This report covers the statistical analysis and analysis of control of processes to optimize processes of a company using their provided sales data. The data is sorted and then analysed with mathematically and visually.

ESCA PROJECT REPORT

QUALITY ASSURANCE 344

Titoti, LI, Mnr [22853480@sun.ac.za]

Contents

Introduction

The aim of this report was to statistically analyse the client data of an online business. The data is vast and spans from the start of 2021 to the end of 2029. The data focuses on the businesses sales with each sale entry including the ID number, age of customer, the cost of the product and the date (year, month and day) that the product was sold. The delivery time and reason of purchase and the class of the product was included as well.

The data was then sorted and all invalid data entries were separated, in order for the analysis of the data to be accurate. The data was

The analysis of the data was broken up into six parts.

Part 1: Data Wrangling

The data analysis began with the sorting the data and removing all invalid data entries. This was done by creating two separate data files with the valid and invalid data, this is done so to ensure work is only done with the relevant data.

```
> #check incomplete cases
> colSums(is.na(Sales))
          X          ID         AGE       Class       Price
          0           0           0           0          17
       Year       Month         Day Delivery.time  Why.Bought
          0           0           0           0           0
>
```

The figure above shows the number of invalid data, this entry points are removed as previously stated.

Part 2: Descriptive Statistics

After all invalid entries have been removed the data is suitable for analysis. The total number of valid sales entries are 179 978 and they start from the beginning of 2021 till the end of 2029.

The figure below shows the summary of the valid data.

```
> summary(Sales.good)
       X                ID             AGE            Class
 Min.   :     1   Min.   :11126   Min.   : 18.00   Length:179978
 1st Qu.: 45004   1st Qu.:32700   1st Qu.: 38.00   Class :character
 Median : 90005   Median :55081   Median : 53.00   Mode  :character
 Mean   : 90003   Mean   :55235   Mean   : 54.57
 3rd Qu.:135000   3rd Qu.:77637   3rd Qu.: 70.00
 Max.   :180000   Max.   :99992   Max.   :108.00
     Price                Year          Month            Day
 Min.   :    35.65   Min.   :2021   Min.   : 1.000   Min.   : 1.00
 1st Qu.:   482.31   1st Qu.:2022   1st Qu.: 4.000   1st Qu.: 8.00
 Median :  2259.63   Median :2025   Median : 7.000   Median :16.00
 Mean   : 12294.10   Mean   :2025   Mean   : 6.521   Mean   :15.54
 3rd Qu.: 15270.97   3rd Qu.:2027   3rd Qu.:10.000   3rd Qu.:23.00
 Max.   :116618.97   Max.   :2029   Max.   :12.000   Max.   :30.00
 Delivery.time   Why.Bought
 Min.   : 0.5   Length:179978
 1st Qu.: 3.0   Class :character
 Median :10.0   Mode  :character
 Mean   :14.5
 3rd Qu.:18.5
 Max.   :75.0
```

This summary shows that the company has a wide range of products, with a large range of price, delivery time and class.

To measure how a process performs with the customers requirements for criteria, the process capability indices are used. The indices include the capability of the process (Cp), upper (Cpu) and lower (Cpl) capability and the adjusted capability index (Cpk). These values were calculated to be as follows:

