ECSA Project

Quality Assurance 344

by Student, Kyla Meyer

3rd year BEng Industrial

University of Stellenbosch

17 October 2022

Abstract

A sales dataset of an online business is visualised and analysed to make conclusions about the business's performance and to identify where some improvements can be made to enhance efficiency and effectiveness. It is observed that some instances in the dataset contain missing values and that the *Price* feature has some negative values which are removed to ensure accurate analysis. Using different descriptive statistics techniques, gifts- and technology items are the most frequently sold, most sales made are priced below R10 000 and are made due to items being recommended via some source, luxury items contribute the most to the overall revenue made, and at last, household items take the longest to be delivered. From the Statistical Process Control X-charts for all sales per class, one can observe that the delivery process of household, gifts, and luxury is out of control. It is advised that the latter should be investigated to identify the reasons for this behaviour. To obtain the best profit the delivery time process' center of twenty hours should be reduced by three hours. Reliable delivery can be expected for 361 days within a year consisting of 365 days.

Table of Contents

Abstract 2

Table of Figures 5

Introduction 6

1. Data wrangling 7

2. Descriptive statistics 9

2.1.1. Class 9

2.1.2. Age 10

2.1.3. Price 11

2.1.4. Why bought 12

2.1.5. Sales vs Year 13

2.1.6. Price vs Class 14

2.1.7. Price vs Age 14

2.1.8. Price vs Delivery time 15

2.1.9. Age vs Class 16

2.1.10. Delivery time vs Class 17

2.1.11. Delivery time vs Year 18

2.2. Process Capabilities 18

2.2.1. Process capability indices for the process delivery times (in hours) of the technology class items 19

3. Statistical Process Control (SPC) 20

3.1. SPC for first 30 samples 20

3.1.1. X-chart values 20

3.1.2. s-chart values 20

3.1.3. X- and s-charts (first 30 samples) 21

3.2. SPC for all samples, including the first 30 samples 23

3.2.1. X- and s-charts (all the samples) 23

4. Optimising the delivery processes 27

4.1. A. Sample means outside control limits (X-charts) 27

4.1. B. Most consecutive sample standard deviations between -0.3 sigma- and +0.4 sigma control limits (s-charts) 27

4.2. Hypothesis testing (Type I error for A and B) 28

4.2.1. Type I error (A) 29

4.2.2. Type I error (B) 29

4.3. Delivery process's center for the best profit 29

4.4. Hypothesis testing (Type II error for A) 30

5. Multivariate Analysis of Variance (MANOVA) 31

5.2.1. MANOVA 1 31

5.2.2. MANOVA 2 32
6. Reliability of the service and products 33
6.1. Taguchi Loss Function 33
6.2. System reliability 35
6.3. Delivery reliability 35
6.3.1. Vehicle reliability 36
6.3.2. Driver reliability 37
6.3.3. Vehicle and driver reliability 38
Conclusion 39
References 40

Table of Figures

Figure 1: Sales frequency vs Class 9
Figure 2: Sales frequency vs Age 10
Figure 3: Sales frequency vs Price 11
Figure 4: Sales frequency vs Why bought 12
Figure 5: Sales frequency vs Years 13
Figure 6: Prices vs Class 14
Figure 7: Price vs Age 14
Figure 8: Price vs Delivery time 15
Figure 9: Age vs Class 16
Figure 10: Delivery time vs Class 17
Figure 11: Delivery time vs Year 18
Figure 12: Sweets s-chart (30) 21
Figure 13: Sweets X-chart (30) 21
Figure 14: Household X-chart (30) 21
Figure 15: Household s-chart (30) 21
Figure 16: Gifts X-chart (30) 21
Figure 17: Gifts s-chart (30) 22
Figure 18: Technology X-chart (30) 22
Figure 19: Technology s-chart (30) 22
Figure 20: Luxury X-chart (30) 22
Figure 21: Luxury s-chart (30) 22
Figure 22: Clothing X-chart (30) 22
Figure 23: Clothing s-chart (30) 22
Figure 24: Food X-chart (30) 23
Figure 25: Food s-chart (30) 23
Figure 26: Sweets X-chart (all) 23
Figure 27: Sweets s-chart (all) 23
Figure 28: Household s-chart (all) 24
Figure 29: Household X-chart (all) 24
Figure 30: Gifts s-chart (all) 24
Figure 31: Gifts X-chart (all) 24
Figure 32: Technology X-chart (all) 25
Figure 33: Technology s-chart (all) 25
Figure 34: Luxury X-chart (all) 25
Figure 35: Luxury s-chart (all) 25
Figure 36: Clothing X-chart (all) 26
Figure 37: Clothing s-chart (all) 26
Figure 38: Food X-chart (all) 26
Figure 39: Food s-chart (all) 26
Figure 40: Standardised Normal Distribution 28
Figure 41: Total cost vs Hours to improve 30
Figure 42: Likelihood of making type II error for A 31
Figure 46: Link between Delivery time and Year 32
Figure 47: Taguchi Loss Function (Problem 6) 33
Figure 48: Taguchi Loss Function (Problem 7a) 34

Introduction

The importance of data analysis in assisting organisations in gaining a competitive edge in the market has grown significantly. The process of gathering and analysing data to identify patterns and trends and inform decision-making is known as statistical analysis.

A sales dataset of an online business selling a variety of items is analysed to obtain results needed to assist in future decision-making as this will enhance overall efficiency and effectiveness.

Before proceeding with descriptive statistics, the data is wrangled and cleaned to ensure accurate observations and analysis. The data is visualised and explained based on what can be observed from respective graphs.

Statistical Process Control (SPC) for the X- and s-charts is done for each class of items' delivery process times to identify any unusual process variations that in return are used by manufacturers to test for potential causes. Thereafter, the delivery processes are optimised and the probability of making a Type I and Type II error is estimated.

In addition to the latter, multivariate analysis of variance (MANOVA) is done on specific data and finally, the probability of having reliable delivery within a year is calculated and analysed.

1. Data wrangling

By correcting errors and improving the format of the data, data wrangling helps an organisation obtain the most accurate analytical and predictive findings (Furche et al., 2016:474). The issue is that you are unable to manage cleaning your data with no longer a manual procedure. In the modern world where datasets are not just large but continuous as well, data wrangling needs to be done quite carefully to ensure accurate predictions. Companies may swiftly make informed decisions to meet the needs of the business and their clients due to data wrangling. Avoiding data wrangling can have a lot of unintended repercussions. Organizations can more effectively handle upcoming discrepancies by properly preparing their data.

Since the dataset, *salesTable2022*, is already available, data collection is not needed. It is very unlikely for a dataset to be free of any data quality issues. Upon visual inspection, it may appear that *salesTable2022* does not suffer from any data quality issues but proceeding with this assumption might result in inaccurate analysis and prediction.

salesTable2022 consists of 180 000 rows/instances and 10 columns/descriptive features. The features can be split into qualitative - and quantitative features *Class* and *Why.Bought* and *X, ID, AGE, Price, Year, Month, Day,* and *Delivery.time*, respectively. The *summary()* function in RStudio is used to obtain the characteristics of the quantitative features, more specifically, each feature’s minimum - and maximum value, the 1st and 3rd quartile, and the mean, and median. “*X*” and “*ID*” can be considered irrelevant features thus no observation regarding their respective characteristics is needed.

Feature	**Minimum**	**1^{st} Quartile**	**Mean**	**3^{rd} Quartile**	**Maximum**	**Median**
AGE	18	38	54.57	70	108	53
Price	-588.8	482.3	12293.7	15270.7	116619	2259.6
Year	2021	2022	2025	2027	2029	2025
Month	1	4	6.521	10	12	7
Day	1	8	15.54	23	30	16
Delivery.time	0.5	3.0	14.5	18.5	75	3

Table 1: Continuous features's characteristics

Each feature's summarised characteristics have been analysed and it is noted that the feature "*Price*" has a minus minimum value (see Table 1 above). Since it is not possible to pay a negative amount for an item being purchased, negative values for "*Price*" are considered a potential data quality issue and are thus removed from the dataset. Data instances containing missing values (*NA*) for any feature are also omitted from the dataset. *salesTable2022* is split into two separate datasets, *invalidData*, containing the instances with negative "*Price*" values and *NA* values, and *validData,* consisting of the remaining 17 978 instances. *validData* is used for further analysis and prediction.

2. Descriptive statistics

Descriptive statistics are used to summarise data by explaining the relationship between features in a sample or population. Making inferential statistical comparisons should never occur before calculating descriptive statistics, which is an essential initial step in conducting research (Yellapu, 2018).

2.1.1. Class

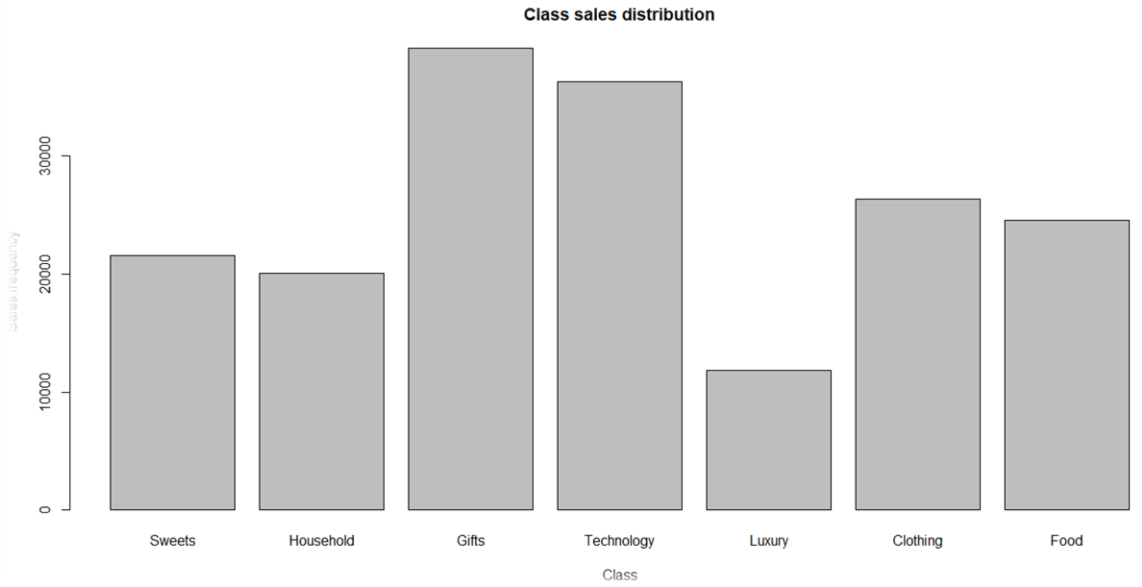


Figure 1: Sales frequency vs Class

The frequency of the different classes of items bought is unevenly distributed. The latter can be a result of several items being more affordable than others or the demand for specific goods might be higher than others. The frequency of sales for gifts and technology is the largest and this can be due to websites making it easy to order and purchase these types of items irrespective of where one is located. Clothing and food are basic needs, and it explains why these items are bought regularly as well. Household- and luxury items, and sweets have a lower frequency of sales possibly due to lower demand for them.

2.1.2. Age

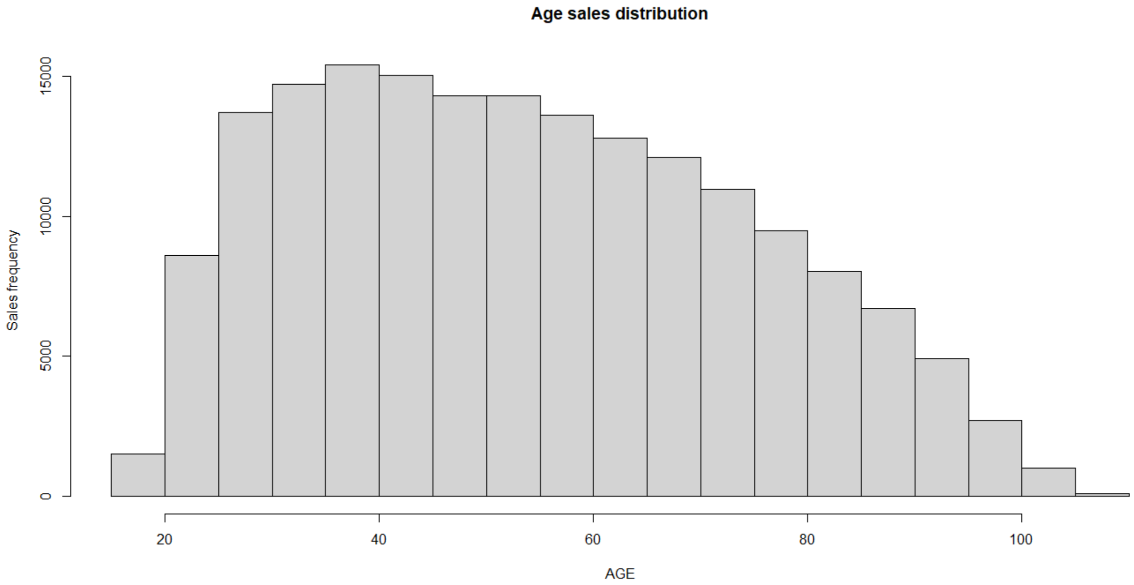


Figure 2: Sales frequency vs Age

The distribution of age (age of people buying goods) is skewed to the right due to the mean being larger than the median. A strong negative correlation is present considering age larger than forty because as age increases, the frequency of sales decreases. Most purchases are made by people aged between twenty-five and sixty-five. People aged between thirty-five and forty are responsible for the highest frequency of sales. The low frequency of sales for people aged younger than twenty and older than eighty-five is because younger people do not necessarily earn a salary and therefore depend on their parents, aged between thirty and fifty, and people aged above eighty-five do not visit shopping centers regularly.

