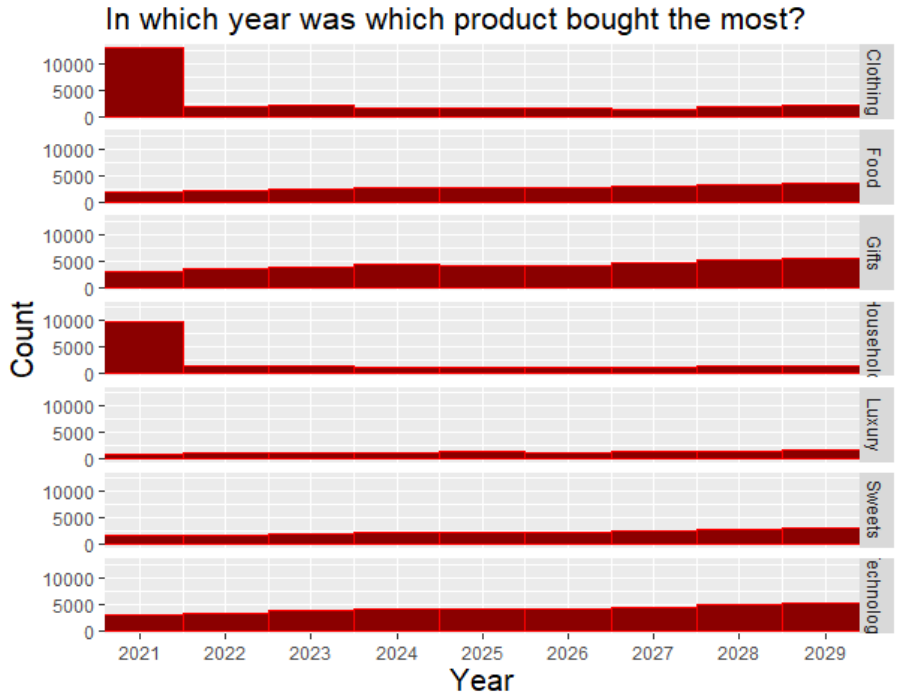


Figure 17: year vs class

Comment: Figure 17 visually indicates the year in which specific classes was bought the most. Clothing and household were bought most frequently in 2021 whereafter it took a steep decline and stayed mostly consistent until 2029. Food gifts, luxury, sweets and technology is distributed with a gradual increase as the years increase.

Part 3: Statistical Process control (SPC)

The X-and-S chart are constructed for delivery times using a sample size of 15 for 30 initial samples. In order to use these sample, the valid data had to be ordered from oldest to latest by year, month and then day.

X-Chart table:

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Sweets	2.896798	2.757125	2.617451	2.477778	2.338105	2.198431	2.058758
Household	50.246180	49.018193	47.790207	46.562220	45.334233	44.106247	42.878260
Gifts	9.487909	9.112310	8.736710	8.361111	7.895512	7.609912	7.234313
Technology	22.973100	22.106880	21.240660	20.374440	19.508220	18.642000	17.775790
Luxury	5.493524	5.240868	4.988212	4.735556	4.482900	4.230244	3.977587
Food	2.709330	2.636220	2.563110	2.490000	2.416890	2.343780	2.270670
Clothes	9.404681	9.259787	9.114894	8.970000	8.825106	2.270670	8.535319

Table 3: X-chart table

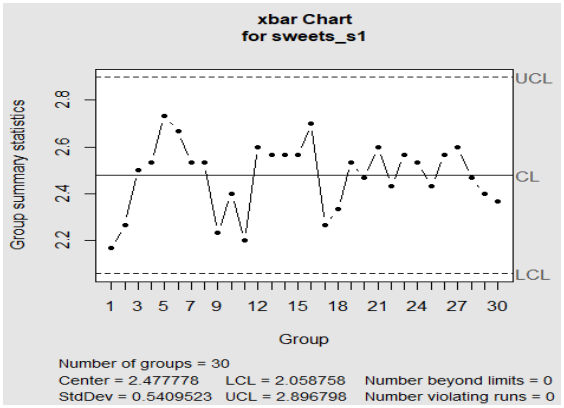
S-Chart table:

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Sweets	0.8352331	0.7339508	0.6326685	0.5313862	0.4301039	0.3288216	0.2275393
Household	7.3432480	6.4527887	5.5623293	4.6718700	3.7814107	2.8909513	2.0004930
Gifts	2.2460480	1.9736870	1.7013260	1.4289650	1.1566040	0.8842430	0.6118823
Technology	5.1799120	4.5517840	3.9236560	3.2955280	2.6674000	2.0392720	1.4111430
Luxury	1.5108600	1.3276496	1.1444393	0.9612289	0.7780185	0.5948082	0.4115978
Food	0.4371911	0.3841763	0.3311615	0.2781467	0.2251319	0.1721171	0.1191023
Clothes	0.8664496	0.7613819	0.6563142	0.5512465	0.4461788	0..3411111	0.2360435

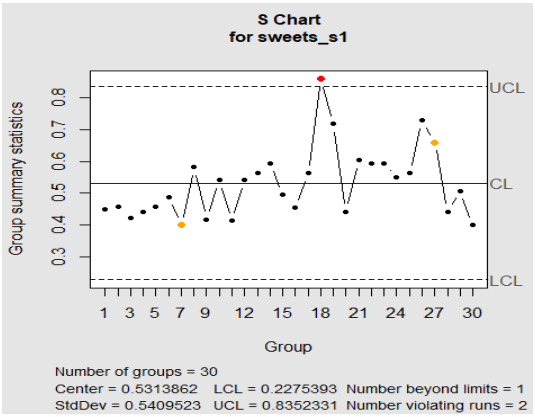
Table 4: S-chart table

Graphs for the first 30 samples for each class

Sweets:

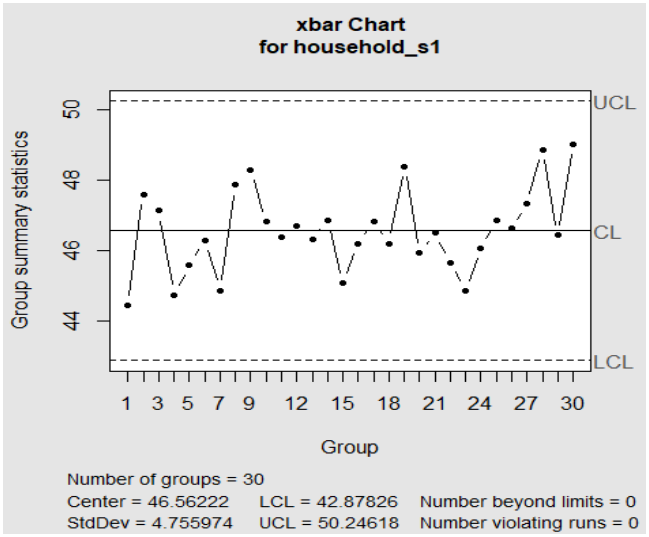


This chart shows that all of the sample means are in control.