Cp = 1.142207

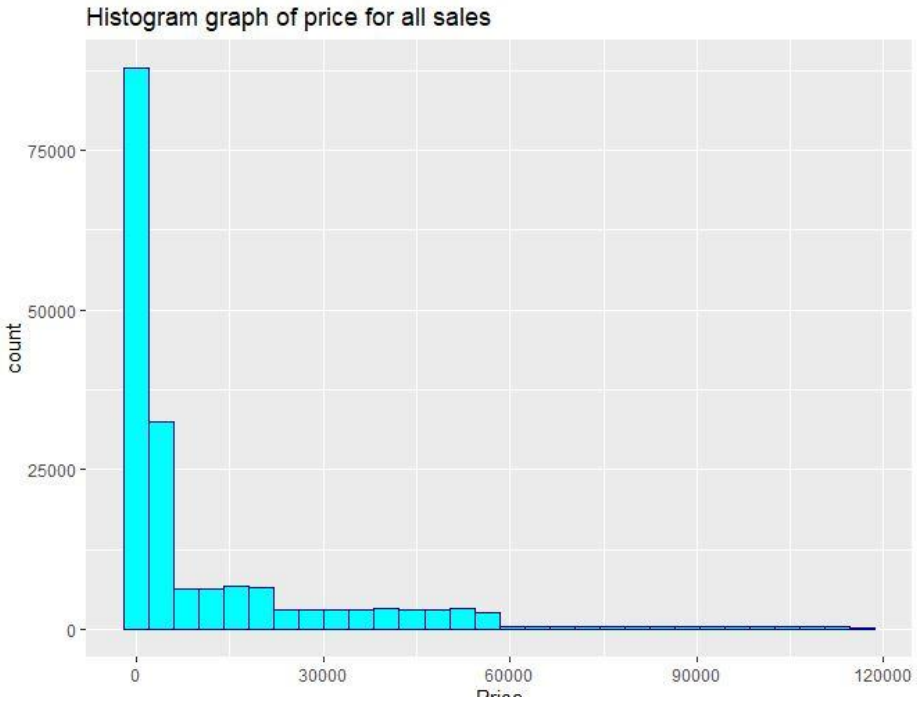
CpU = 0.3796933

CpL = 1.90472

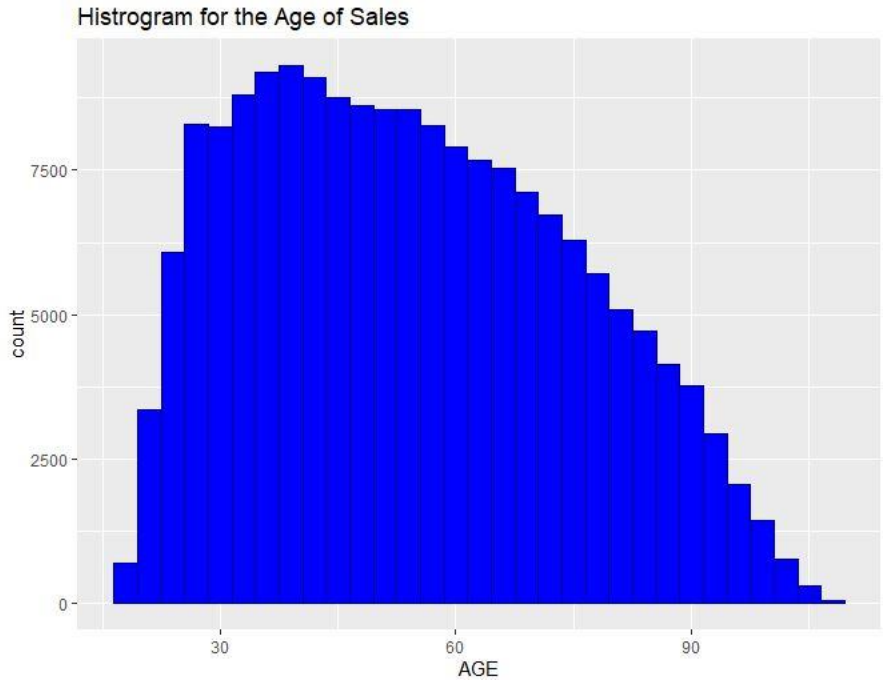
Cpk = 0.3796933

There are two different values for the Cp and the Cpk. Given the size of the discrepancy between the Cpu and Cpl, it follows that the average as a whole is not centered. When the Cp value exceeds 1, the process is capable of satisfying the specifications; when it falls below 1, the process is only partially capable of achieving the specifications. As a result, the delivery process falls short of completely fulfilling the requirements.

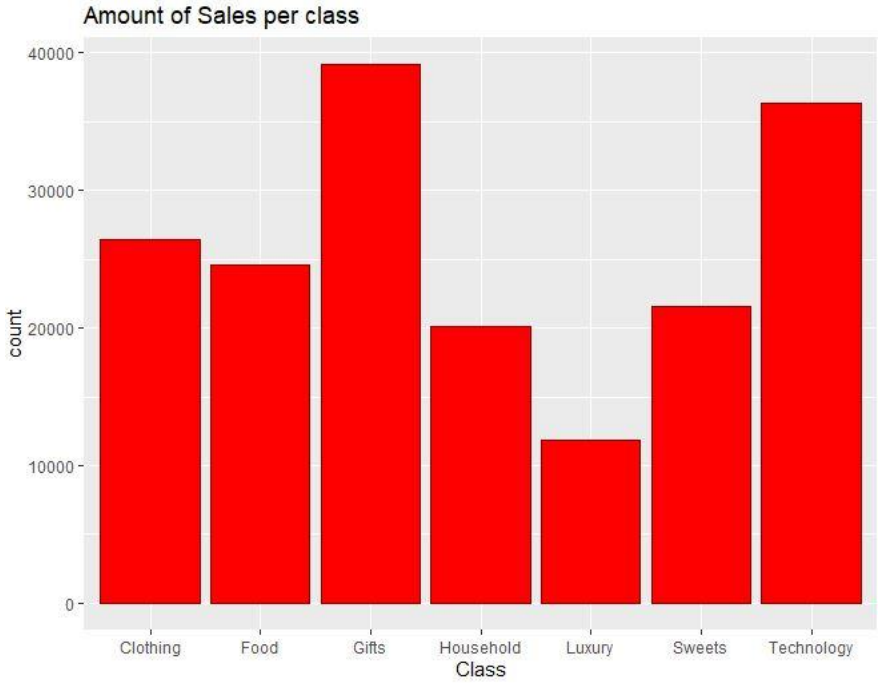
Upon analysis the data can be visually represented through different graphs and diagrams. A handful of graphs were made to help us understand the data.



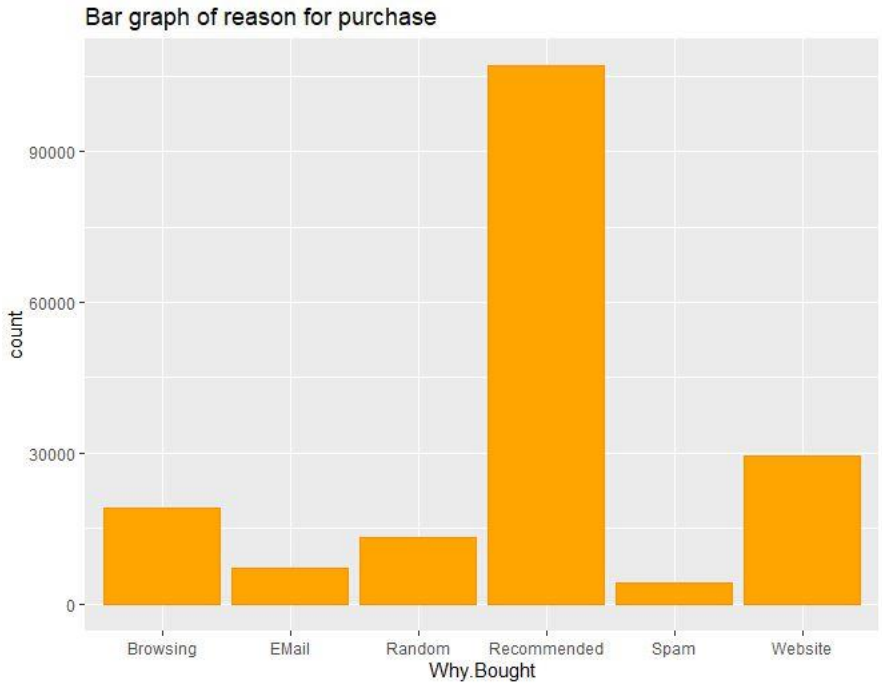
The graph above shows the large range of costs for the products being sold.