2.1.3. Price

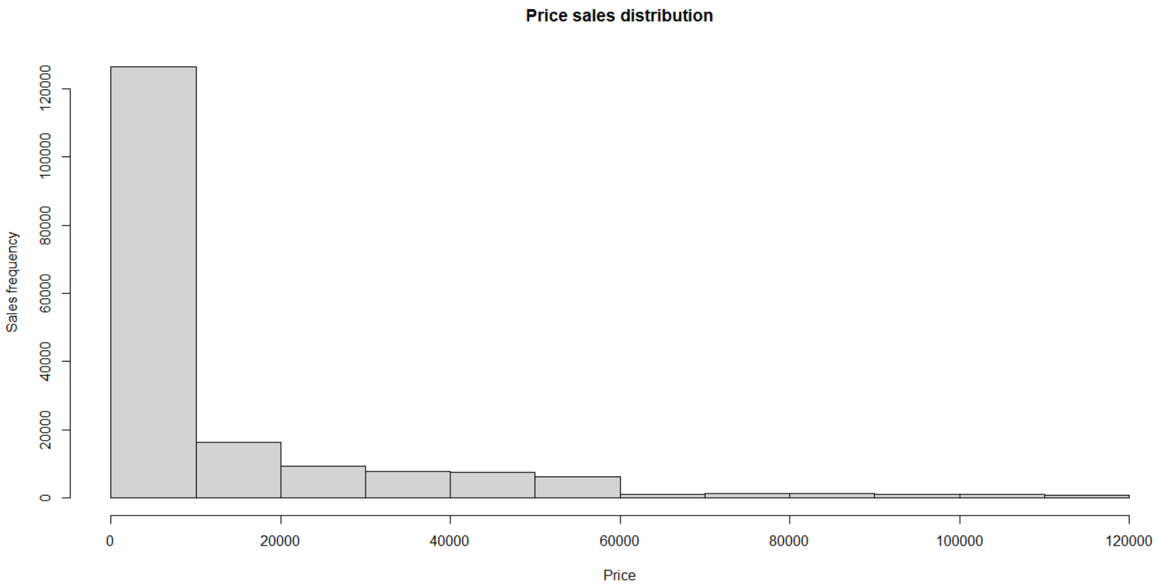


Figure 3: Sales frequency vs Price

Goods priced below R10 000, food, gifts, several luxury items, and technology items, are bought the most frequently as it is more affordable, resulting in uneven distribution. Goods having a value greater than R60 000, household items, and expensive luxury and technology items can potentially be considered outliers since the frequency of them being purchased are significantly lower than the rest of the goods valued below R60 000.

2.1.4. Why bought

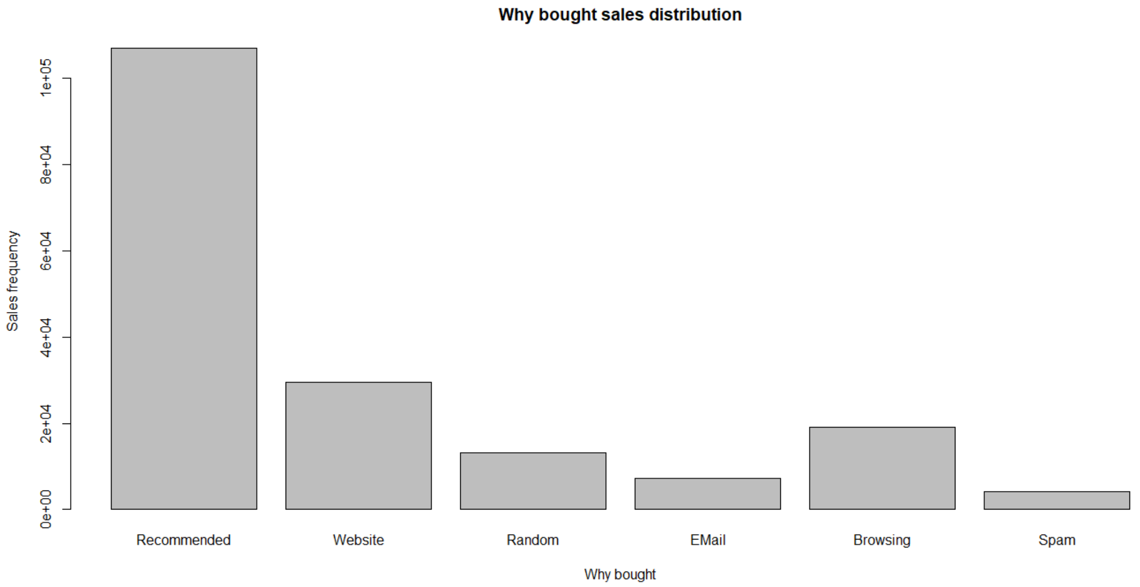


Figure 4: Sales frequency vs Why bought

Roughly 110 000 sales are made because of the items being recommended via some sort of communication. The distribution of the frequency of reasons for purchasing an item is uneven. 50 000 sales were made because of people browsing the internet and visiting the website. Email barely contributes to the overall number of sales and thus the business should not invest a lot in it.

2.1.5. Sales vs Year

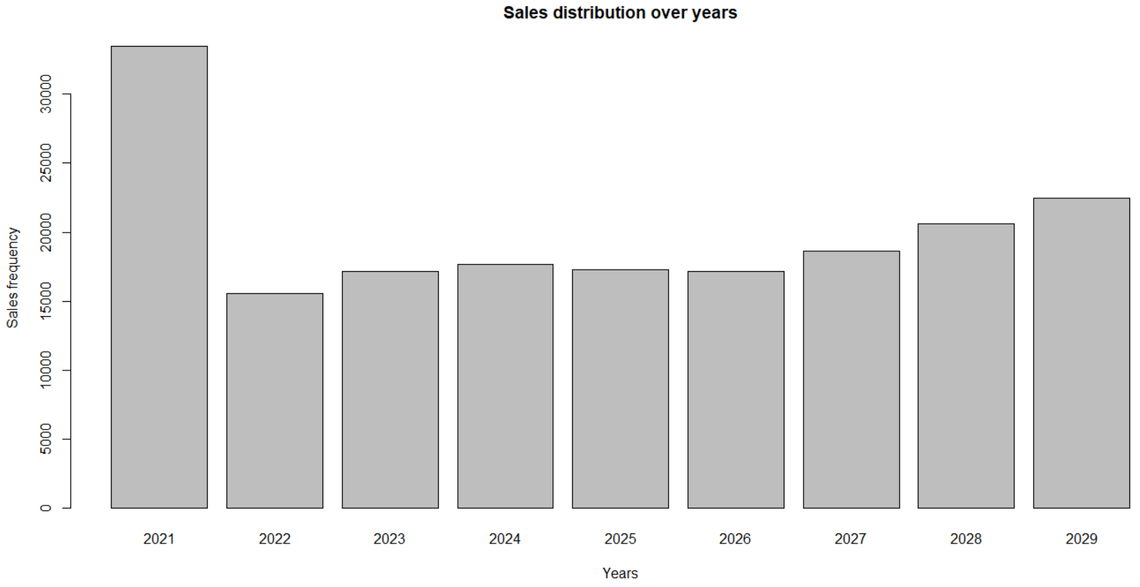


Figure 5: Sales frequency vs Years

From the year 2026 through to the year 2029, there is a clear positive correlation between the number of sales made and the year in which the sales are made. The latter can be due to items being recommended and advertised more as the company expands. In 2021, one-quarter of the total amount of sales was made followed by an aggressive decrease in the frequency of sales in the year 2022. The sales increased again up until the year 2024 and slightly dipped again in the years 2025 and 2026.

2.1.6. Price vs Class

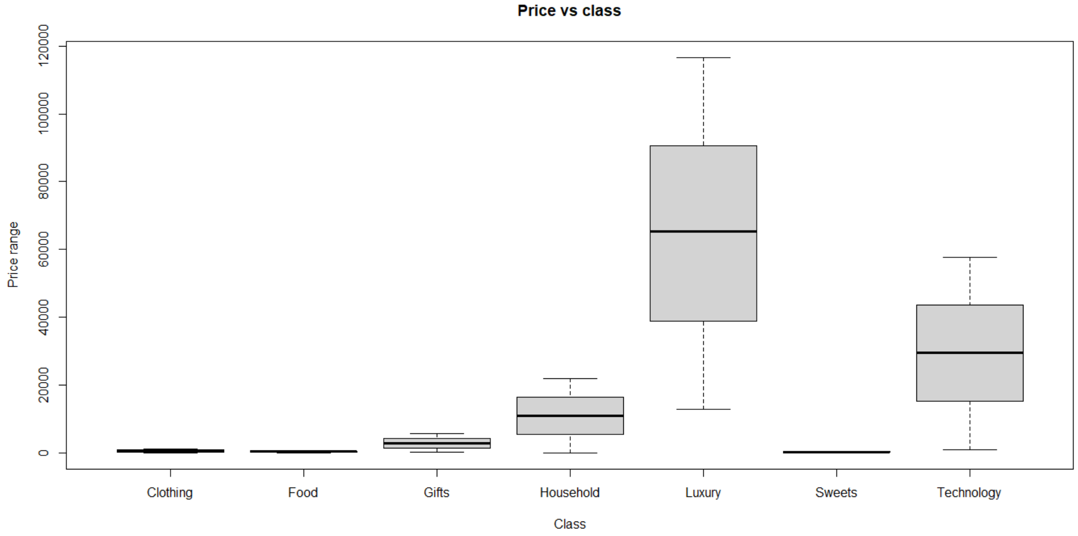


Figure 6: Prices vs Class

The price range for each class is normally distributed. The boxplots are symmetric and indicate that the mean price for each class is equal to the median price of each class, respectively. Luxury items are the most expensive with technology items being the second most expensive; the mean price for luxury and technology is extremely greater than that of clothing, food, gifts, and sweets. The price range of one class is independent of the price range of a different class.

2.1.7. Price vs Age

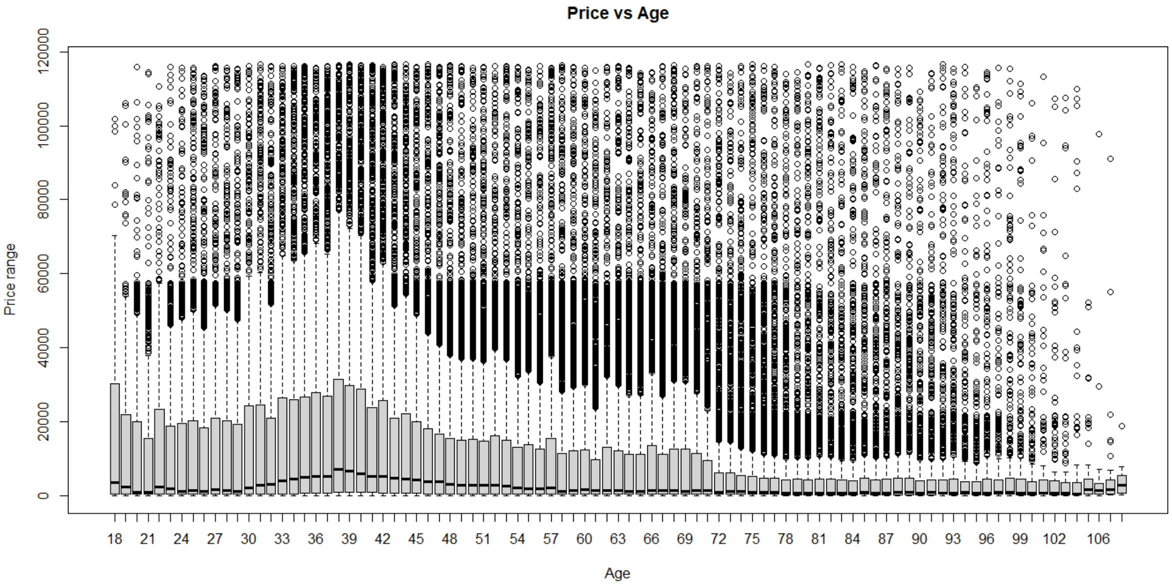


Figure 7: Price vs Age

The boxplots illustrate the price range for the different ages of the purchasers respectively. The distribution of the price range is skewed to the right for each age. The 1st and 3rd quartile values depend on if the subset of data has outliers. The 3rd quartile value for the price range of each age is large, confirming the presence of outliers. Without considering the average values for price for age groups eighteen to twenty-eight, the average value of the price increases up to and including age thirty-five and decreases thereafter, thus confirming that people between thirty and fifty-seven are willing or able to afford to pay more for items. It is clear, considering the latter observation, that business should see them, people between the ages of thirty and fifty-seven, as a priority because they contribute more to the overall income made.