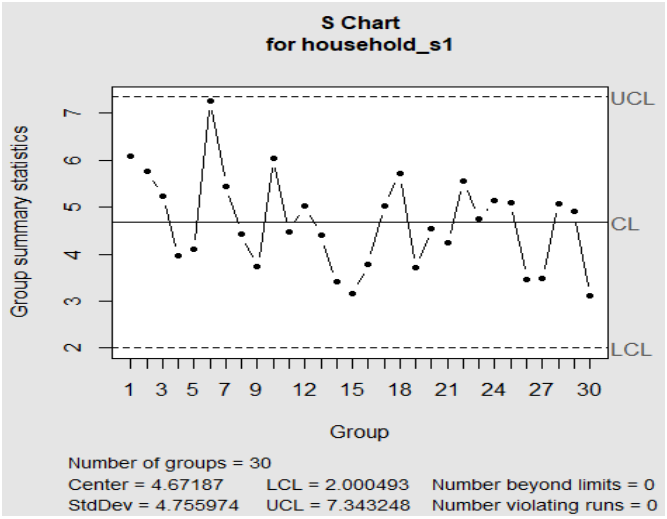


This chart shows that all of the sample standard deviations are in control except for sample 16, therefor sample 16 should be removed.

Household:

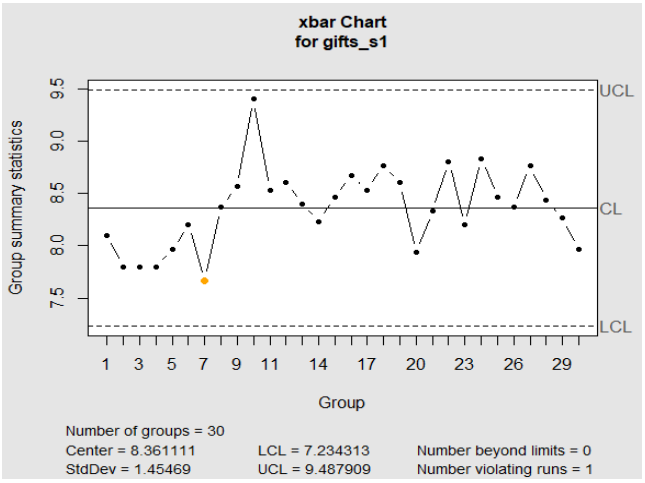


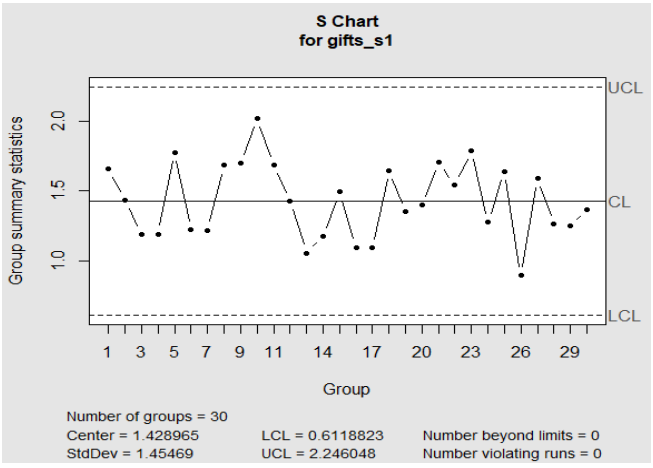
This chart shows that all of the sample means are in control.



This chart shows that all of the sample standard deviations are in control. There is no clear variation in the sales of household items.

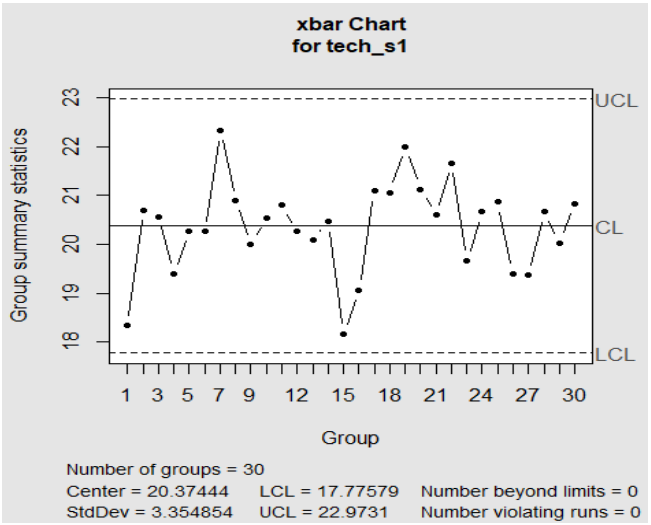
Gifts:

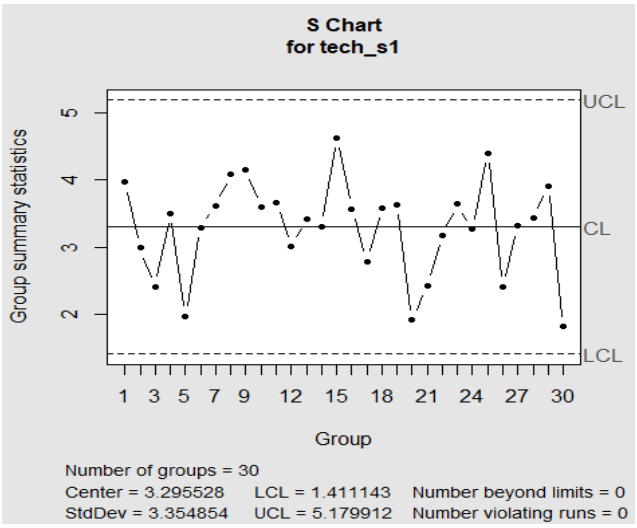




This chart shows that all of the sample means are in control.	This chart shows that all of the sample standard deviations are in control. There is no clear variation in the sales of gifts.

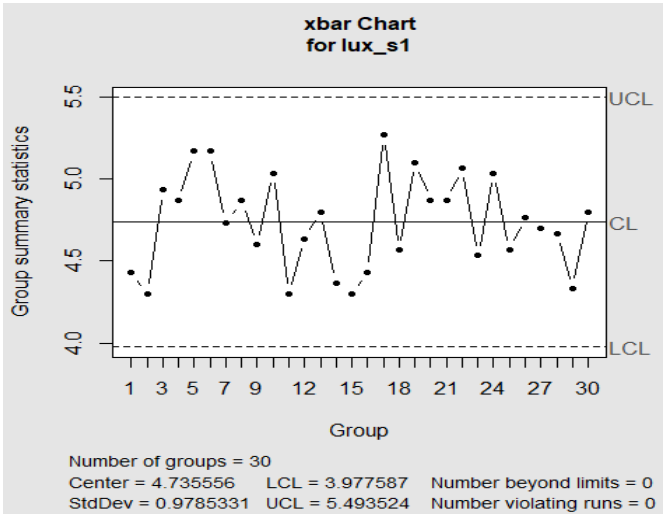
Technology:

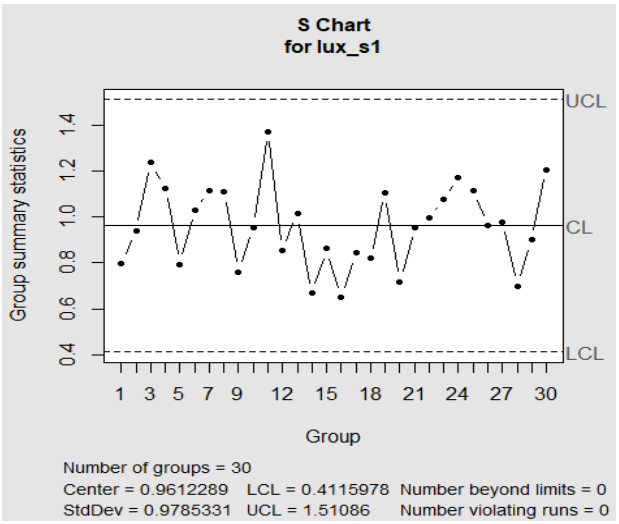




This chart shows that all of the sample means are in control.	This chart shows that all of the sample standard deviations are in control. There is no clear variation in the sales of technology.