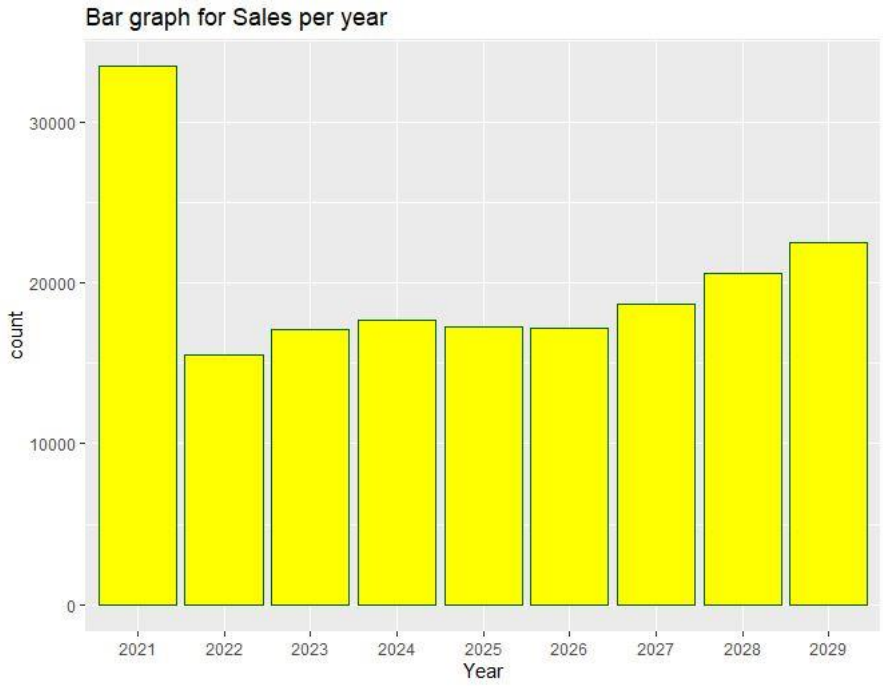
The histogram graph above shows the age range of customers buying products from the company. It can be seen that majority of the customers are adults between ages 25 and 65.



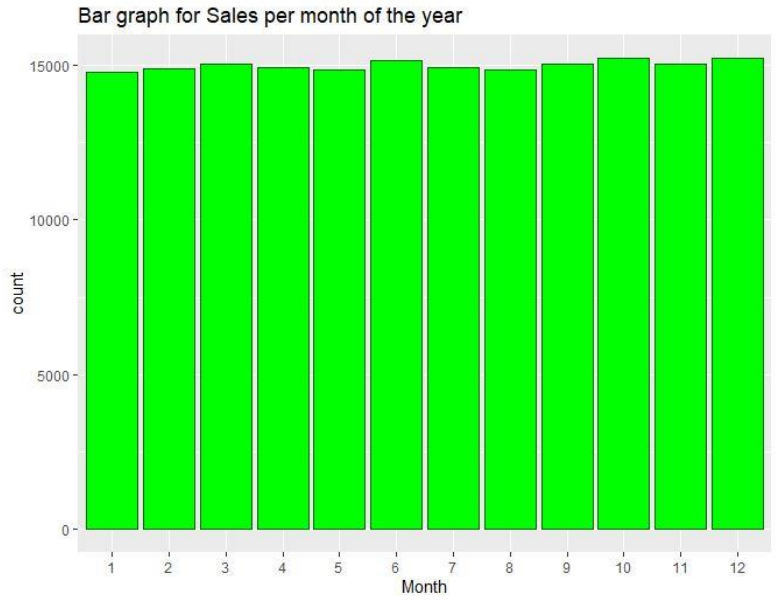
In the graph above it is evident that the gift class is the most popular of products sold, followed by technology. Luxury is the class with the lowest amount of sales and is likely due to the high expense of products in this class.



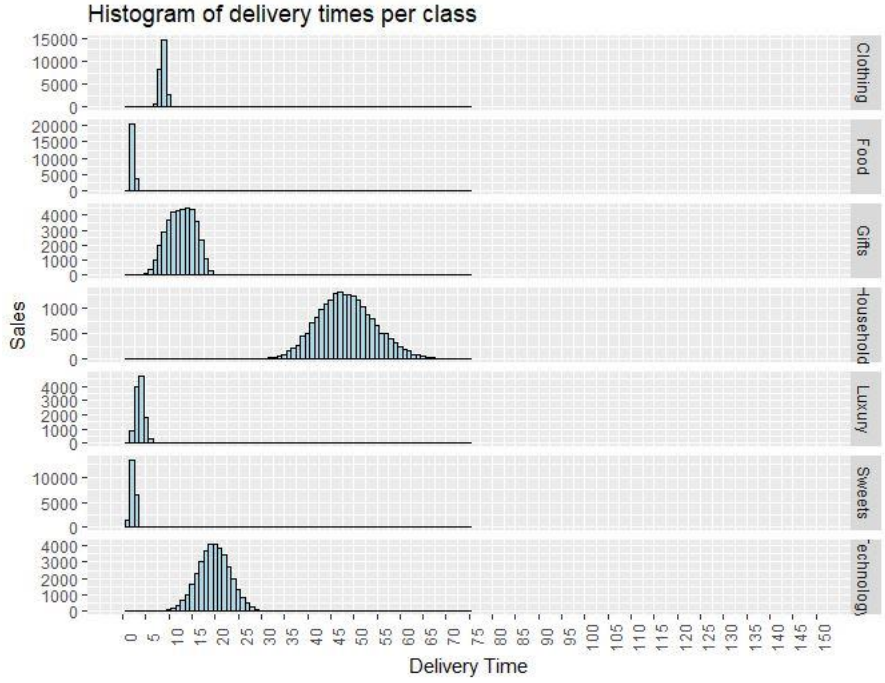
The graph above shows the number of sales for the different customer reasons for purchase. According to the data, recommendations are the most common reason customers choose the company's products. This may be a sign that the business values its relationships with its clients. If so many people are praising the business, then they must have had a positive experience.