2.1.8. Price vs Delivery time

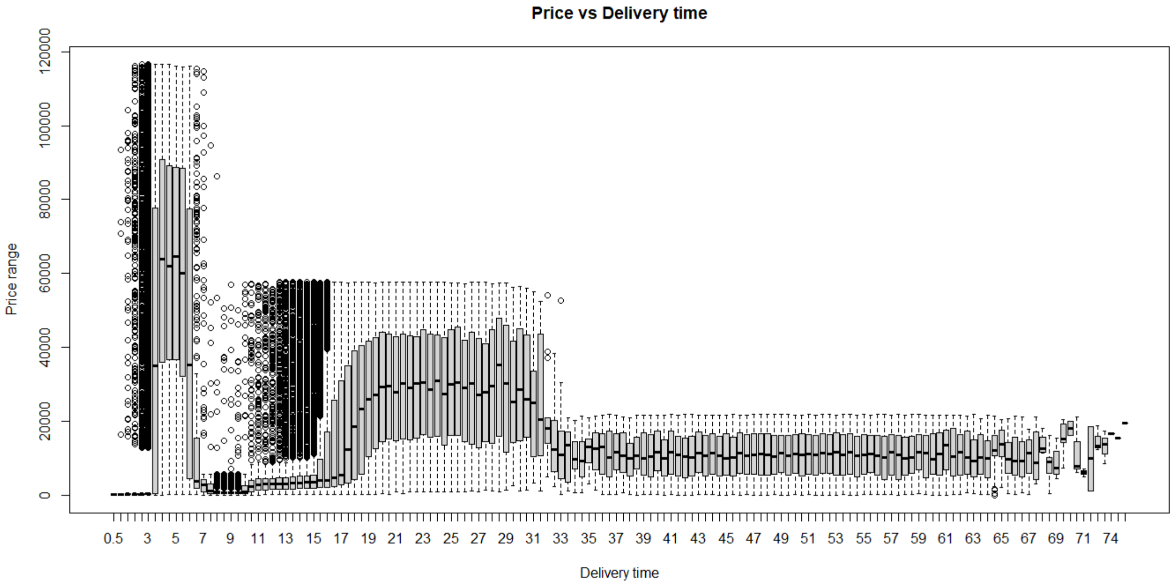


Figure 8: Price vs Delivery time

The boxplots for the price range corresponding to a set of delivery times are grouped as seen on the graph. Each group of boxplots represents a specific class. The time it takes to deliver items from one class differs from that of another and will be confirmed when analysing the Delivery time vs Class graph.

2.1.9. Age vs Class

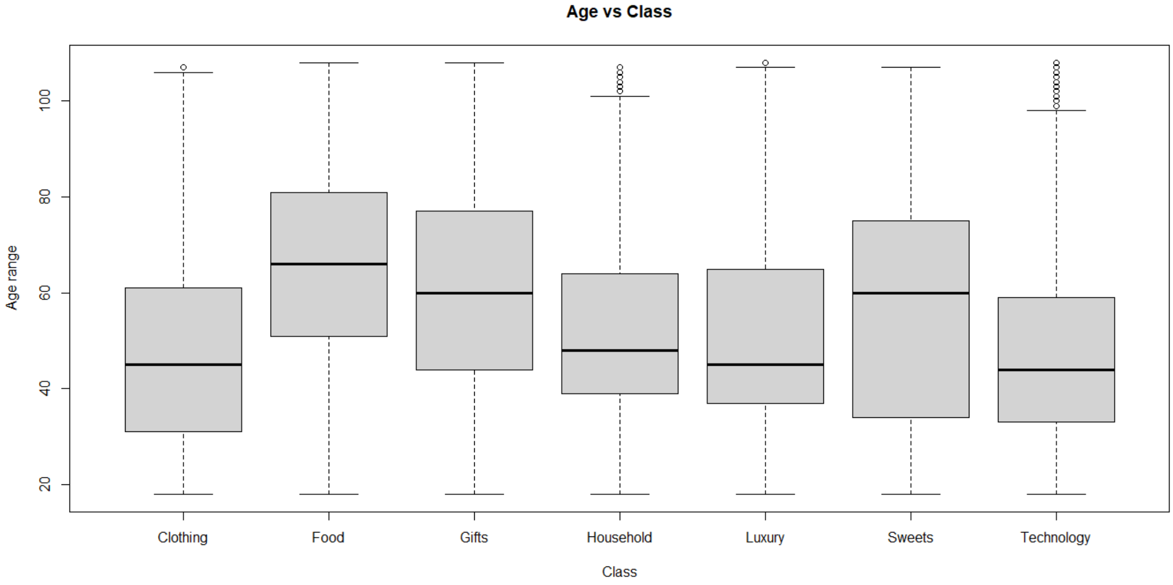


Figure 9: Age vs Class

The age range of people buying different class items is unevenly distributed but several observations can be made. The age range for household items, luxury items, and technology items is skewed to the right whereas the age range for sweets is skewed to the left. The age range for clothing, food, and gifts, considering ages between the 1st and 3rd quartiles, is normally distributed. These boxplots provide information regarding which age groups the business should prioritise for each respective class.

2.1.10. Delivery time vs Class

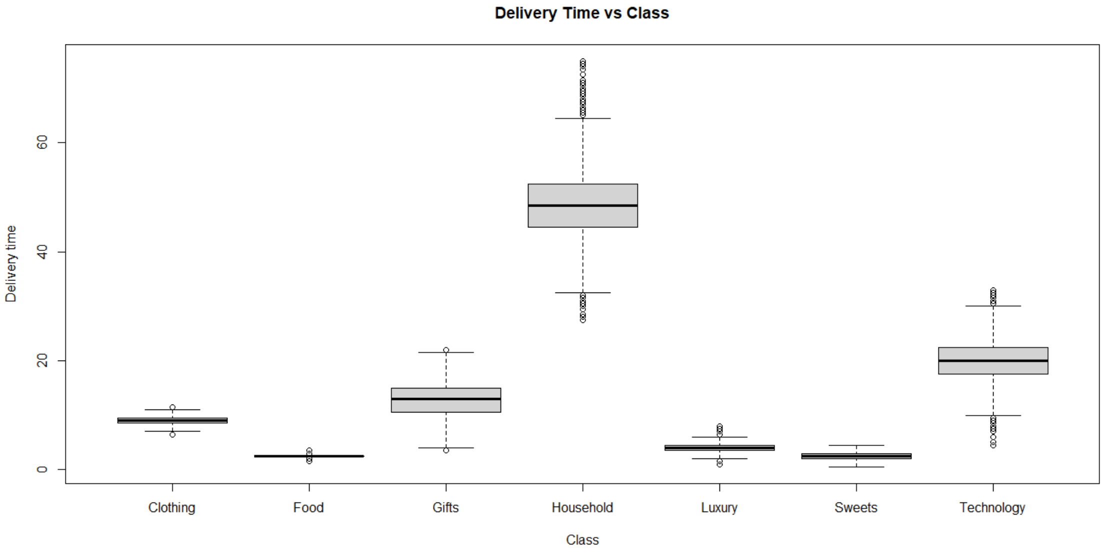


Figure 10: Delivery time vs Class

Each boxplot represents the delivery time range for items from the respective classes. Household items, technology items, and gifts take longer to deliver than clothing, food, luxury, and sweets. Household items have the largest mean delivery time whereas food and sweets have the shortest. The open circles confirm that there are several outliers present for each class and these cases should be investigated to obtain a more stable delivery time range approximation.

2.1.11. Delivery time vs Year

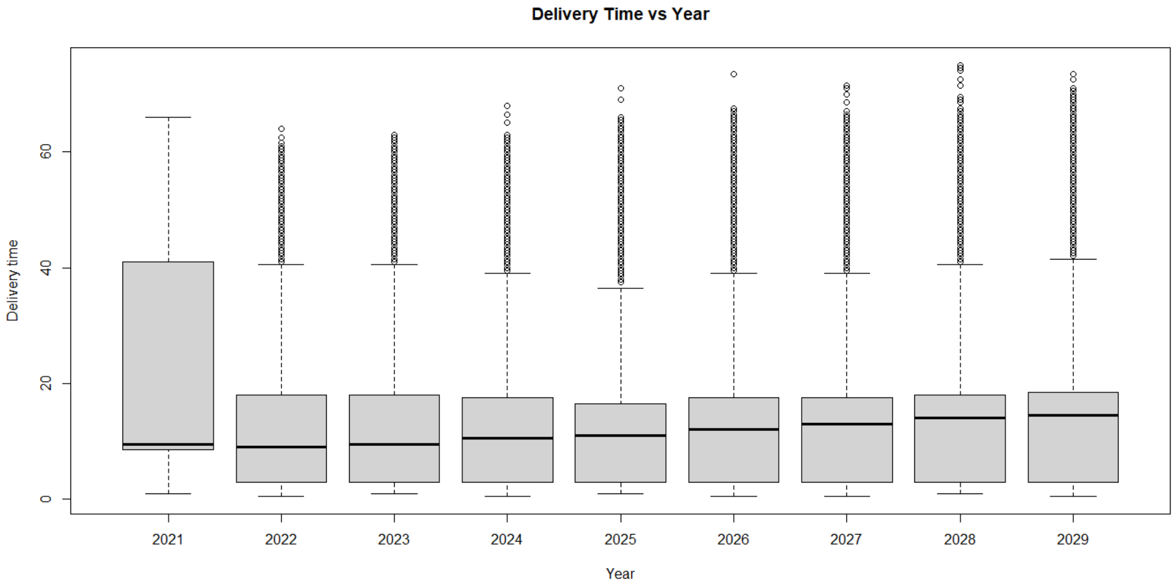


Figure 11: Delivery time vs Year

There is a positive correlation between the delivery time and the year in which items are delivered. As the year increases the average delivery time slightly increases. The distribution of the subsets of data for each year changes from skewed to the right to skewed to the left; the average delivery time becomes shorter than the median. The increase in the mean delivery time confirms the observation that the number of sales increases with time resulting in longer mean delivery times for each year.

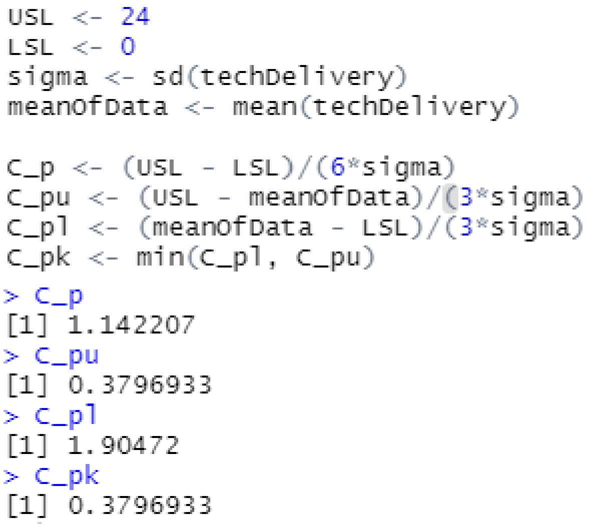
2.2. Process Capabilities

Utilising capability indices, process capability compares an in-control process's output to its specification's upper and lower bounds. When a process' capability is recognised and recorded, it may be used to prioritise the order of process improvements to be made, measure constant improvement using trends over time, and determine whether a process can achieve client needs (MacKay, Abraham, and Steiner, 1995).

Index	Equation	Definition
Cp	$\frac{(USL - LSL)}{6\sigma}$	Process capability for two-sided specification limits; does not consider where the process is centered (i.e., what the process average Xbar is)
Cpu	$\frac{(USL - Xbar)}{3\sigma}$	Process capability based on the upper specification limit
Cpl	$\frac{(Xbar - LSL)}{3\sigma}$	Process capability based on the lower specification limit
Cpk	$\min(Cpu, Cpk)$	Process capability for two-sided specification limits; consider where the process is centered (i.e., what the process average Xbar is)

Table 2: Process capabilities indices formulas

2.2.1. Process capability indices for the process delivery times (in hours) of the technology class items



```
USL <- 24
LSL <- 0
sigma <- sd(techDelivery)
meanOfData <- mean(techDelivery)

C_p <- (USL - LSL)/(6*sigma)
C_pu <- (USL - meanOfData)/(3*sigma)
C_pl <- (meanOfData - LSL)/(3*sigma)
C_pk <- min(C_pl, C_pu)
> C_p
[1] 1.142207
> C_pu
[1] 0.3796933
> C_pl
[1] 1.90472
> C_pk
[1] 0.3796933
```

Since the Cp value is greater than one, the process fits inside the respective specification limits. Although Cpl is greater than one, indicating that the process is capable at the lower tail of the distribution, Cpu is less than one indicating that the delivery time process for technology class items longs improvement.

An LSL of zero for the process delivery times of the technology class items is logical because when measuring time, it is impossible to obtain a negative value as time has a starting point of zero. In addition to the latter, it is possible to deliver an item instantly.

3. Statistical Process Control (SPC)

3.1. SPC for first 30 samples

Inline data obtained from the processes that produce the products are used in real time by statistical process control (SPC) during the production process. The process's state of control is then determined using statistical techniques. By offering a graphical representation of the process variance, this statistically based process information can aid in improving knowledge of the process (Fogliatto and Peres, 2018).

The validData dataset, containing the information regarding the sales made, is sorted according to date (year, month, and day) with the first instance representing the oldest sale made. Thirty samples having a sample size of fifteen instances are used to construct the X - and s- charts for the delivery process times of the seven processes (i.e., classes) by using the oldest data first.