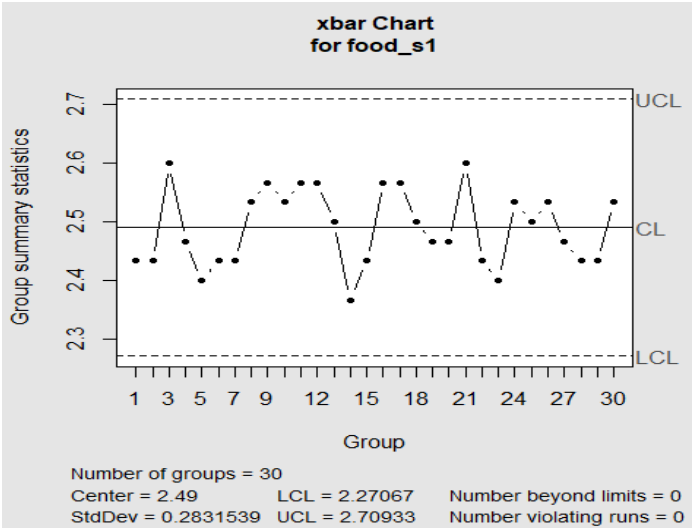
Luxury:

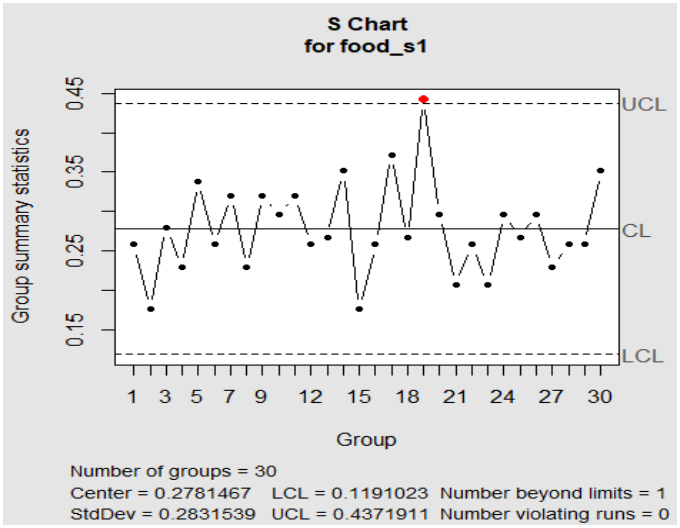




This chart shows that all of the sample means are in control.	This chart shows that all of the sample standard deviations are in control. There is no clear variation in the sales of luxury items.

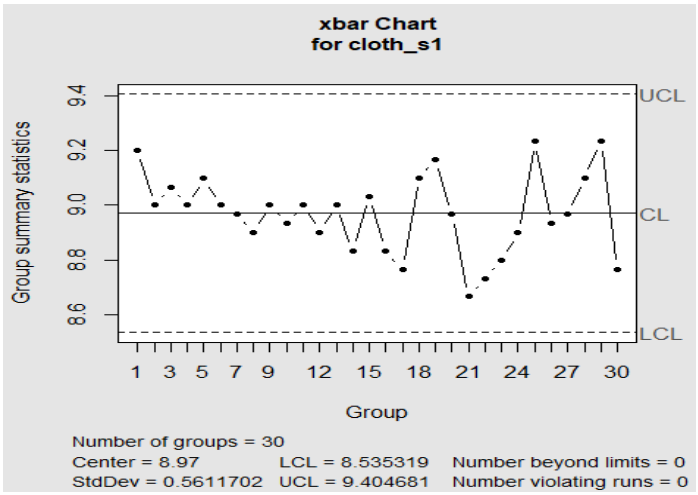
Food:

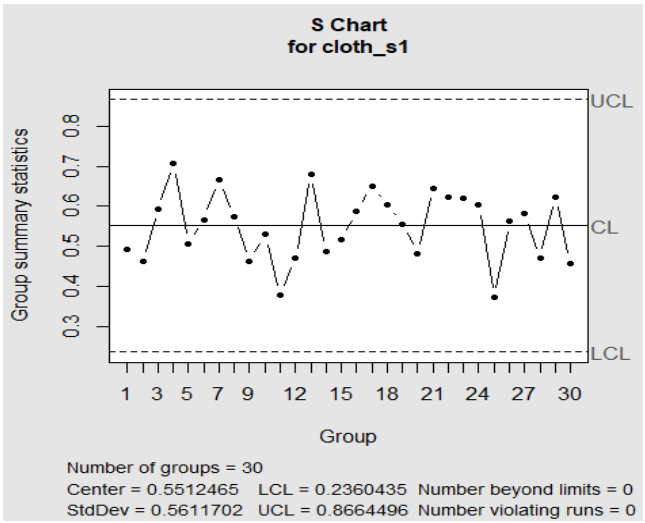




This chart shows that all of the sample means are in control.	This chart shows that all of the sample standard deviations are in control except for sample 19, therefor sample 19 should be removed.

Clothing:

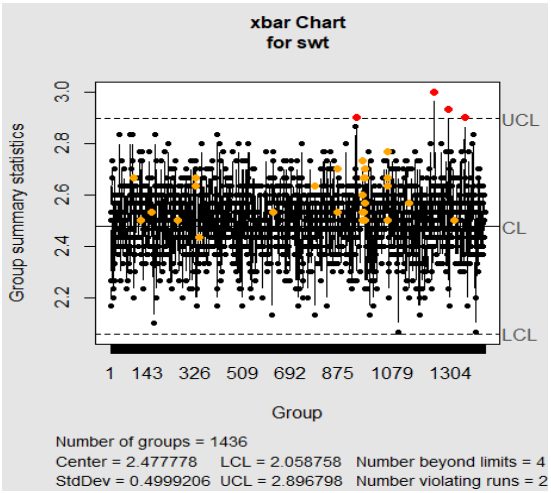
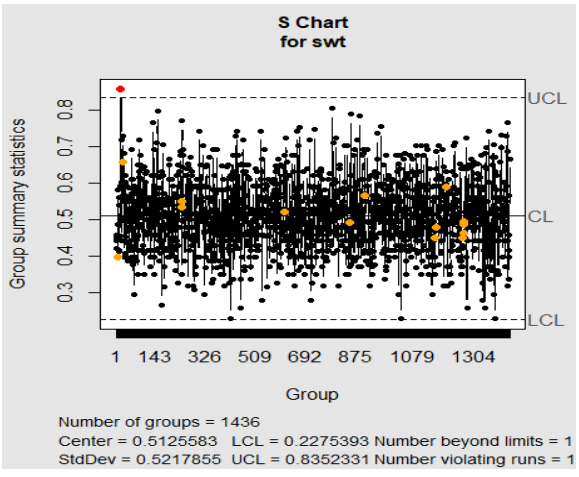




This chart shows that all of the sample means are in control.	This chart shows that all of the sample standard deviations are in control. There is no clear variation in the sales of clothing items.