The sales data cannot be used to infer the cause of the much higher number of sales in 2021 compared to later years, as shown in the graph above. Sales show a general increase tendency from 2022 to 2029; this is a sign of a company with good prospectives.



The graph above shows the sales for each month during the relevant time and demonstrates that the business does not experience any kind of seasonality and instead has a constant volume of sales all year round. This should result in a consistent income stream that makes it easier for them to create a budget.



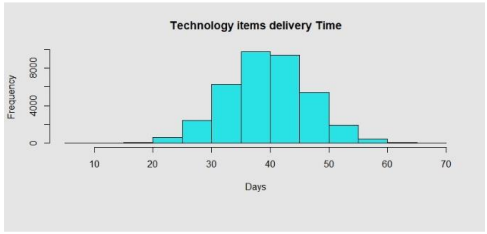
The delivery times for each lesson are shown visually in the graph above. This histogram shows that the household items class's delivery times are significantly longer than those of the other classes'. The majority of purchases take place around every 60 days, although the average delivery period for household goods is 100 days. This emphasizes the need for careful consideration in the household item distribution process.

The graph also demonstrates that the delivery times follow a normal distribution, which is to be anticipated. The lack of notable outliers also suggests that the processes are reliable.

Part 3: Statistical Process Control

As variability in a process can have numerous detrimental implications on production and the company, statistical process control is a tool used to measure the control and consistency of a process.

The delivery timing information for the various classes will be the main focus of the statistical process control. The distribution of the delivery times for the corresponding classes is more clearly displayed in the graphs below.



3.1 Initialization of s charts and x-bar charts

The delivery time data must be divided into its appropriate classifications and organized chronologically in order to create x-bar and s-charts. The initial 30 samples from a total of 16 were examined for the control charts. The investigation produced the following control charts.

S-charts, also referred to as standard deviation control charts, show the standard deviation of a variable over time from various subgroups or samples.

X-bar graphs depict the average difference in a variable (in this case, delivery time) across time from various subgroups. This is used to determine whether the process' average stays inside a set of desired boundary values.

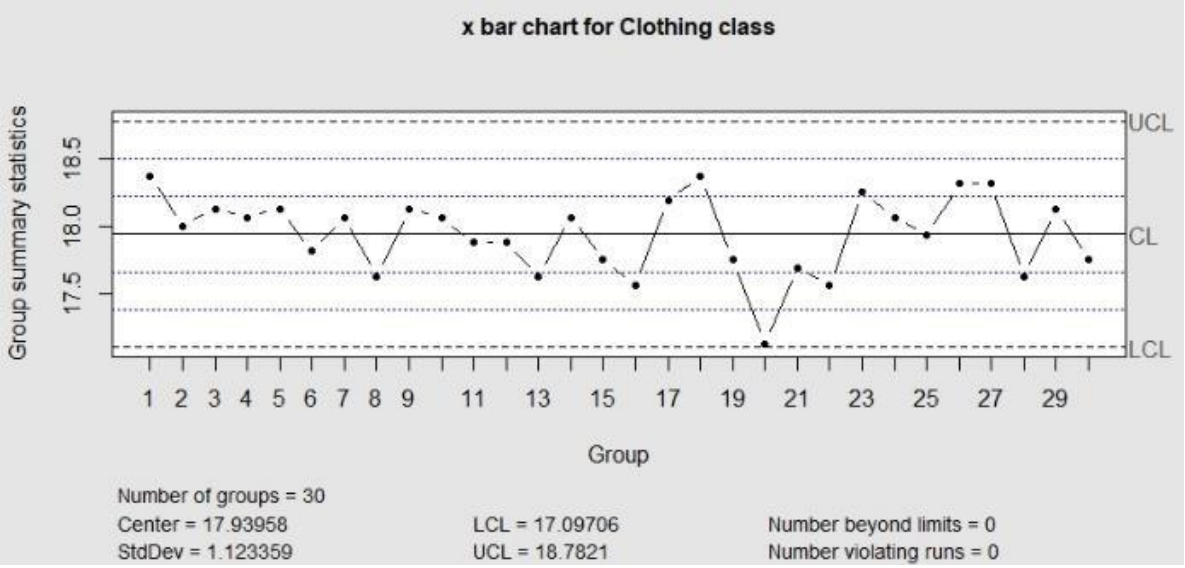
The charts are organized below according to their categories; we begin by analyzing the s-chart because it shows the process variation that affects how the xbar chart is studied.

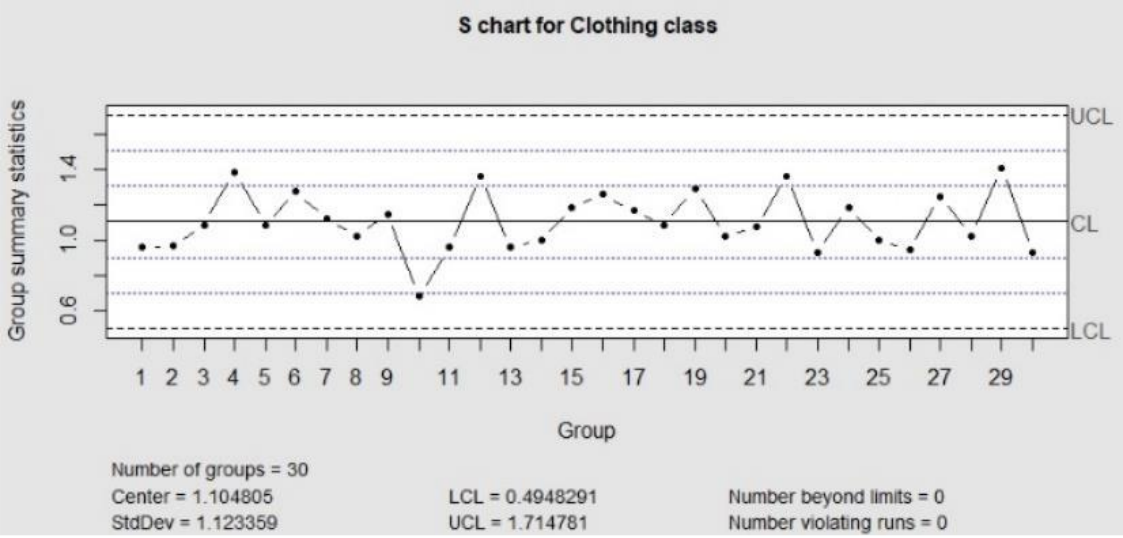
A word on the control chart comments: If there isn't a comment on a particular control chart, the process is in control and everything is going as planned. This means that the x-bar chart can be recognized as genuine if there are no comments on the s-chart.