3.1.1. X-chart values

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Sweets	2.897	2.7573	2.6175	2.4778	2.338	2.1983	2.0585
Household	50.2483	49.0196	47.7909	46.5622	45.3335	44.1048	42.8761
Gifts	9.4886	9.1127	8.7369	8.3611	7.9853	7.6095	7.2337
Technology	22.9746	22.1079	21.2412	20.3744	19.5077	18.641	17.7743
Luxury	5.494	5.2412	4.9884	4.7356	4.4828	4.2299	3.9771
Clothing	9.4049	9.26	9.115	8.97	8.825	8.68	8.5351
Food	2.7095	2.6363	2.5632	2.49	2.4168	2.3437	2.2705

Table 3: X-chart values

3.1.2. s-chart values

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Sweets	0.8353	0.734	0.6327	0.5314	0.4301	0.3288	0.2274
Household	7.3442	6.4534	5.5626	4.6719	3.7811	2.8903	1.9996
Gifts	2.2463	1.9739	1.7014	1.429	1.1565	0.8841	0.6116
Technology	5.1806	4.5522	3.9239	3.2955	2.6672	2.0388	1.4105
Luxury	1.5111	1.3278	1.1445	0.9612	0.778	0.5947	0.4114
Clothing	0.8666	0.7615	0.6564	0.5512	0.4461	0.341	0.2359
Food	0.4372	0.3842	0.3312	0.2781	0.2251	0.1721	0.119

Table 4: s-chart values

3.1.3. X- and s-charts (first 30 samples)

🟦 CL 🟩 U1/L1Sigma 🟥 U2/L2Sigma 🟨 UCL/LCL

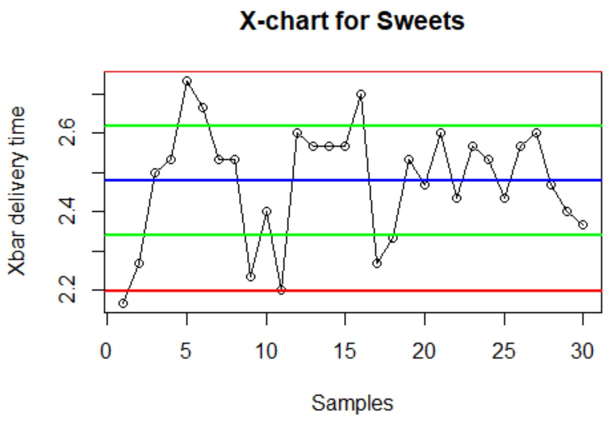


Figure 12: Sweets X-chart (30)

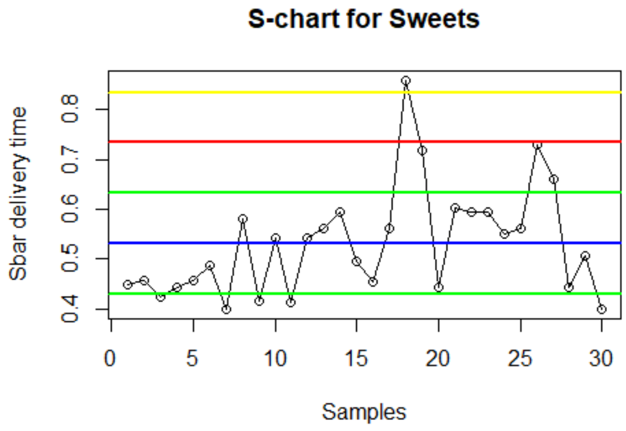


Figure 13: Sweets s-chart (30)

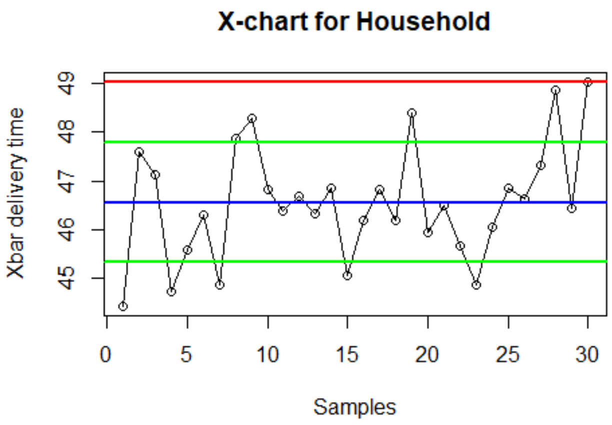


Figure 14: Household X-chart (30)

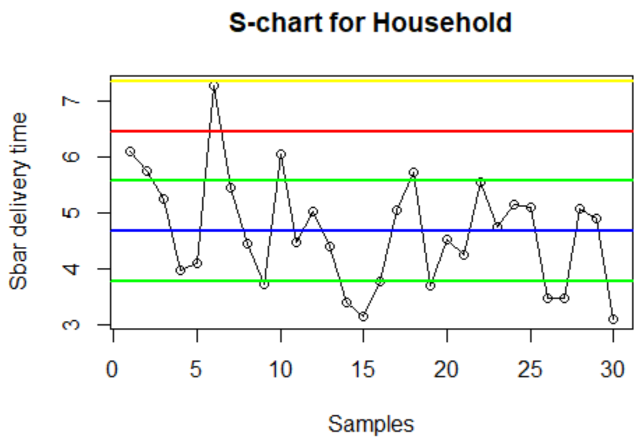


Figure 15: Household s-chart (30)

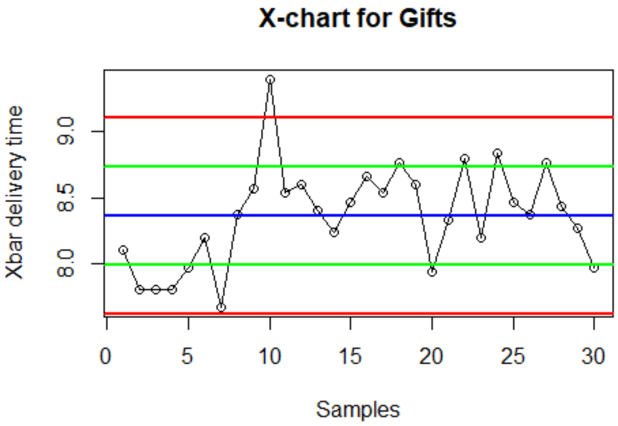
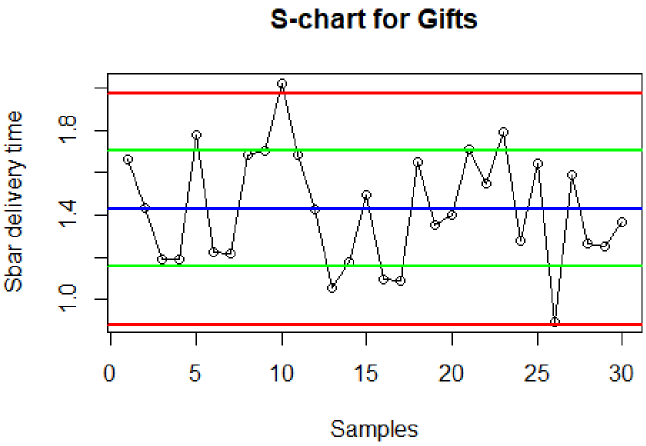


Figure 16: Gifts X-chart (30)



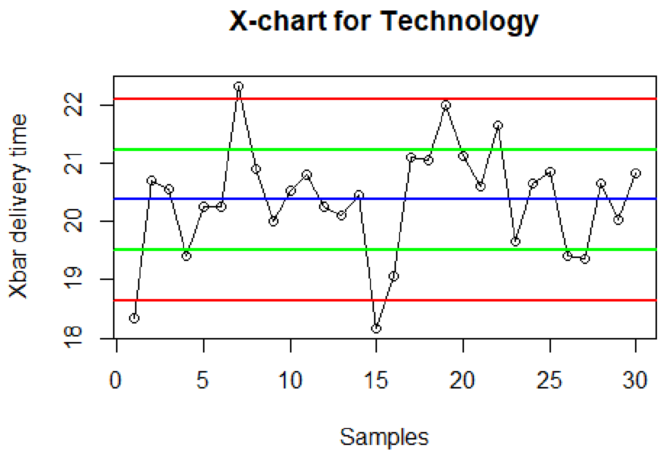


Figure 18: Technology X-chart (30)

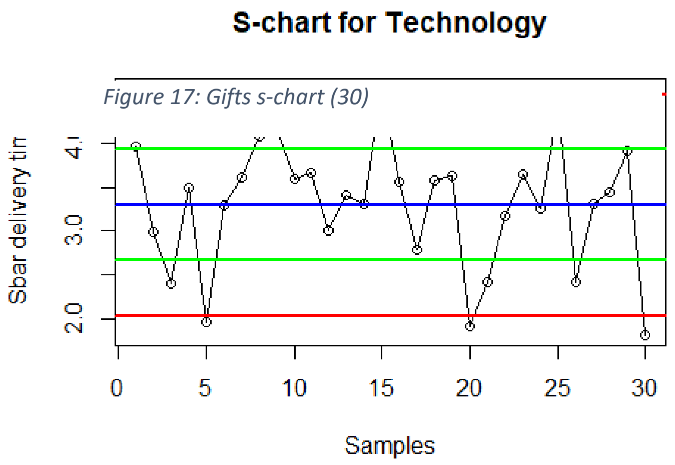


Figure 17: Gifts s-chart (30)

Figure 19: Technology s-chart (30)

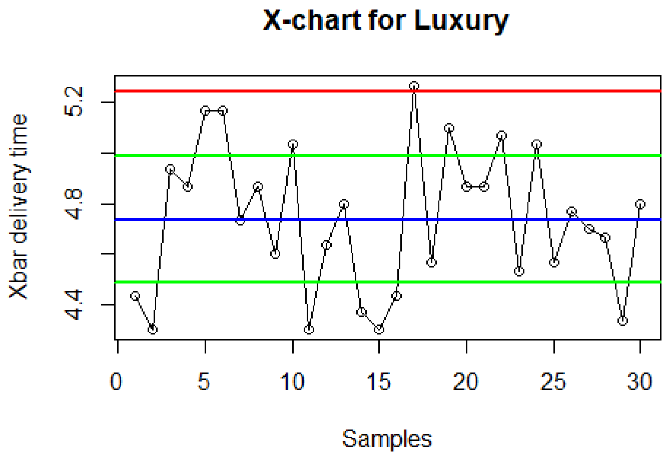


Figure 20: Luxury X-chart (30)

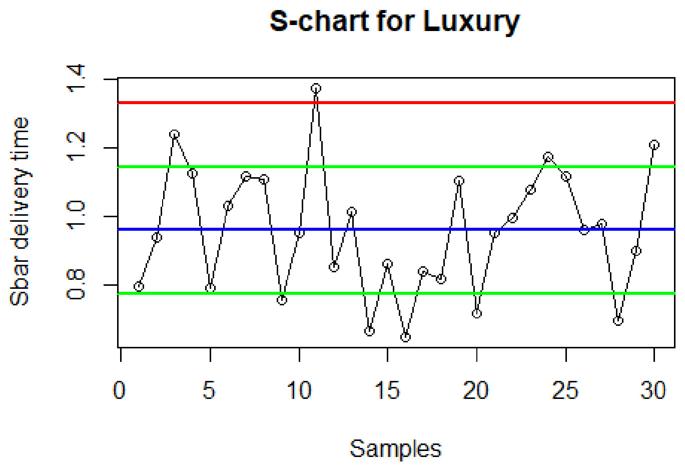


Figure 21: Luxury s-chart (30)

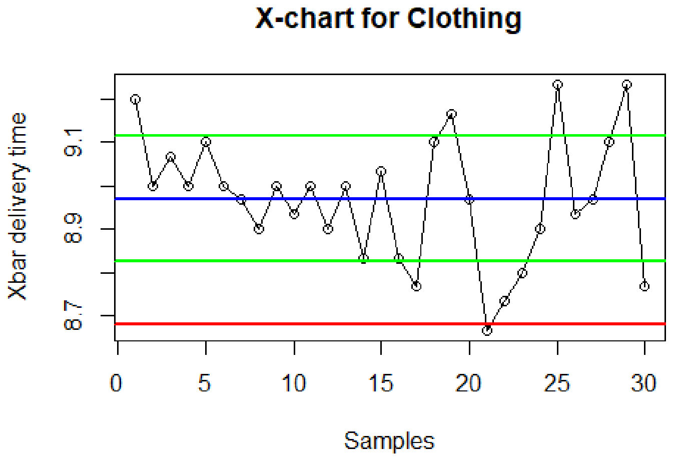


Figure 22: Clothing X-chart (30)

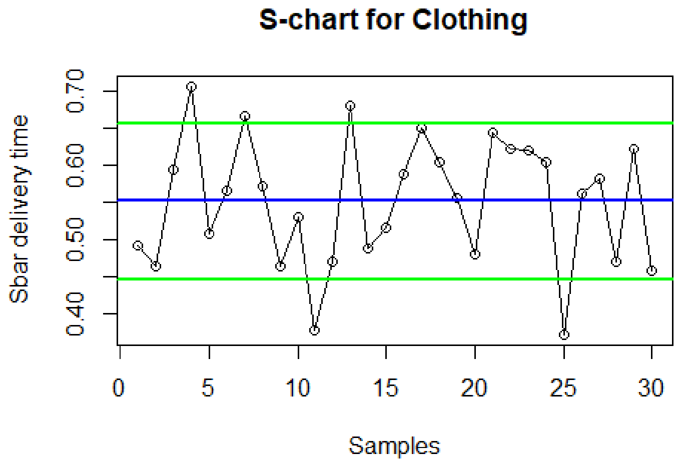


Figure 23: Clothing s-chart (30)

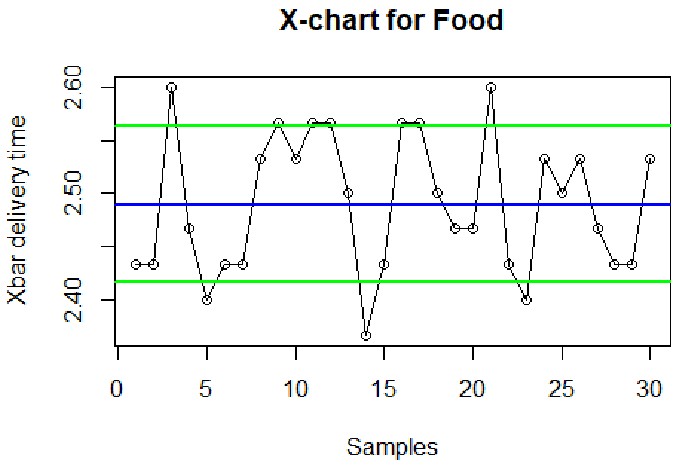


Figure 24: Food X-chart (30)

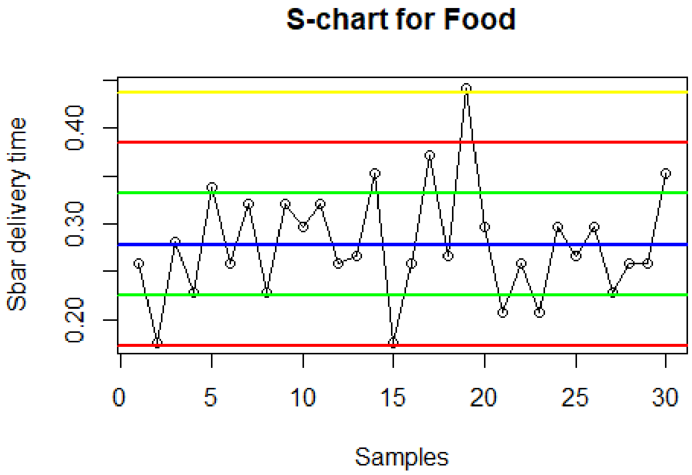


Figure 25: Food s-chart (30)

The upper control limit (UCL) and lower control limit (LCL) are placed 3 sigmas (of the samples plotted) above and below the center line (CL) respectively. If a high relative percentage of samples falls outside these limits, it can be said that the process is out of control. It is clear from the above control charts that nearly all samples fall within UCL and the LCL. The latter is an indication that the process for the delivery times of the respective classes is in control and stable thus one can conclude that little to no variation is present.