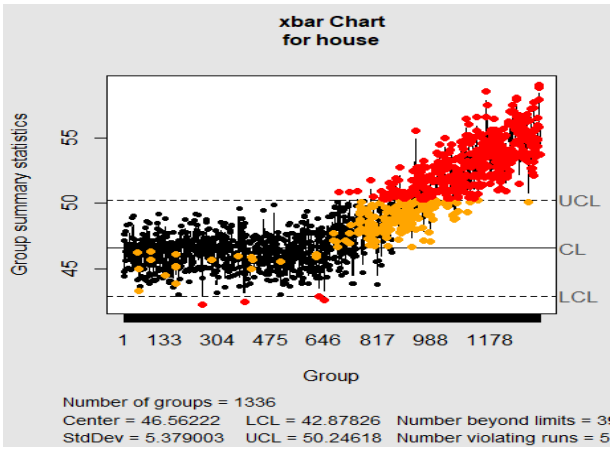
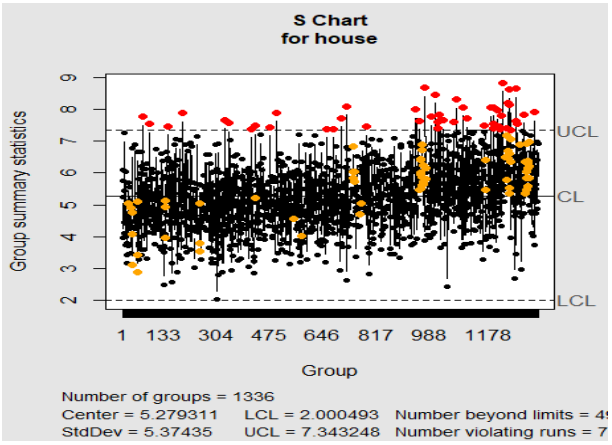
Part 3.2: X-and-S bar charts for all samples.

Sweets:

X-chart	S-chart
	

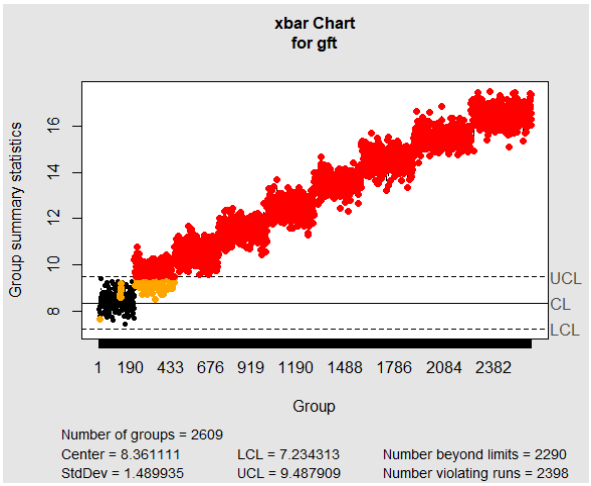
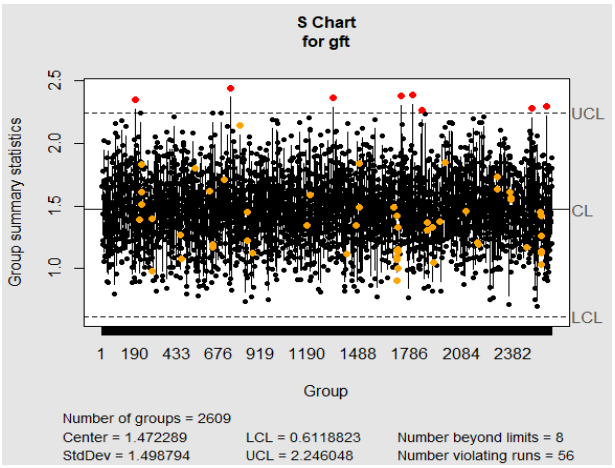
The mean and variance of Sweets is not a function of time and it is clear that Sweets is under control with very few instances that fall outside of the outer control limits.

Household

X-chart	S-chart
	

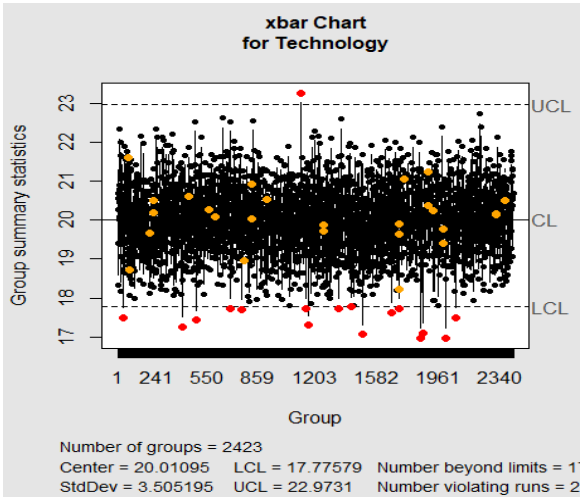
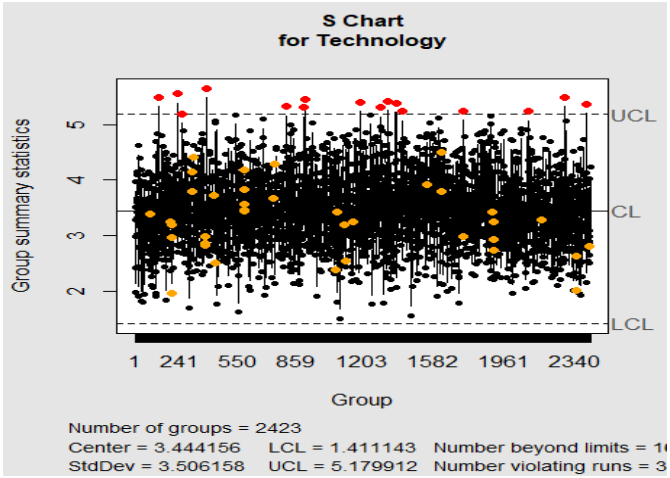
The delivery time for household items seem to have an increasing trend that starts more or less in the middle of the number of instances. This means that Household is out of control and the reason for this increase in delivery time should be investigated further.

Gifts

X-chart	S-chart
	

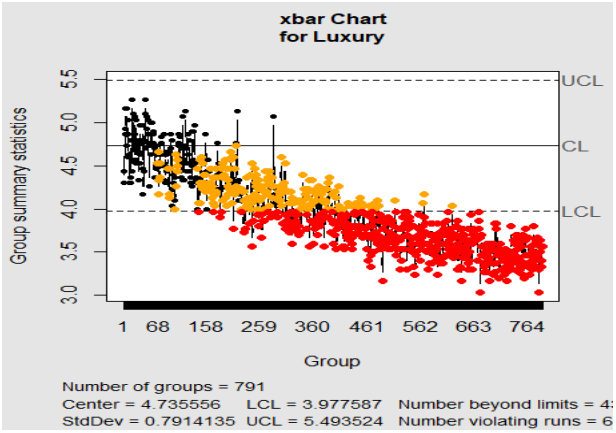
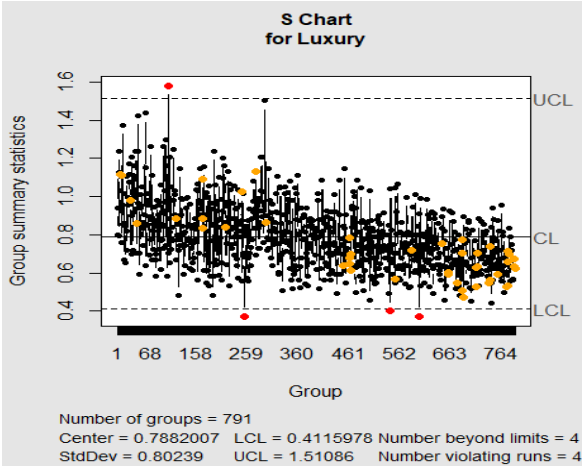
The delivery time of Gifts indicates an increasing linear trend, which means that Gifts are not under control and should therefor be further investigated in order to understand this increase.