Class	Lower Limit	Upper Limit	Values between limits	Number of consecutive cases	Ending sample number
Clothing	1.206468	1.348796	297	4	238
Food	0.6020975	0.6020975	269	4	119
Sweets	1.1703	1.308361	196	3	159
Luxury	2.073728	2.318367	29	4	4
Gifts	3.178991	3.554018	474	6	314
Technology	7.11626	7.955769	481	8	284
Household	10.17114	11.37104	292	6	191

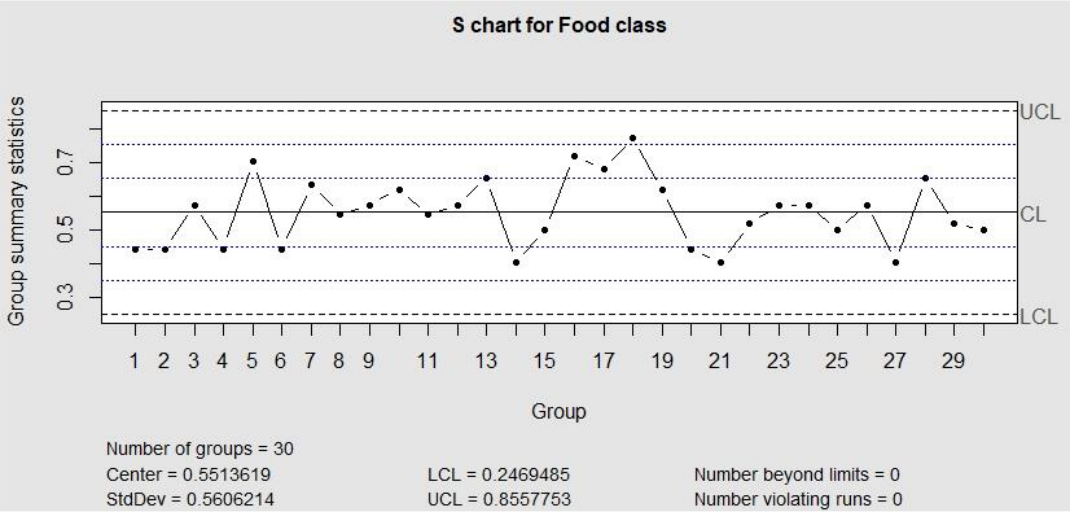
Class	UCL	U2 Sigma	U1 Sigma	CL	L1 Sigma	L2 Sigma	LCL
Clothing	18.7821	18.34623	18.14291	17.93958	17.73626	17.53293	17.0971
Food	5.401716	5.184192	5.082721	4.98125	4.879779	4.778308	4.56078
Gifts	18.98042	17.83192	17.29617	16.76042	16.22466	15.68891	14.5404
Household	100.2237	96.54911	94.83497	93.12083	91.4067	89.69256	86.018
Luxury	10.90857	10.15939	9.809901	9.460417	9.110932	8.761448	8.01226
Sweets	5.767262	5.34446	5.14723	4.95	4.75277	4.55554	4.13274
Technology	45.60704	43.0361	41.8368	40.6375	39.4382	38.2389	35.668