3.2. SPC for all samples, including the first 30 samples

3.2.1. X- and s-charts (all the samples)

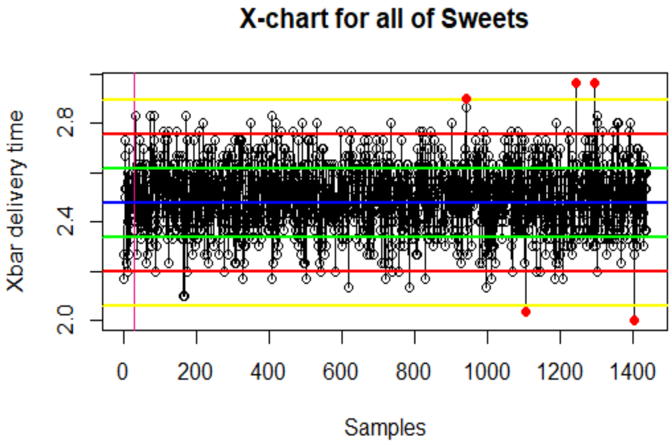


Figure 26: Sweets X-chart (all)

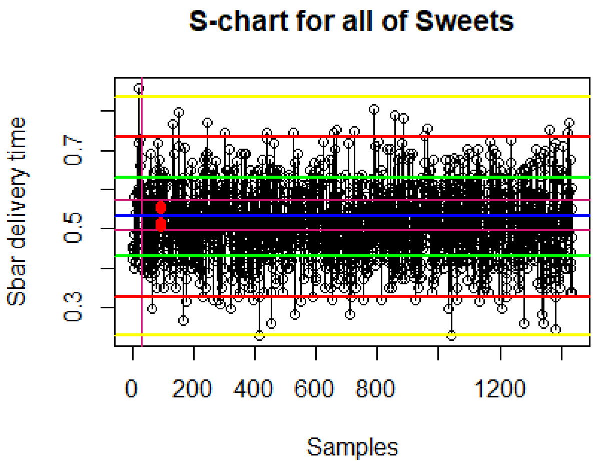


Figure 27: Sweets s-chart (all)

Number of samples = 1437

Five outliers are present in Figure 26 for sweets indicating that the process is relatively stable.

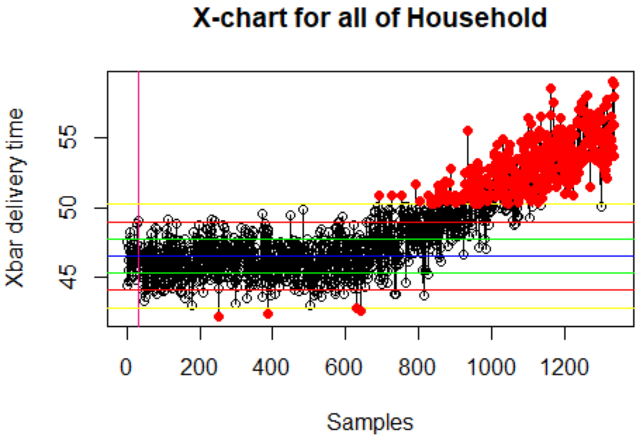


Figure 29: Household X-chart (all)

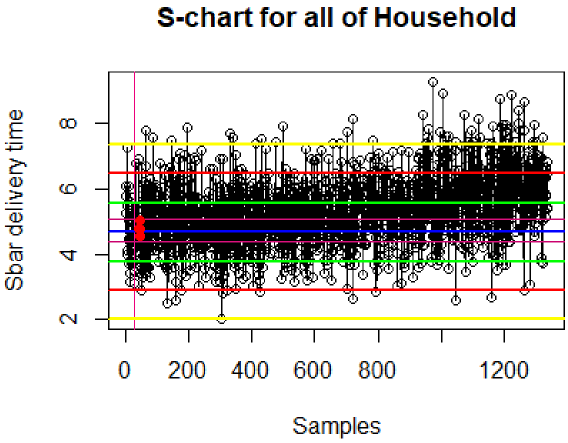


Figure 28: Household s-chart (all)

Number of samples = 1337

As can be seen from Figure 27, there are only five outliers in the first 650 samples. The process is thus relatively stable up until the 650th sample but becomes out of control as the mean delivery time of samples increases as time goes by. The latter is not desired, and an investigation of this process is recommended.

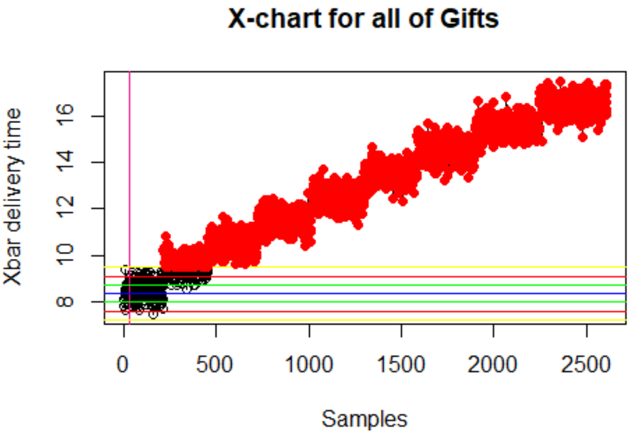


Figure 31: Gifts X-chart (all)

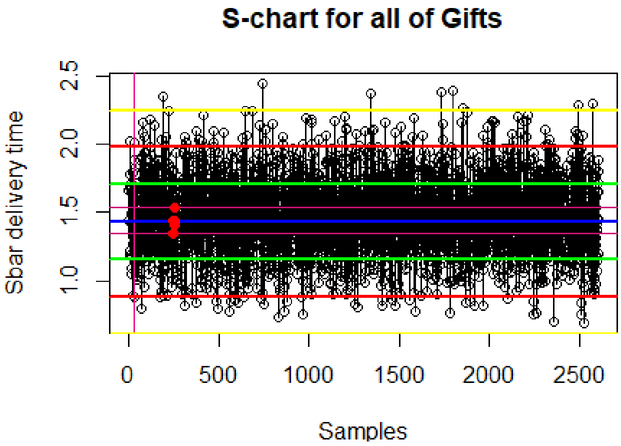


Figure 30: Gifts s-chart (all)

Number of samples = 2609

The mean delivery time for gift samples within the same year increases each year as illustrated in Figure 28. The reason behind this should be found because the process is out of control and unstable. One reason could be that gifts are not seen as a priority because it does not contribute much to the overall revenue made.

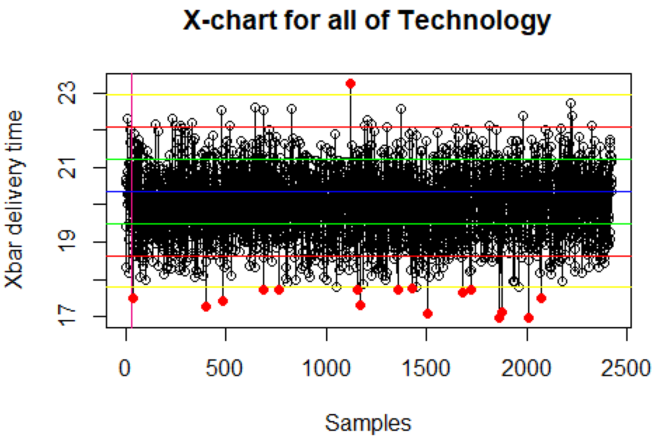


Figure 32: Technology X-chart (all)

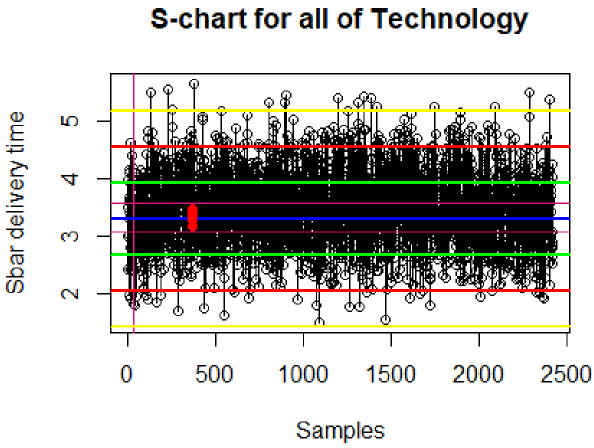


Figure 33: Technology s-chart (all)

Number of samples = 2423

In Figure 29, only seventeen outliers are present and thus make up a small percentage of the overall samples present in the process. It can be concluded that the process is relatively in control and stable with little variation.

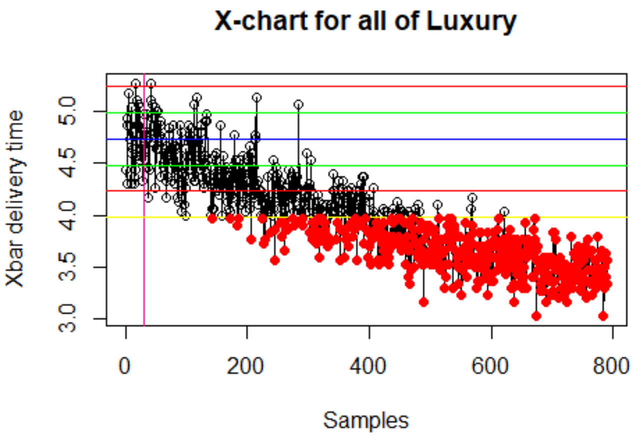


Figure 34: Luxury X-chart (all)

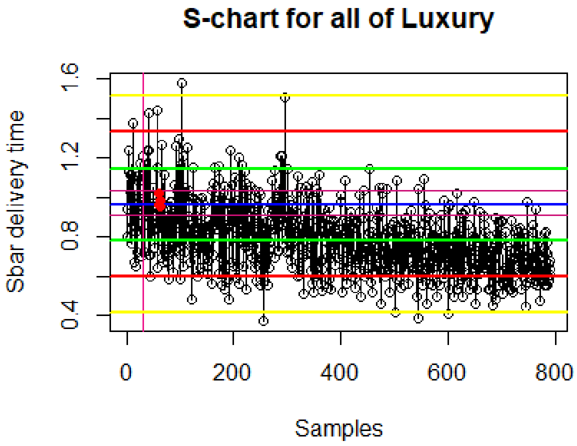


Figure 35: Luxury s-chart (all)

Number of samples = 791

The luxury delivery time process is stable for approximately the first hundred samples. As the years pass, the process systematically becomes unstable and out of control. It is clear from Figure 6 that luxury makes a significant contribution to the overall revenue made by the company and is thus seen as a top priority. The decrease in delivery times for luxury items is desired because of the reasons mentioned.

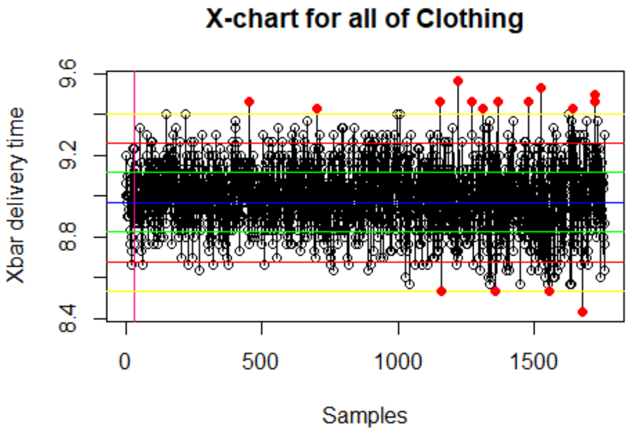


Figure 36: Clothing X-chart (all)

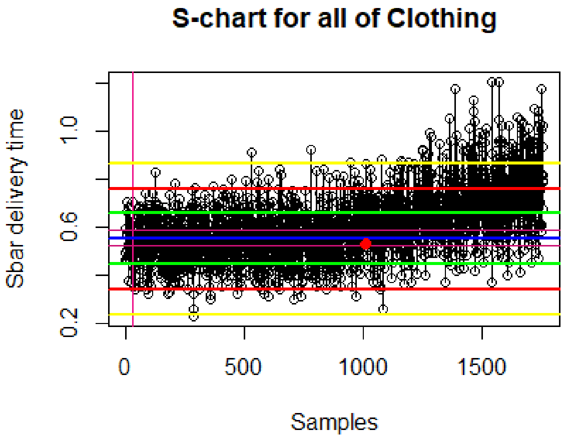


Figure 37: Clothing s-chart (all)

Number of samples = 1760

Seventeen samples are lying outside the UCL and LCL as indicated in Figure 31, indicating some variation within the process. The process is somewhat in control.