Technology

X-chart	S-chart
	

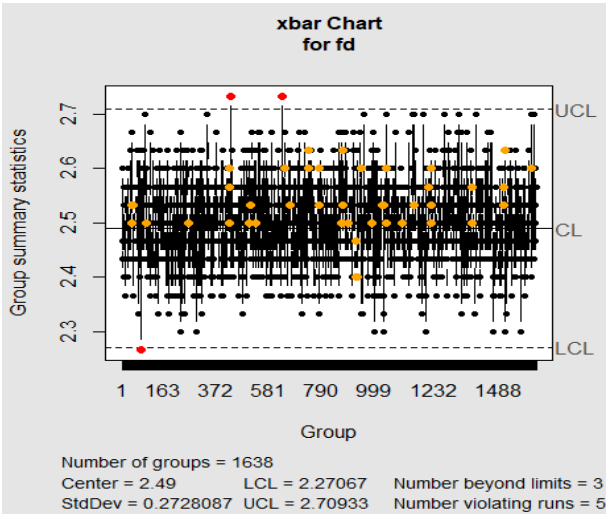
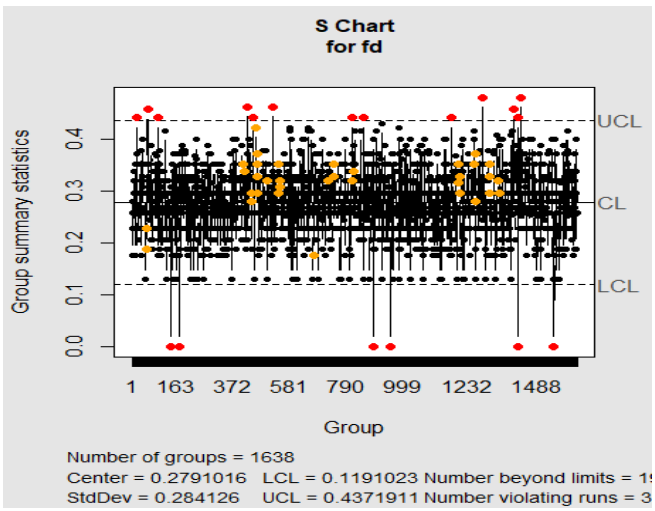
The delivery time of Technology is mostly under control except for few instances that fall outside of the outer control limits.

Luxury

X-chart	S-chart
	

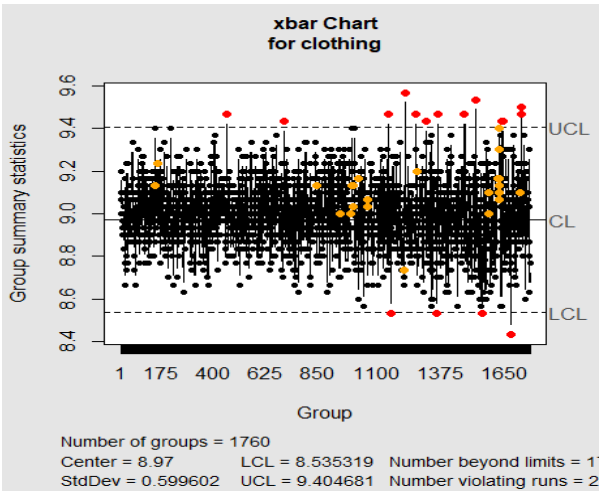
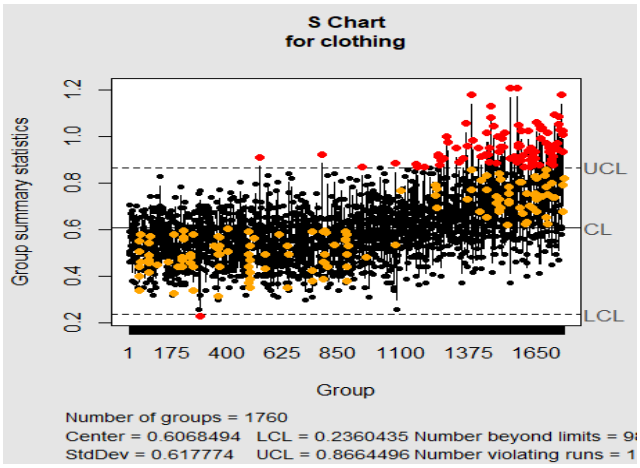
The delivery time for Luxury items indicates a decreasing linear trend, which means it is considered to be out of control. One reason for this may be because of the fact that luxury items are expected to have fast delivery times, lower delivery times are usually considered to be a good thing therefor the reason for this decrease should be investigated further. On the other hand, the variation of luxury is considered to be under control.

Food

X-chart	S-chart
	

The delivery times of Food seem to be under control with very few instances that fall outside of the outer control limits.

Clothing

X-chart	S-chart
	

For clothing is it clear that most of the instances are within in the control limits when looking at the X-chart, thus the conclusion can be made that clothing is under control, but when looking at the S-chart it is clear that the variance of clothing has a large increase at later instances meaning that is might not be completely under control. Therefor further investigation is needed to under this increase in variation. One of the reasons might be seasonal changes.

Part 4: Optimising the delivery processes

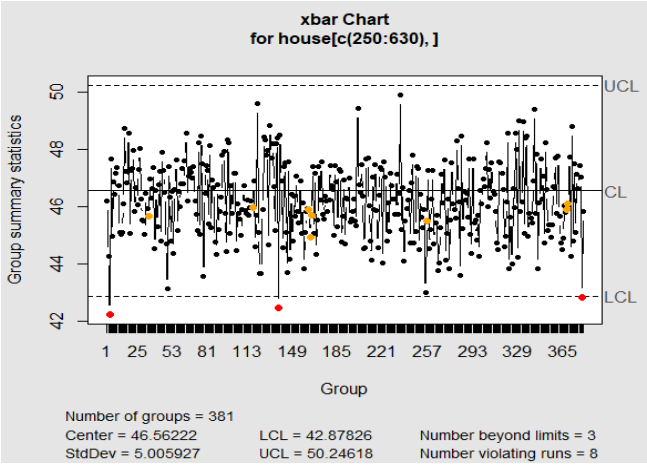
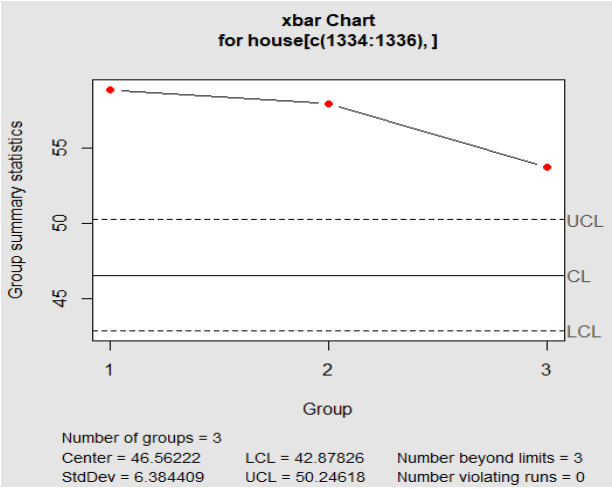
4.1 A) Samples that are outside of outer control limits