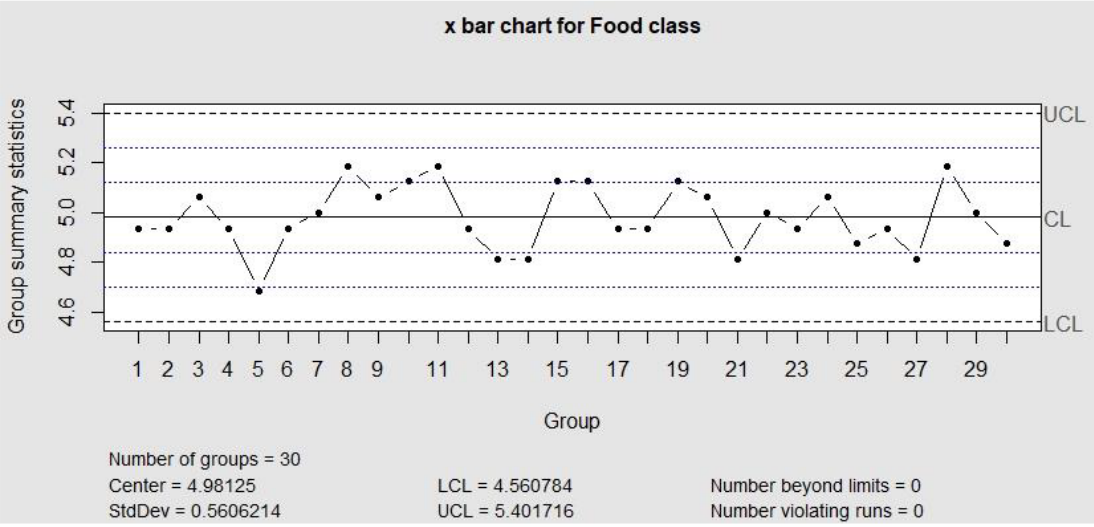
Clothing class charts



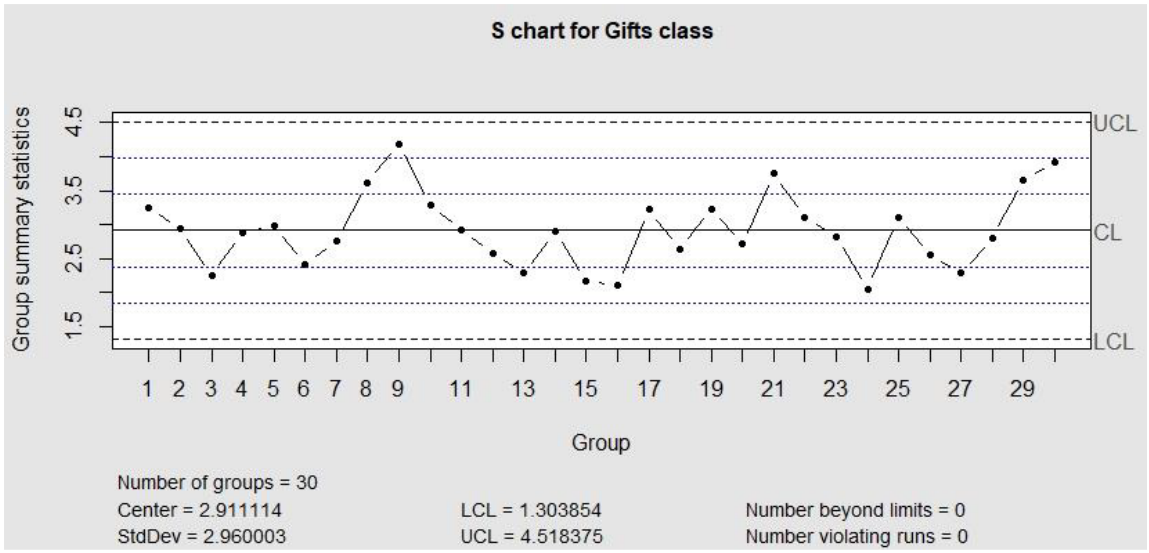


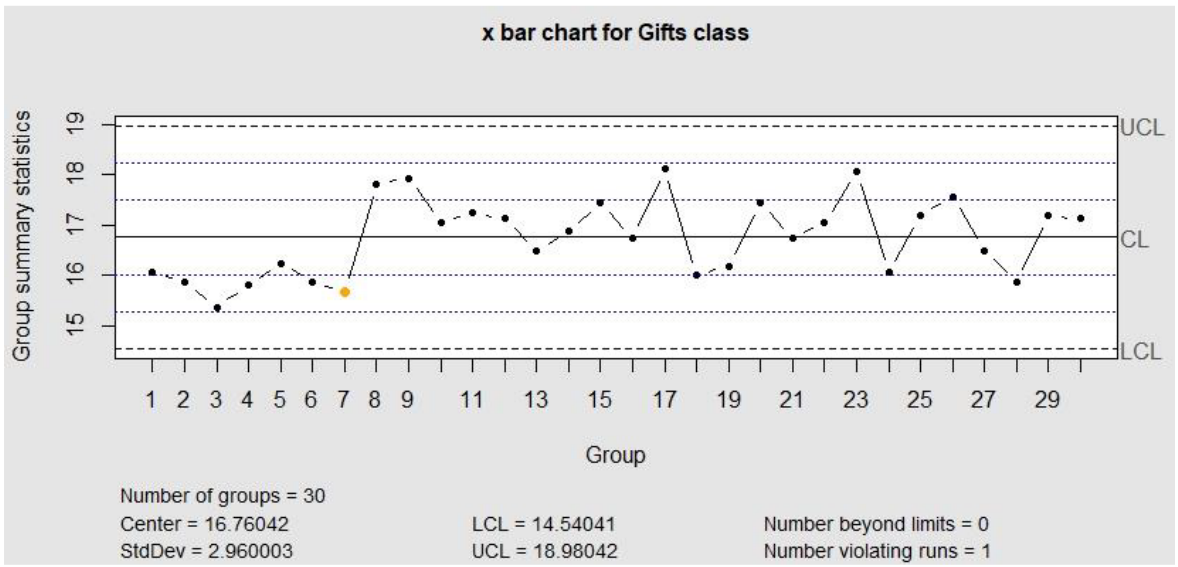
Food class charts