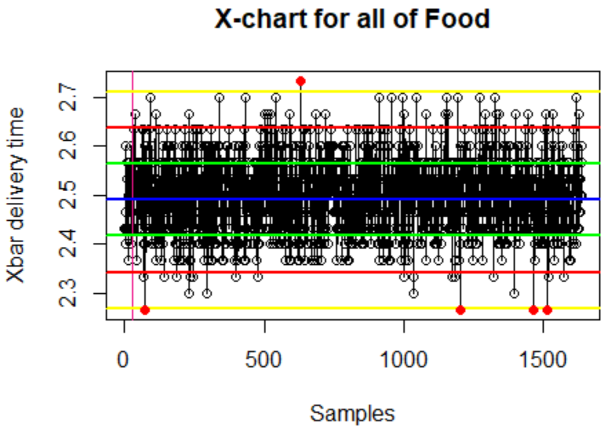


Figure 38: Food X-chart (all)

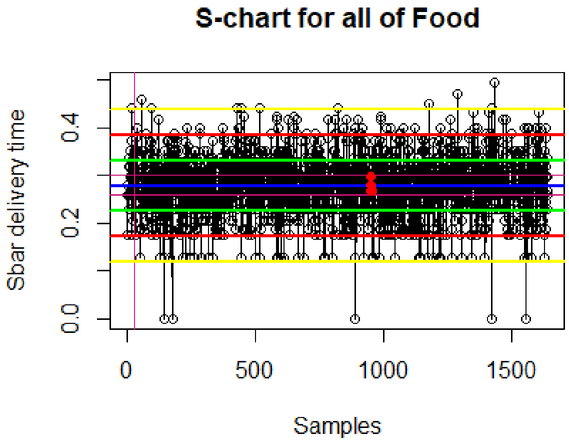


Figure 39: Food s-chart (all)

Number of samples = 1638

Five samples are not within the UCL and LCL, less than 0.5% of all samples. Only some variation is present in the process. The process is in control with no variation, specifically in the most present years.

4. Optimising the delivery processes

4.1. A. Sample means outside control limits (X-charts)

Class	**Total**	**1st**	**2nd**	**3rd**	**3rd last**	**2nd last**	**last**
Sweets	5	942	-	-	-	-	1403
Household	400	252	387	629	1335	1336	1337
Gifts	2290	213	216	218	2604	2605	2609
Technology	17	37	398	483	1872	2009	2071
Luxury	434	142	171	184	789	790	791
Clothing	17	455	702	1152	1677	1723	1724
Food	5	75	-	-	-	-	1515

Table 5: Sample means outside control limits

As indicated by red-filled circles on the X-charts and explained in 3.2., a disproportionate amount of delivery time sample means for household, gifts, and luxury lie outside the respective control limits resulting in excessive variation within the processes. For household and gifts specifically, the sample mean increases with time (years) thus, the processes should be investigated as this is not desired. In contrast to the latter, the sample means for luxury delivery time decreases over time and is not considered a problem because, as already mentioned, luxury is seen as a top priority and a shorter delivery time is desired.

4.1. B. Most consecutive sample standard deviations between -0.3 sigma- and +0.4 sigma control limits (s-charts)

Class	**Total samples within the range**	**Last sample within the range**
Sweets	4	94
Household	3	45
Gifts	5	254
Technology	6	372
Luxury	4	63
Clothing	4	1013
Food	7	952

Table 6:Consecutive sample standard deviations

The red-filled circles on the s-charts in 3.2. illustrate the most consecutive samples which fall within the -0.3 and +0.4 sigma-control limits. As seen in Table 6, the maximum consecutive sample standard deviations are seven for food items. For every class a very low percentage of sample standard deviations are consecutive.

4.2. Hypothesis testing (Type I error for A and B)

A type of statistical inference known as hypothesis testing uses data from a sample to make inferences about a population parameter or population probability distribution. An educated guess is first made on the parameter or distribution. The null hypothesis is also known as Ho because it is the default assumption. The opposite of what is said in the null hypothesis is then specified as an alternative hypothesis (designated Ha). Using sample data, the hypothesis-testing technique determines whether Ho may be rejected. The statistical conclusion is that the alternative hypothesis Ha is true if Ho is rejected (Mourougan and Sethuraman, 2017).

If a researcher rejects a null hypothesis that is true in the population, this is known as a type I error (false-positive). To calculate the likelihood of making a type 1 (Manufacture's) error for 4.1. A. and 4.1. B., it can be assumed Ho is: "the process is in control and centered on the centreline calculated using the first 30 samples".

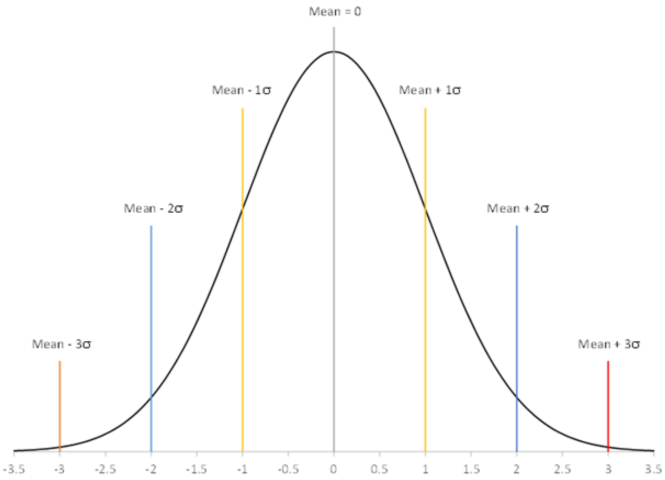
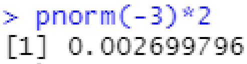


Figure 40: Standardised Normal Distribution

4.2.1. Type I error (A)

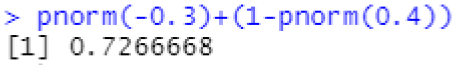
The probability of making a type 1 error for A, that is when a sample lies outside the lower- and upper control limits (“Mean - 3*sigma” and “Mean + 3*sigma” respectively) is 0.002699 (0.2699%) and is calculated using the R code, provided below.



```
> pnorm(-3)*2
[1] 0.002699796
```

4.2.2. Type I error (B)

The probability of making a type 1 error for B, that is when a sample lies outside the lower- and upper control limits (“Mean – 0.3*sigma” and “Mean + 0.4*sigma” respectively) is 0.7266668 (72.67%) and is calculated using the R code, provided below.



```
> pnorm(-0.3)+(1-pnorm(0.4))
[1] 0.7266668
```

4.3. Delivery process’s center for the best profit

It is every business’ desire to generate the best profit possible. When a technology item is not delivered within twenty-six hours a penalty cost of R329 per item per hour must be paid. It requires a cost of R2.5 per item per hour to reduce the delivery time. Therefore, the business must investigate if it is worth it to reduce the delivery time for each item.

45 347 technology sales were made and 1356 thereof had a delivery time of more than twenty-six hours. Since the penalty to be paid is much greater than the reduction cost, it is expected that the total amount to be paid for late deliveries will exceed the total reduction cost. It can be assumed that the shape of the process output distribution remains the same when the center line is moved. Therefore, the company must decide on how many hours the delivery process center should be adjusted to obtain the best profit.

As can be seen in Figure 40 below, after approximately three hours, there is a strong positive linear relationship between Total cost and Hours to improve. It is calculated (see R code) and indicated (with a red-filled circle) that if the delivery time for each item is reduced by three hours, it will cost R340 870, which will result in the best profit. Currently the mean delivery time is approximately twenty hours but will then be reduced to seventeen hours based on the calculated three-hour delivery time reduction.

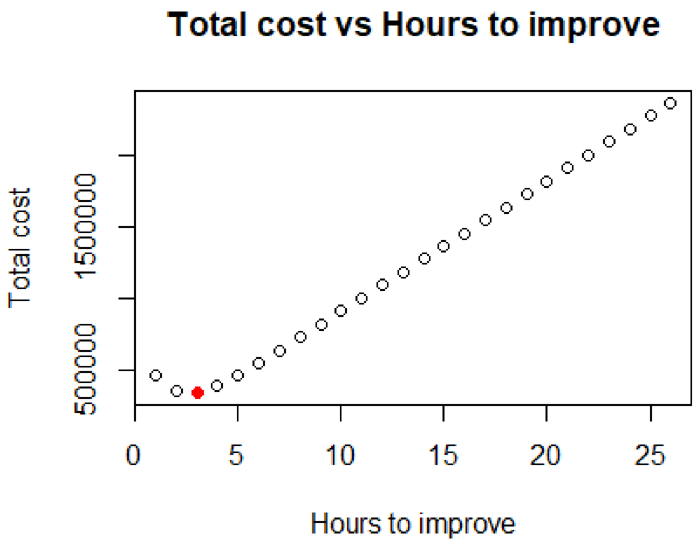
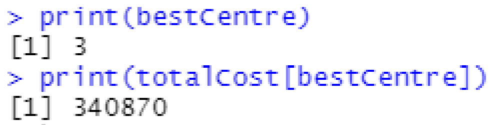


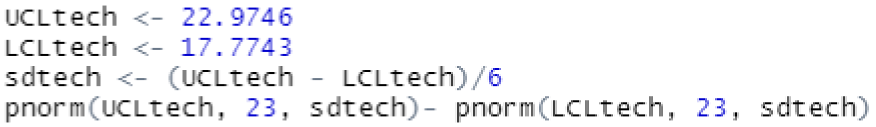
Figure 41: Total cost vs Hours to improve



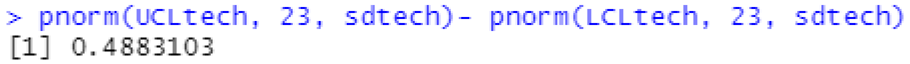
```
> print(bestCentre)
[1] 3
> print(totalCost[bestCentre])
[1] 340870
```

4.4. Hypothesis testing (Type II error for A)

When one fails to reject a wrong null hypothesis, this error is known statistically as a type II error. The probability of making a type II (Consumer's) error for 4.1. A. for technology class items, given that the delivery process average moves to 23 hours, is calculated in R (see below) to be 0.4883103 (48.83%). Thus, the likelihood of making a type II error is significantly high.



```
UCLtech <- 22.9746
LCLtech <- 17.7743
sdtech <- (UCLtech - LCLtech)/6
pnorm(UCLtech, 23, sdtech)- pnorm(LCLtech, 23, sdtech)
```



```
> pnorm(UCLtech, 23, sdtech)- pnorm(LCLtech, 23, sdtech)
[1] 0.4883103
```

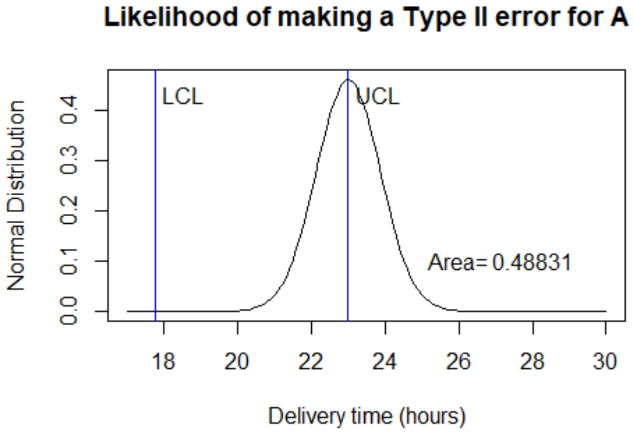


Figure 42: Likelihood of making type II error for A

5. Multivariate Analysis of Variance (MANOVA)

The statistical significance of the impact of one or more independent variables on a collection of two or more dependent variables is evaluated using MANOVA. The main goal of MANOVA is to find out if various levels of independent variables have an impact on the dependent variables individually or in combination with one another (Weinfurt*,* 2000).

Based on earlier analysis it seems that there is a link between the *Class* feature and the *Price* and *Delivery.time* features. A link between the *Year* feature and the *Sales* and *Delivery.time* features do exist as well. Two MANOVAs are performed to assist with the confirmation of the mentioned assumptions.

5.2.1. MANOVA 1

Dependent variables: *Price* and *Delivery.time*
Independent variable: Class

Null hypothesis: The price and delivery time of a sold item **do not depend** on the class of the item.
Alternative hypothesis: The price and delivery time of a sold item **depend** on the class of the item.

The p-value obtained for *Price* vs *Class* and *Delivery.time* vs *Class* is 2.2e-16 which is less than the significance level of 0.05. The latter confirms that the dependent variables do depend on the independent variable and that the null hypothesis should be rejected. Figures 42 and 43 below clearly show that a link between the respective dependent variable and independent variable exists.