Class	Total	1st sample	2nd sample	3rd sample	3rd last sample	2nd last sample	Last sample
Sweets	4	942	1243	1294	NA	NA	1358
Household	396	252	387	629	1335	1336	1337
Gifts	2290	213	216	218	2607	2608	2609
Technology	17	37	398	483	1872	2009	2071
Luxury	433	142	171	184	788	789	790
Food	3	75	432	633	NA	NA	NA
Clothes	17	455	702	1152	1677	1123	1724

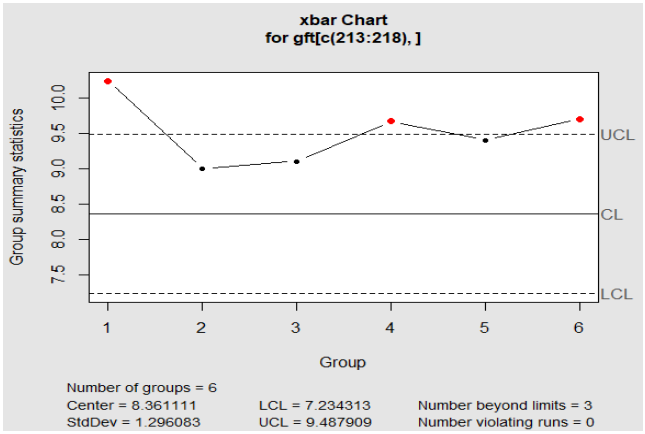
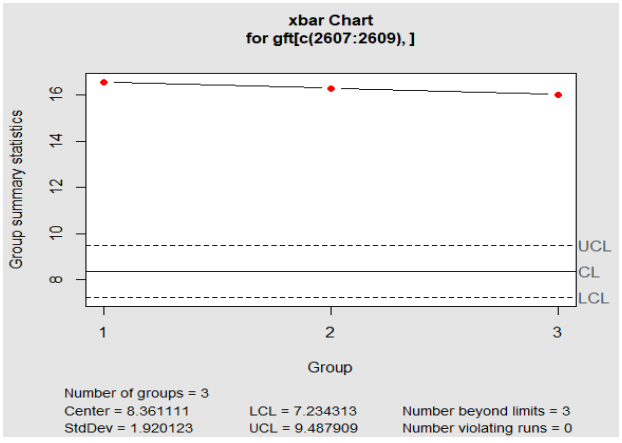
Table 5: Samples that are outside of outer control limits

From table 5 it is clear that sweets, technology, food and clothes are in control because of the fact that these classes have very few samples that are outside of the outer control limits. On the other hand, the classes such as household, gifts and luxury have a lot of samples that fall outside of the outer control limits, which means that these classes are considered to be out of control. This means that these classes should be further investigated to determine what causes the delivery time of these classes to variate outside of the outer control limits.

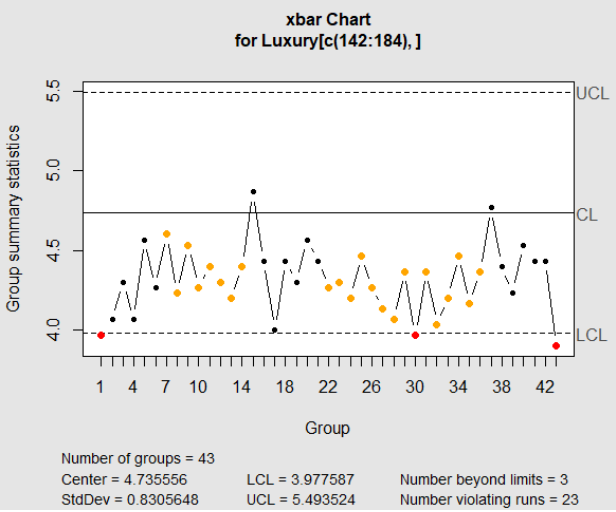
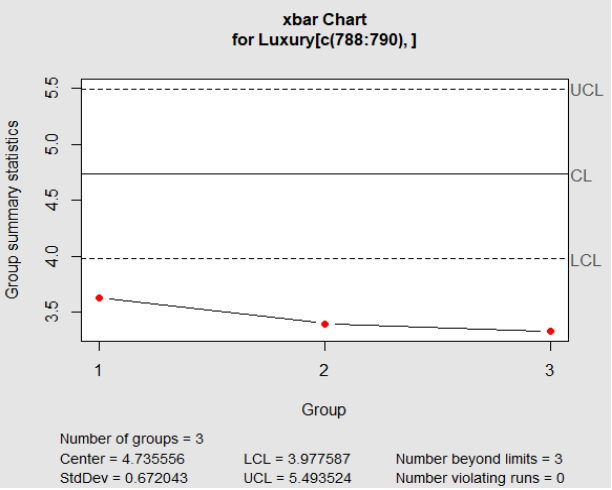
Household

First 3	Last 3
	
Displaying samples from 250 to 630	Displaying samples from 1335 to 1337

Gifts

First 3	Last 3
	
Displaying samples from 213 to 218	Displaying samples from 2607 to 2609

Luxury

First 3	Last 3
	
Displaying samples from 142 to 148	Displaying samples from 788 to 790

4.1 B) Most consecutive samples

Class	Max # consecutives	Ending sample index number
Sweets	4	971
Household	3	843
Gifts	5	1651
Technology	6	372
Luxury	4	63
Food	6	441
Clothes	4	1031

4.2) Probability of making a Type I error for A and B

H0: The process mean is within the control limits and centred on the centreline.
H1: The process mean is not within the control limits and has moved from the centreline or has increased/decreased in variation.

P(Type I error for A) = (1-pnorm(3)) + (pnorm(-3))

= 0.002699796

= 0.27%

P(Type I error for B) = (1-pnorm(0.4)) + pnorm(-0.3)

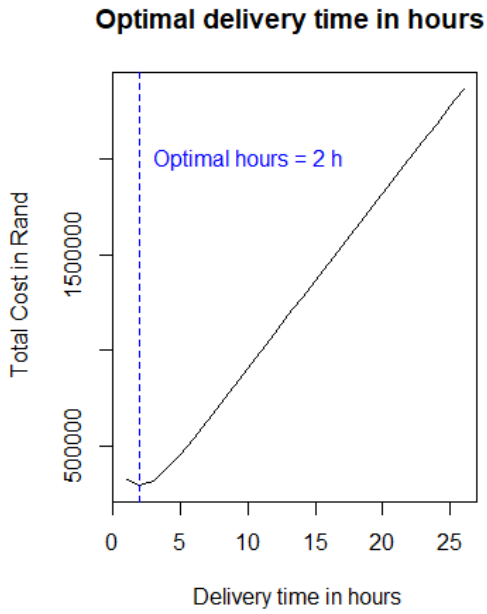
= 0.7266668

= 72.67%

4.3) Calculating a new centre for the delivery process of technology items.

The current mean in hours for delivering technology products are 20. 01095 hours. The total amount of delivery hours over the 26-hour limit are 1356 hours. At a lost sale cost of R329, this cost will lead to total loss of R446 124. It costs R2.5 per hour to move the mean to the left. This means that when the entire distribution is taking into account it will cost R636072.5 to move the mean.

The cost for different hours is compared by looping through every hour. By doing this is was found that the mean should move 2 hours to the left, in other words the optimal delivery time centre is 18. 01095 hours.