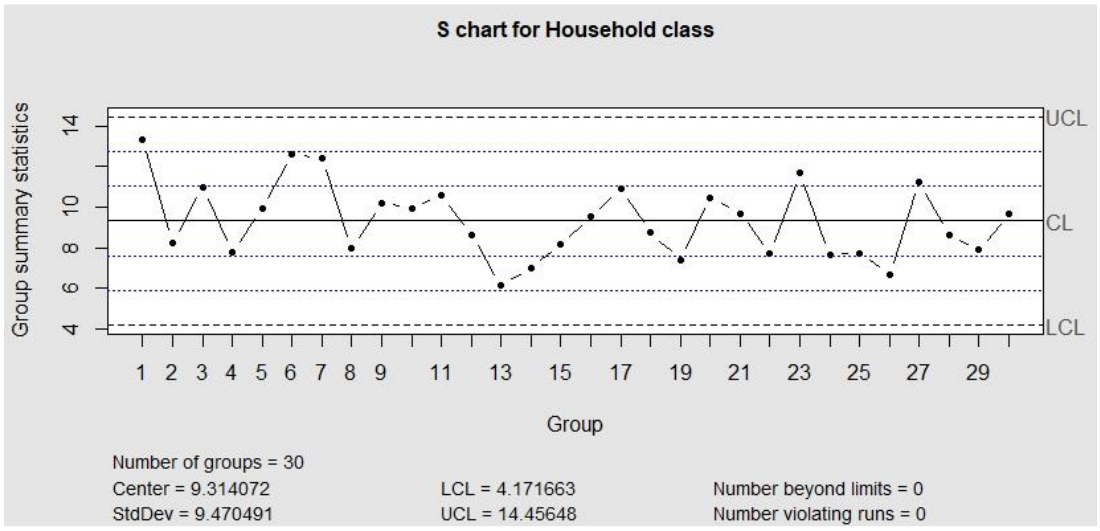


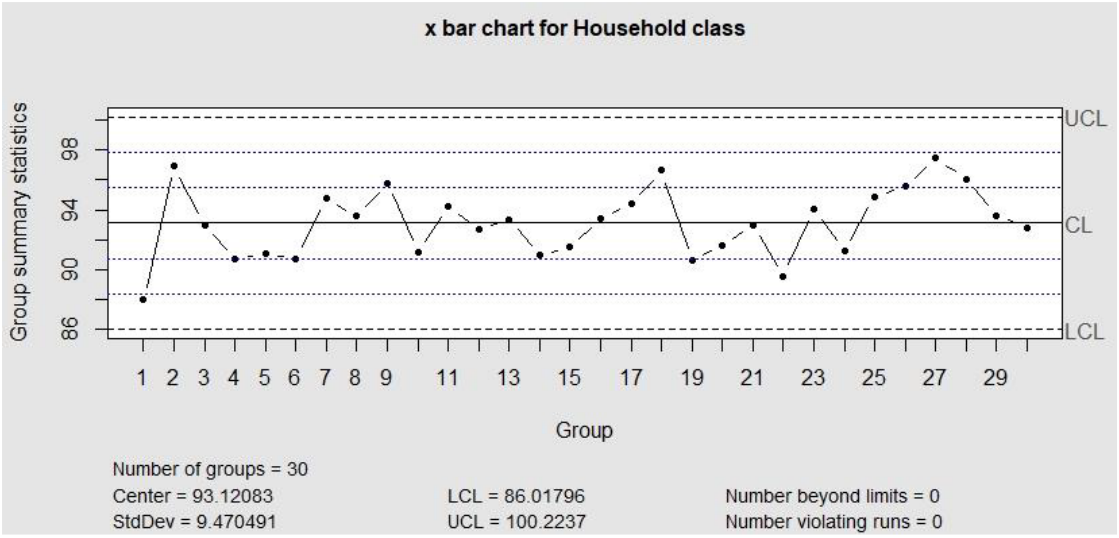
Gifts class charts



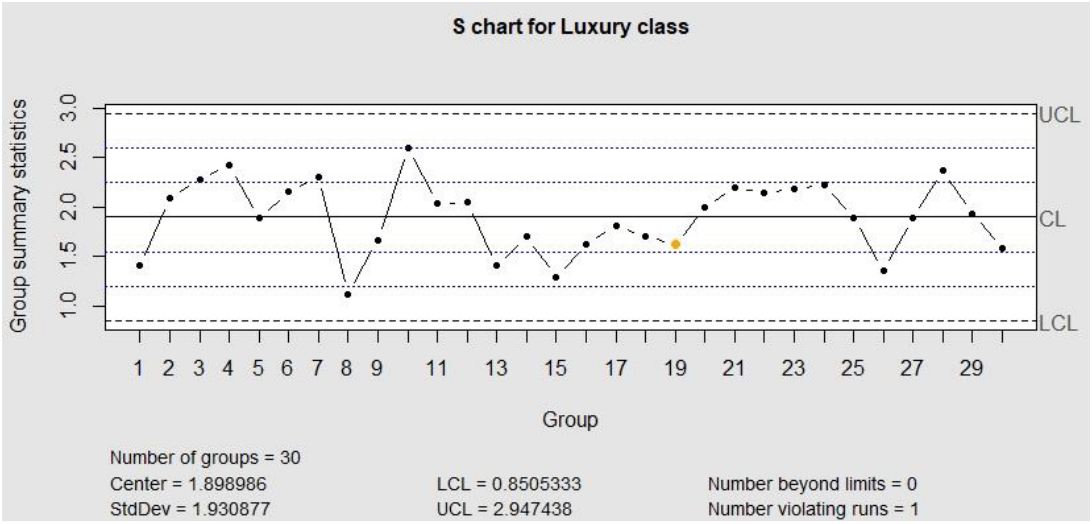


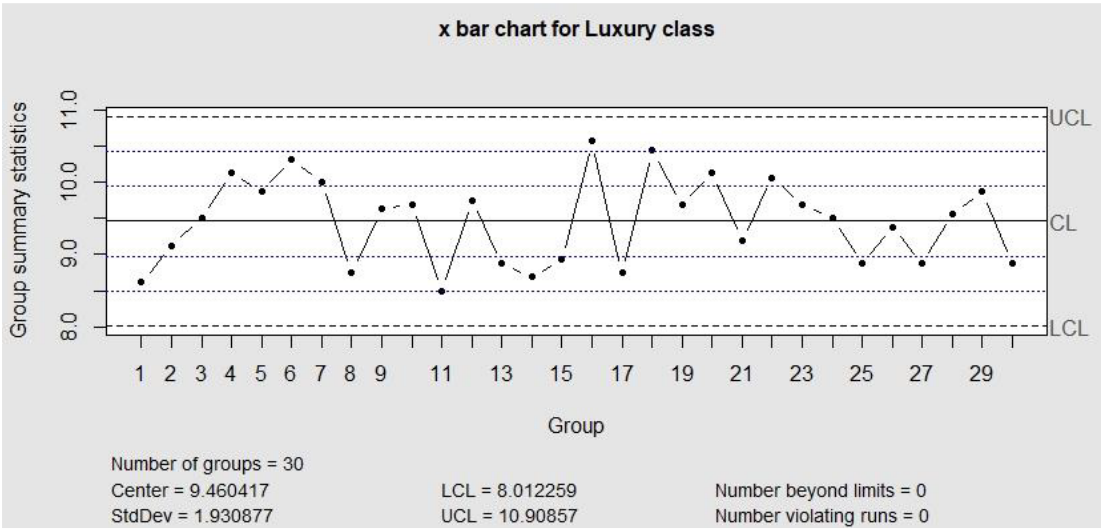