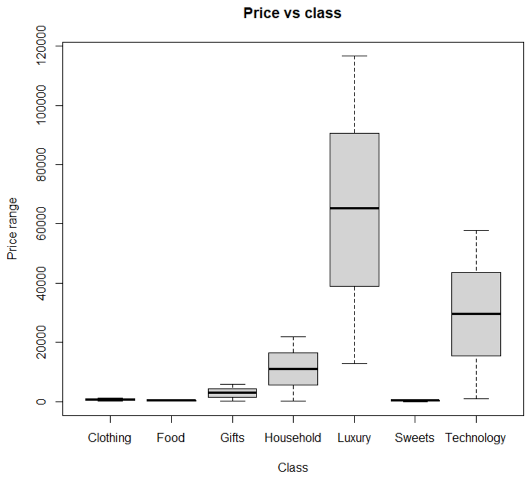


Figure 43: Link between Price and Class

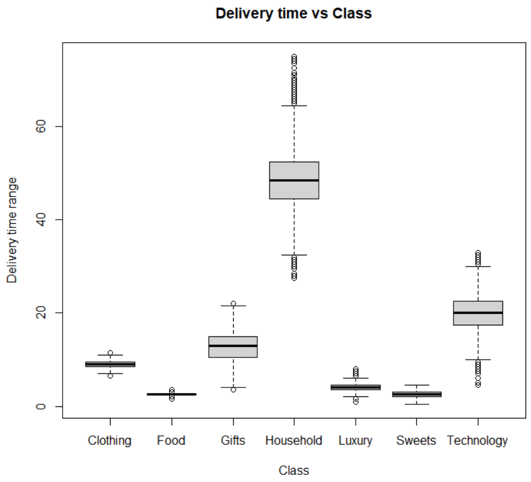


Figure 44: Link between Delivery time and Class

5.2.2. MANOVA 2

Dependent variables: Sales and *Delivery.time*

Independent variable: Year

Null hypothesis: The number of sales made per year and the delivery time of a sold item **do not depend** on the year.

Alternative hypothesis: The number of sales made per year and the delivery time of sold items **do depend** on the year.

The resulting p-values of 2.2e-16 for *Sales* vs *Year* and *Delivery.time* vs *Year* is much less than the significance level of 0.05 and the null hypothesis should be rejected. As seen in Figures 44 and 45 below, the number of sales made, and the delivery time of sold items do depend on the year in which the sales are made.

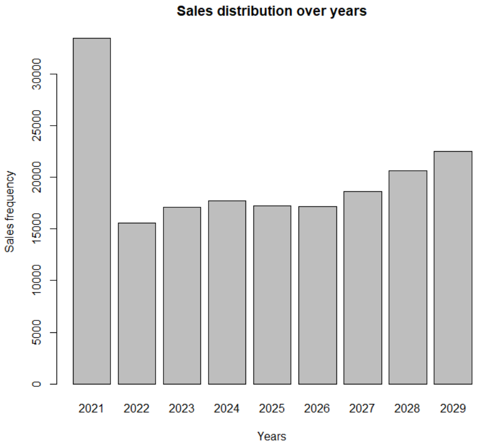


Figure 45: Link between Sales and Year

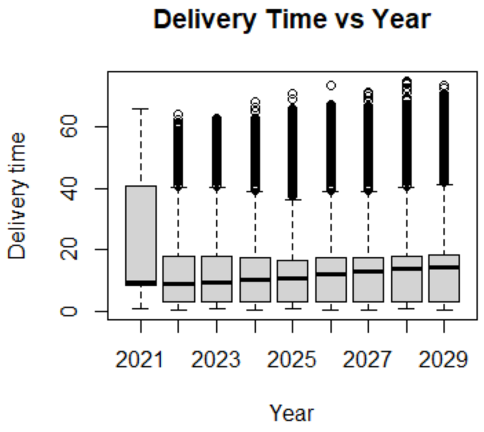
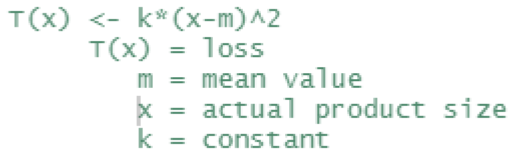


Figure 4643: Link between Delivery time and Year

6. Reliability of the service and products

6.1. Taguchi Loss Function

The Taguchi Loss Function is a graphical illustration of how an increase in variance within specification boundaries simultaneously increases customer dissatisfaction and the penalty/scrap cost to be paid exponentially.



```
T(x) <- k*(x-m)^2
     T(x) = loss
        m = mean value
        x = actual product size
        k = constant
```

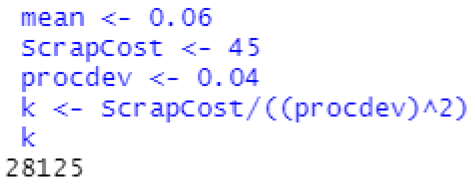
Problem 6

Question:

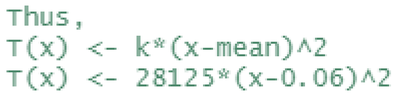
A blueprint specification for the thickness of a refrigerator part at Cool Food, Inc. is 0.06 +/- 0.04 centimeters (cm). It costs $45 to scrap a part that is outside the specifications.

Determine the Taguchi loss function for this situation.

Answer:



```
 mean <- 0.06
 scrapCost <- 45
 procdev <- 0.04
 k <- scrapCost/((procdev)^2)
 k
28125
```



```
Thus,
T(x) <- k*(x-mean)^2
T(x) <- 28125*(x-0.06)^2
```

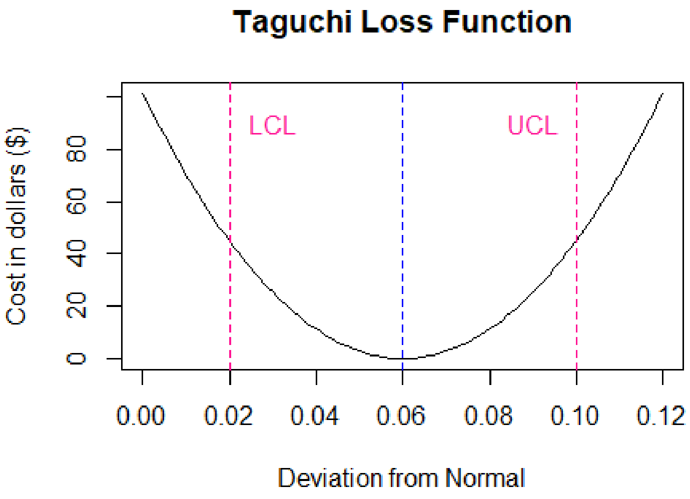


Figure 44: Taguchi Loss Function (Problem 6)

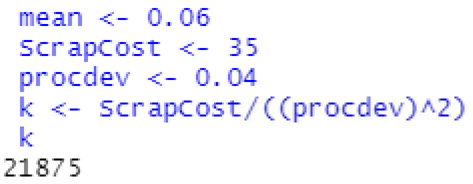
Problem 7

A team was formed to study the refrigerator part at Cool Food, Inc. described in Problem 6. While continuing to work to find the root cause of the scrap, they found a way to reduce the scrap cost to $35 per part.

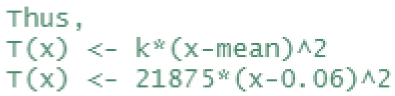
Question:

a. Determine the Taguchi loss function for this situation.

Answer:



```
mean <- 0.06
ScrapCost <- 35
procdev <- 0.04
k <- ScrapCost/((procdev)^2)
k
21875
```



```
Thus,
T(x) <- k*(x-mean)^2
T(x) <- 21875*(x-0.06)^2
```

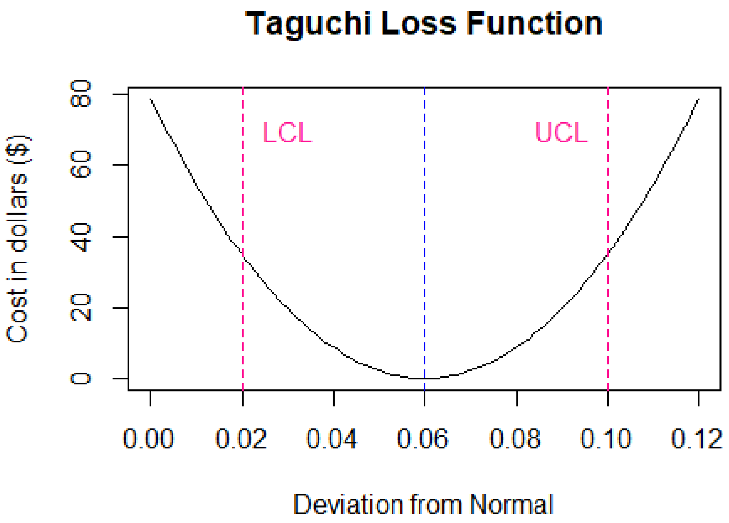
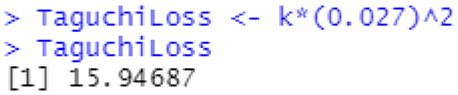


Figure 45: Taguchi Loss Function (Problem 7a)

Question:

b. If the process deviation from the target can be reduced to 0.027 cm, what is the Taguchi loss?

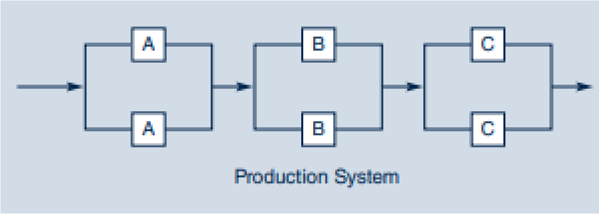
Answer:

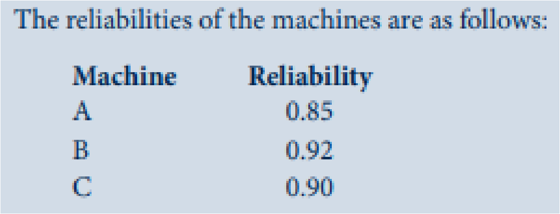


```
> TaguchiLoss <- k*(0.027)^2
> TaguchiLoss
[1] 15.94687
```

6.2. System reliability

Problem 27



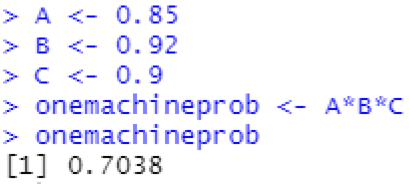

The reliabilities of the machines are as follows:

Machine	Reliability
A	0.85
B	0.92
C	0.90

Question:

a. Analyse the system reliability, assuming only one machine at each stage (all the backup machines are out of operation).

Answer:



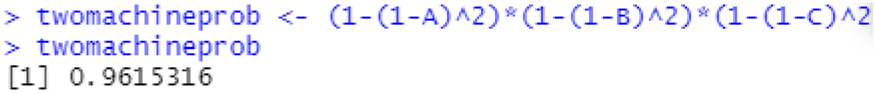
```
> A <- 0.85
> B <- 0.92
> C <- 0.9
> onemachineprob <- A*B*C
> onemachineprob
[1] 0.7038
```

With only one machine working at each stage, the production reliability is 70.38%.

Question:

b. How much is the reliability improved by having two machines at each stage

Answer:



```
> twomachineprob <- (1-(1-A)^2)*(1-(1-B)^2)*(1-(1-C)^2
> twomachineprob
[1] 0.9615316
```

Production reliability increases to 96.15% when both machines work in parallel at each stage.

6.3. Delivery reliability

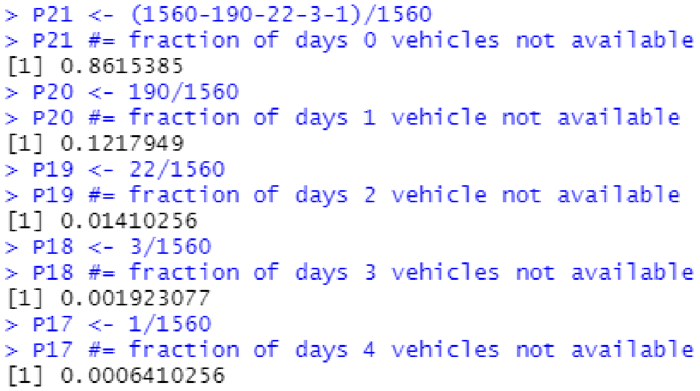
Vehicles available (/21)	Days available (/1560)
20	190
19	22
18	3
17	1

Drivers available (/21)	Days available (/1560)
20	95
19	6
18	1
17	-

For the rest of the 1560 days, it can be assumed that twenty-one vehicles and twenty-one drivers are available. It is required that at least twenty vehicles and twenty drivers should be available for the process to be classified as being reliable.