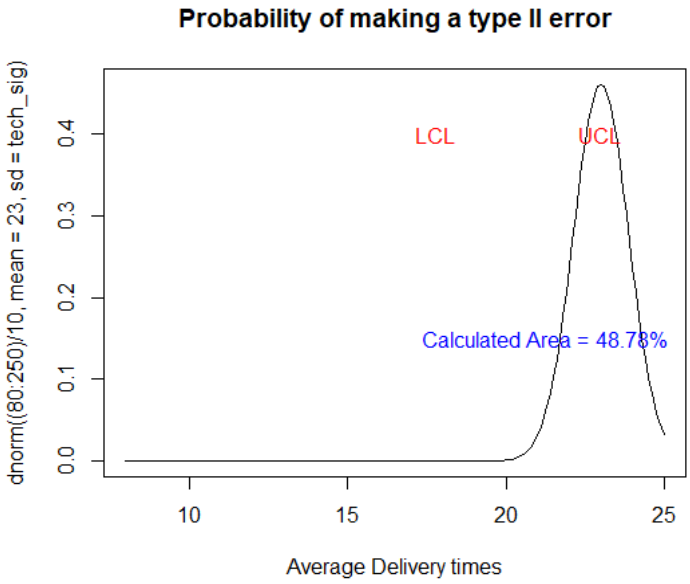


4.4) Probability of making a Type II Error for A in Class=Technology

```
#### Part 4.4 #####
tech_sig <- (UCL_mean[4]-LCL_mean[4])/6

#probability of sample being inside the limits
pnorm(UCL_mean[4],mean=23,sd=tech_sig)-
  pnorm(LCL_mean[4],mean=23,sd=tech_sig)
#0.487613
```

The figure above was used to calculated the probability for making a type two error for A when Claas+ technology. A type two error occurs when the product is delivered late but according to the company the product was delivered on time. The probability of this happening is 48.76%, which is quite high and the company might have to look at methods to prevent a type two error.



Part 5: MANOVA testing

For the MANOVA test, a p value of 0.05 is chosen.

5.1 Hypothesis test 1

Dependent variables	Age, Price and Delivery times
Independent variable	Why Bought
H0	Age, Price and Delivery times made no remarkable change to the reason why the product was bought.
H1	At least one of these features had an influence on the reason why the product was bought.

Manova test: test whether there is a feature that has an influence

P value	<2.2e-16 Reject Null Hypotheses. At least one of these features had an influence on the reason why the product was bought.

Each dependent variable and class:

Dependent variable	P value	Analyses
Age	2.2e-16	Age's P value is smaller than 0.05. This means that Age differs depending on the reason why the product was bought.
Price	2.2e-16	Price's P value is smaller than 0.05. This means that Price differs depending on the reason why the product was bought.
Delivery times	2.2e-16	Delivery time's P value is smaller than 0.05. This means that Delivery times differs depending on the reason why the product was bought.

5.2 Hypothesis test 2

Dependent variables	Day, Month and Year
Independent variable	Why Bought
H0	Day, Month and Year made no remarkable change to the reason why the product was bought.
H1	At least one of these features had an influence on the reason why the product was bought.

Manova test: test whether there is a feature that has an influence

P value	<2.2e-16 Reject Null Hypotheses. At least one of these features had an influence on the reason why the product was bought.

Each dependent variable and class:

Dependent variable	P value	Analyses
Day	0.5585	Day's P value is bigger than 0.05. This means that Day has no influence on why the product was bought.
Month	0.7902	Month's P value is bigger than 0.05. This means that Month has no influence on why the product was bought.
Year	2.2e-16	Year's P value is smaller than 0.05. This means that Year differs depending on the reason why the product was bought.

5.3 Hypothesis test 3

Dependent variables	Day, Month and Year
Independent variable	Class
H0	Day, Month and Year made no remarkable change to the specific class of product that was bought.
H1	At least one of these features had an influence on the specific class of product that was bought.

Manova test: test whether there is a feature that has an influence

P value	<2.2e-16 Reject Null Hypotheses At least one of these features had an influence on the specific class of product that was bought.

Each dependent variable and class:

Dependent variable	P value	Analyses
Day	0.1766	Day's P value is bigger than 0.05. This means that Day has no influence on specific class of product that was bought.
Month	0.2859	Month's P value is bigger than 0.05. This means that Month has no influence on the specific class of product that was bought.
Year	2.2e-16	Year's P value is smaller than 0.05. This means that Year differs depending on the specific class of product that was bought.

Part 6:

6.1) Problem 6

Thickness of a refrigerator part is 0.06+-0.04cm. Costs $30 to scrap a part. Taguchi loss function:

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

$45 = k\ (0.06)^2$

k = 45/ $(0.06)^2$ = 28125

k = 12500

Loss function: $L(x) = k(x - T)^2$

$L(x)$ = 12500 $(x - 0.06)^2$

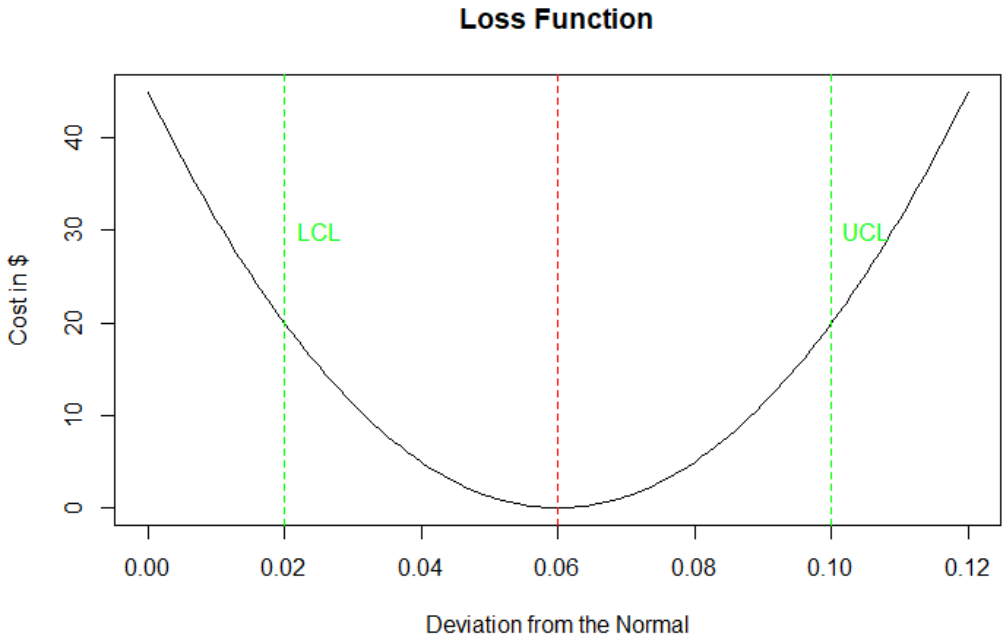


Figure 13: Loss function graph

As seen in figure 13, the more the thickness of a refrigerator deviate from a value of 0.06 the more it will cost Cool Food Inc, which means less profit for the company.