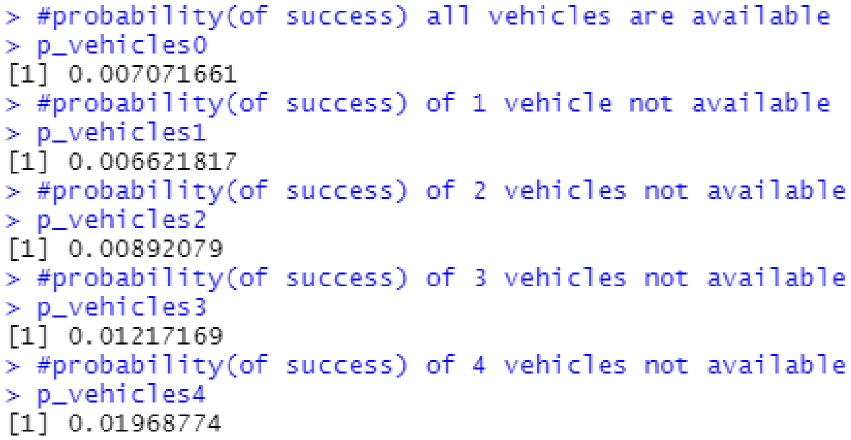
6.3.1. Vehicle reliability

Fraction of days x number of vehicles not available



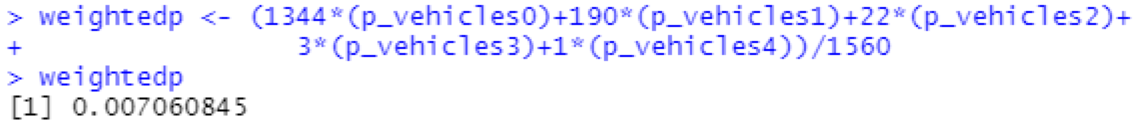
```
> P21 <- (1560-190-22-3-1)/1560
> P21 #= fraction of days 0 vehicles not available
[1] 0.8615385
> P20 <- 190/1560
> P20 #= fraction of days 1 vehicle not available
[1] 0.1217949
> P19 <- 22/1560
> P19 #= fraction of days 2 vehicle not available
[1] 0.01410256
> P18 <- 3/1560
> P18 #= fraction of days 3 vehicles not available
[1] 0.001923077
> P17 <- 1/1560
> P17 #= fraction of days 4 vehicles not available
[1] 0.0006410256
```

Probability of success of x number of vehicles not available



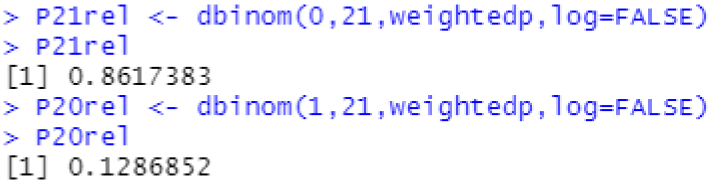
```
> #probability(of success) all vehicles are available
> p_vehicles0
[1] 0.007071661
> #probability(of success) of 1 vehicle not available
> p_vehicles1
[1] 0.006621817
> #probability(of success) of 2 vehicles not available
> p_vehicles2
[1] 0.00892079
> #probability(of success) of 3 vehicles not available
> p_vehicles3
[1] 0.01217169
> #probability(of success) of 4 vehicles not available
> p_vehicles4
[1] 0.01968774
```

Weighted probability of success



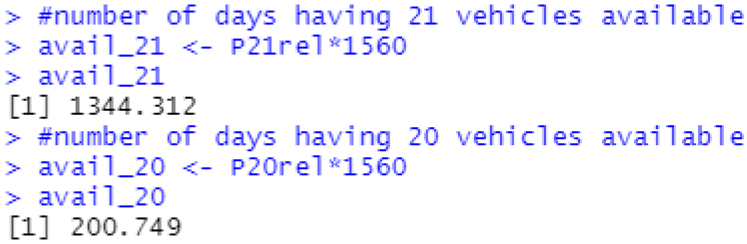
```
> weightedp <- (1344*(p_vehicles0)+190*(p_vehicles1)+22*(p_vehicles2)+
+                   3*(p_vehicles3)+1*(p_vehicles4))/1560
> weightedp
[1] 0.007060845
```

Probability that x number of vehicles will be available



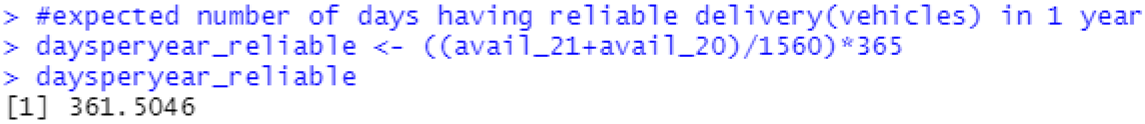
```
> P21rel <- dbinom(0,21,weightedp,log=FALSE)
> P21rel
[1] 0.8617383
> P20rel <- dbinom(1,21,weightedp,log=FALSE)
> P20rel
[1] 0.1286852
```

Number of days (/1560) that x number of vehicles will be available



```
> #number of days having 21 vehicles available
> avail_21 <- P21rel*1560
> avail_21
[1] 1344.312
> #number of days having 20 vehicles available
> avail_20 <- P20rel*1560
> avail_20
[1] 200.749
```

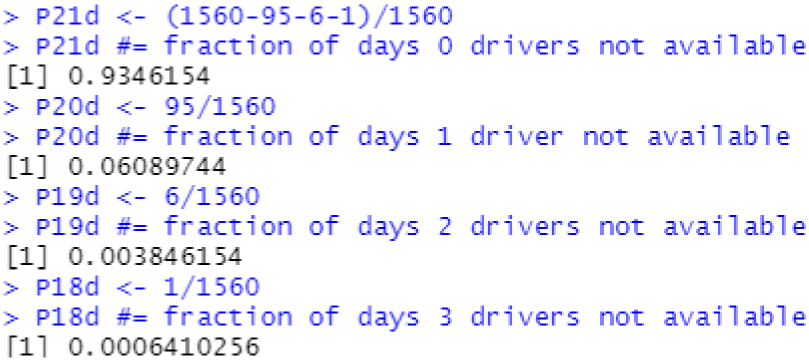
Expected number of days per year having reliable delivery (vehicles)



```
> #expected number of days having reliable delivery(vehicles) in 1 year
> daysperyear_reliable <- ((avail_21+avail_20)/1560)*365
> daysperyear_reliable
[1] 361.5046
```

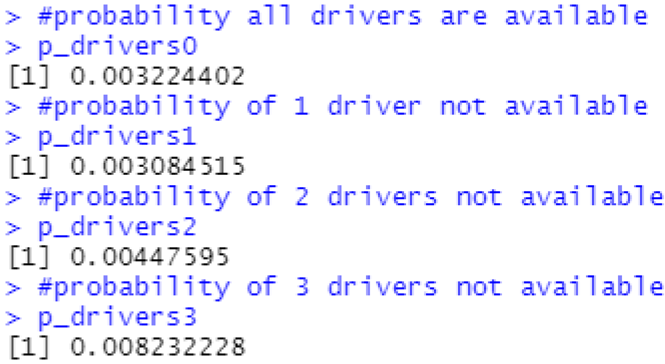
6.3.2. Driver reliability

Fraction of days x number of drivers not available



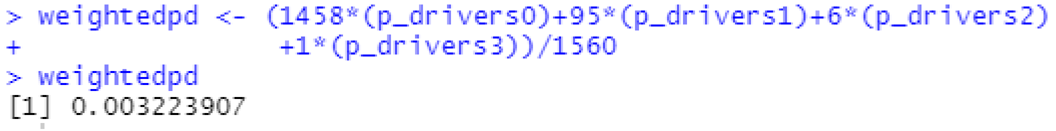
```
> P21d <- (1560-95-6-1)/1560
> P21d #= fraction of days 0 drivers not available
[1] 0.9346154
> P20d <- 95/1560
> P20d #= fraction of days 1 driver not available
[1] 0.06089744
> P19d <- 6/1560
> P19d #= fraction of days 2 drivers not available
[1] 0.003846154
> P18d <- 1/1560
> P18d #= fraction of days 3 drivers not available
[1] 0.0006410256
```

Probability of success of x number of drivers not available



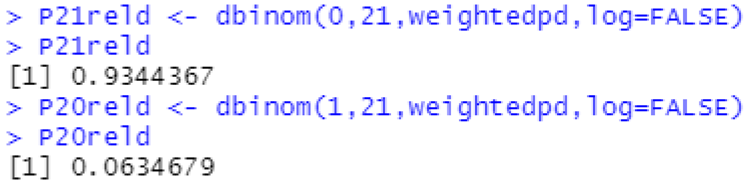
```
> #probability all drivers are available
> p_drivers0
[1] 0.003224402
> #probability of 1 driver not available
> p_drivers1
[1] 0.003084515
> #probability of 2 drivers not available
> p_drivers2
[1] 0.00447595
> #probability of 3 drivers not available
> p_drivers3
[1] 0.008232228
```

Weighted probability of success



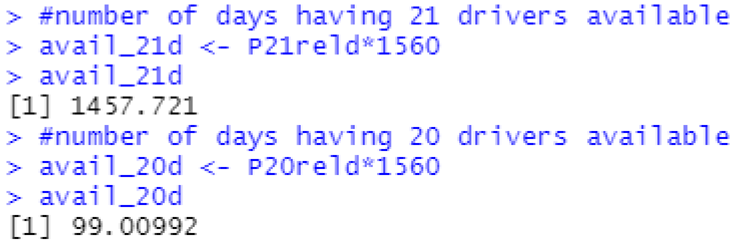
```
> weightedpd <- (1458*(p_drivers0)+95*(p_drivers1)+6*(p_drivers2)
+               +1*(p_drivers3))/1560
> weightedpd
[1] 0.003223907
```

Probability that x number of drivers will be available



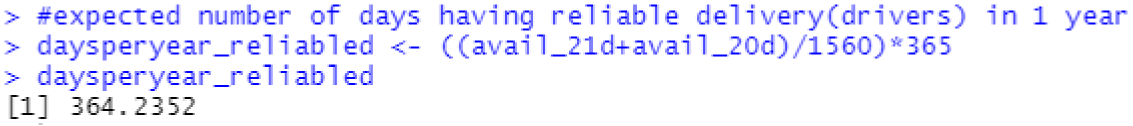
```
> P21reld <- dbinom(0,21,weightedpd,log=FALSE)
> P21reld
[1] 0.9344367
> P20reld <- dbinom(1,21,weightedpd,log=FALSE)
> P20reld
[1] 0.0634679
```

Number of days (/1560) that x number of drivers will be available



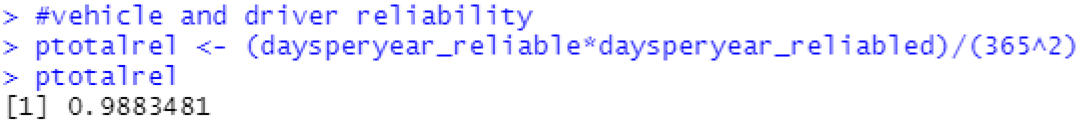
```
> #number of days having 21 drivers available
> avail_21d <- P21reld*1560
> avail_21d
[1] 1457.721
> #number of days having 20 drivers available
> avail_20d <- P20reld*1560
> avail_20d
[1] 99.00992
```

Expected number of days per year having reliable delivery (drivers)



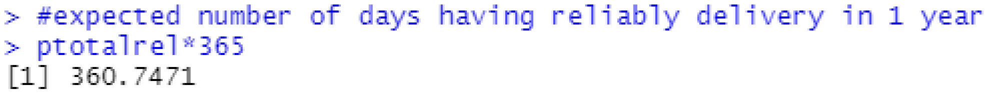
```
> #expected number of days having reliable delivery(drivers) in 1 year
> daysperyear_reliabled <- ((avail_21d+avail_20d)/1560)*365
> daysperyear_reliabled
[1] 364.2352
```

6.3.3. Vehicle and driver reliability



```
> #vehicle and driver reliability
> ptotalrel <- (daysperyear_reliable*daysperyear_reliabled)/(365^2)
> ptotalrel
[1] 0.9883481
```

Expected number of days per year having reliable delivery (vehicles and drivers)



```
> #expected number of days having reliably delivery in 1 year
> ptotalrel*365
[1] 360.7471
```

It is estimated that one can expect reliable delivery 360.75 days per year consisting of 365 days. Delivery reliability depends on the number of vehicles and drivers used and their respective reliability. Thus, if the company seeks to increase the delivery reliability per year, the number of days x number of vehicles available per year should be adjusted to obtain larger weighted probabilities of success. The binomial distribution is used to calculate respective reliabilities.

Conclusion

Insight into the business's sales is provided in the report. The raw data set contained several incorrect entries that were removed. It is clear from the visualised data that some variables are dependent on each other. The SPC charts confirm that some processes are in control while others, such as the delivery time process for household items, gifts, and luxury items, need to be investigated. To generate the best profit possible, the delivery time process is instead centered on seventeen hours. The MANOVA tests yield very low p-values which confirm the observation that certain variables do depend on others. The overall vehicle reliability is quite high as out of the 365 days in a year one can expect there to be approximately 361 days of reliable delivery.

References

1. Abraham, B., Steiner, S. and Mackay, J., (n.d.). *Understanding Process Capability Indices*. [online] Available at: https://sas.uwaterloo.ca/~shsteine/papers/cap.pdf.

2. Fogliatto, F.S. and Peres, F.A.P., 2018. Variable selection methods in multivariate statistical process control: A systematic literature review. *Computers & Industrial Engineering*, 115, pp. 603–619.

3. Furche, T., Gottlob, G., Libkin, L., Orsi, G. and Paton, N., 2016, November. Data wrangling for big data: Challenges and opportunities. In *Advances in Database Technology—EDBT 2016: Proceedings of the 19th International Conference on Extending Database Technology*, pp. 473-478.

4. Mourougan, S. and Sethuraman, Dr.K., 2017. Hypothesis Development and Testing. *IOSR Journal of Business and Management*, 19(5), pp. 34–40.

5. Weinfurt, K*.,* 2000*.* Repeated Measures Analysis ANOVA, MANOVA, and HLM. In L. Grimm, & R. Yarnold (Eds.), *Reading and Understanding MORE Multivariate Statistics,* pp. 317-361*. Washington DC American Psychological Association. - References - Scientific Research Publishing*. [online] Available at: https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1541234.

6. Yellapu, V., 2018. *Descriptive Statistics*, *International Journal of Academic Medicine.* [online] Available at: https://www.researchgate.net/publication/327496870_Descriptive_statistics#:~:text=Descriptive%20statistics%20are%20used%20to,before%20making%20inferential%20statistical%20comparisons.