Problem 7

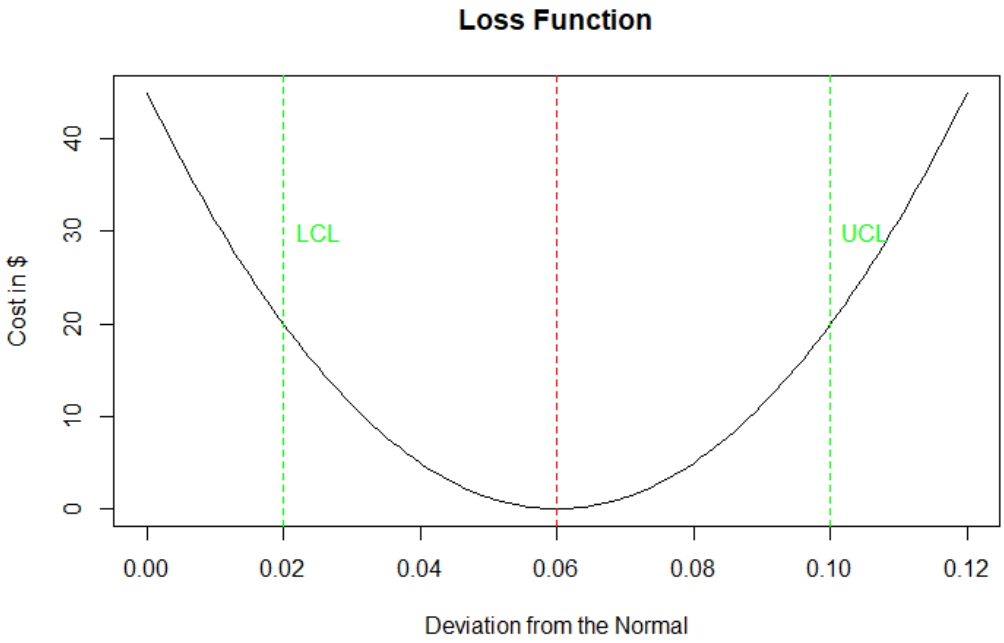
a) $L(x) = k(x - T)^2$

35 = k $(0.06)^2$

k = 35/ $(0.06)^2$ = 9722.22

Loss function: $L(x) = k(x - T)^2$

: $L(x) = 9722.22(x - 0.06)^2$



b) L (0.027) =9722.22 $(0.027-0.06)^2$ =\$ 10.5875

This means that if the refrigerator thickness deviates 0.027mm from the specification it will cost Cool Food Inc. \$ 10.59.

6.2) Problem 27

a) The probability of one machine at each stage

Reliability = Reliability (Machine A) × Reliability (Machine B) × Reliability (Machine C)

= 0.85 × 0.92 × 0.90

= 0.7038

b) Reliability if both machines are used

Reliability = Reliability (A1 and A2) × Reliability (B1 and B2) × Reliability (C1 and C2)

$= (1 - (1 - 0.85)^2) \times (1 - (1 - 0.92)^2) \times (1 - (1 - 0.90)^2)$

= 0.9615

From these calculations its clear that if you put two similar machines in parallel it improves the reliability by more than 20% because of the fact that if one machine breaks the other one can still produce and send product down the production line without the production line coming to a standstill.

6.3) Calculating the expected reliability of a delivery process

Calculate the probability of having reliable vehicles: Using the binomial equation

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

$$\mathrm{P(x < 1) = [20C1 * p^1 * (1 - p)^{20-1}] * 15}$$
$$= 1560 - (190 + 22 + 3 + 1) = 1344$$

p= 0.03280011

$$\mathrm{P(2) = [20C2 * p^2 * (1 - p)^{20-2}] * 1560 = 190}$$

p= 0.0348579

$$\mathrm{P(3) = [20C3 * p^3 * (1 - p)^{20-3}] * 1560 = 21}$$

p= 0.02701039

$$\mathrm{P(4) = [20C4 * p^4 * (1 - p)^{20-4}] * 1560 = 3}$$

p= 0.02812168

$$\mathrm{P(5) = [20C5 * p^5 * (1 - p)^{20-5}] * 1560 = 1}$$

p= 0.03740828

Weighted average:

$$\frac{0.03280011 * 1344 + 0.0348579 * 190 + 0.02701039 * 22 + 0.02812168 * 3 + 0.03740828 * 1}{1560}$$

$= 0.03296304$

Reliable delivery days in a year for vehicles:

$$P(x < 2) = [20C2 * 0.03296304 * (1 - 0.03296304)^{20-2}] = 0.952017$$

$$P(x < 2) = [20C2 * 0.03296304^2 * (1 - 0.03296304)^{20-2}] * 365 = 347.4862$$

Calculate probability of having reliable drivers: Using binomial equation

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

$$P(x < 3) = [20C3 * p^3 * (1 - p)^{25-3}] * 1560 = 1560 - (95 + 6 + 1) = 1458$$

$$p = 0.0779277$$

$$P(4) = [20C4 * p^4 * (1 - p)^{25-4}] * 1560 = 95$$

$$p = 0.08493793$$

$$P(5) = [20C5 * p^5 * (1 - p)^{25-5}] * 1560 = 6$$

$$p = 0.0569216$$

$$P(6) = [20C6 * p^6 * (1 - p)^{25-6}] * 1560 = 1$$

$$p = 0.05803208$$

Weighted average:

$$\frac{0.0779277 * 1458 + 0.08493793 * 95 + 0.0569216 * 6 + 0.05803208 * 1}{1560} = 0.07826106$$

Reliable delivery days in a year for drivers:

$$P(x < 4) = [20C4 * 0.05738262^{4} * (1 - 0.05738262)^{20-4}] = 0.9831905$$

$$P(x < 4) = [20C4 * 0.05738262^{4} * (1 - 0.05738262)^{20-4}] * 365 = 358.8645$$

Total reliable days

P(total) = P(vehicles) * P(drivers) = 0.952017*0.9831905= 0.936014

Total reliable days = 341.6451

Part 2: Increasing the number of vehicles to 21

Using the same binomial equation as before

$$P(21) = [21C0 * p^0 * (1-p)^{21-0}] = 0.4946574$$

$P_{total} = 0.4946574 + 0.936014 = 1.430671$

Number of reliable days in a year: 1.430671*365= 365 days

This means that there will be no unreliable delivery days in the year when the number of vehicles is increased from 20 to 21.

Conclusions

In the first part of this report the data set is cleaned and processed in such a way that the new clean data can be used to make insightful observations. After this a better understanding of the data was gained by using graphs, tables, etc. Constructing control charts was next in line. These charts were used to analyse if classes are in control or out of control. This step in the process in necessary for business to identify problem areas that need investigation.

The reliability of delivery times was then calculated to give the business an indication if they should investigate their delivery process or not. MONAVO was then used to indicate the same results and aid the business in their decision-making process.

The probability of making a type I and II error in terms of delivery times was also calculated and analysed.

In conclusion the report focussed on the importance of explorative data analysis and how a company can use it to their advantage.

References

1) https://towardsdatascience.com/7-data-wrangling-r-functions-for-your-next-data-science-project-in-under-5-minutes-d5a4ad55f99b

2) https://www.techtarget.com/searchenterpriseai/definition/data-splitting#:~:text=Data%20splitting%20is%20when%20data,creating%20models%20based%20on%20data.

3) https://asq.org/quality-resources/statistical-process-control#:~:text=Statistical%20process%20control%20(SPC)%20is,find%20solutions%20for%20production%20issues.

4) https://sixsigmastudyguide.com/x-bar-s-chart/

5) https://asq.org/quality-resources/control-chart#:~:text=The%20control%20chart%20is%20a,are%20determined%20from%20historical%20data.

6) http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance

7) https://www.smartcapitalmind.com/what-is-delivery-reliability.htm