Household class charts



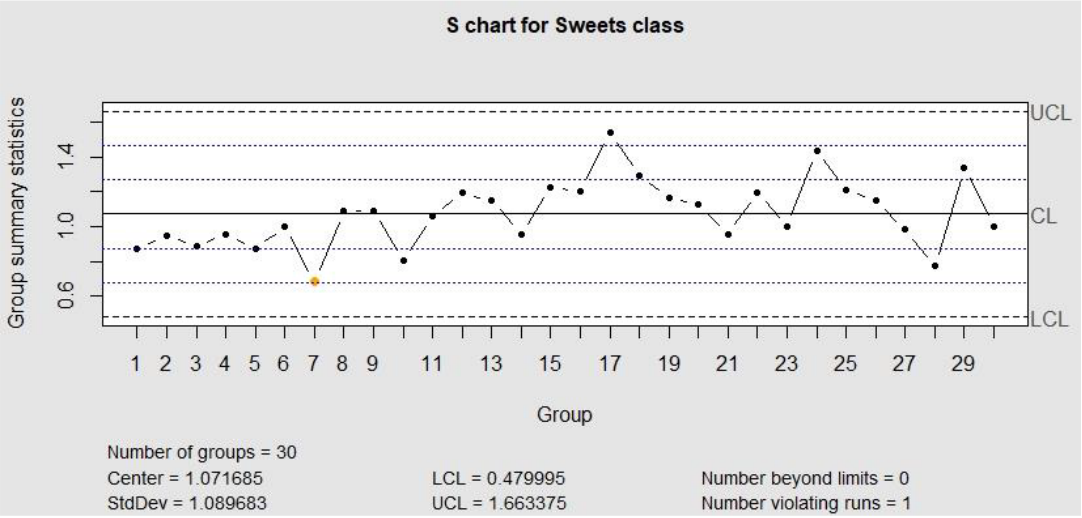


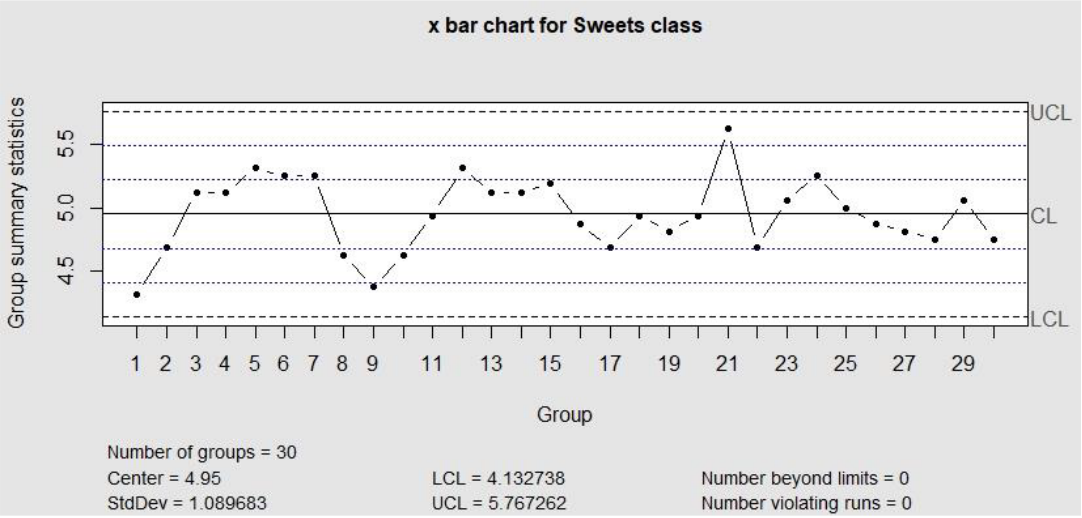
Luxury class charts



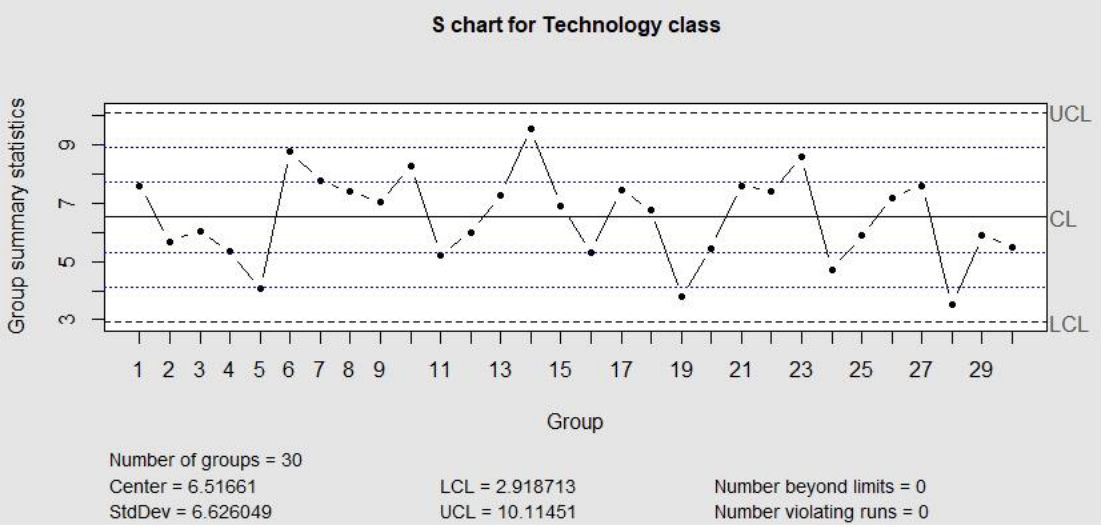


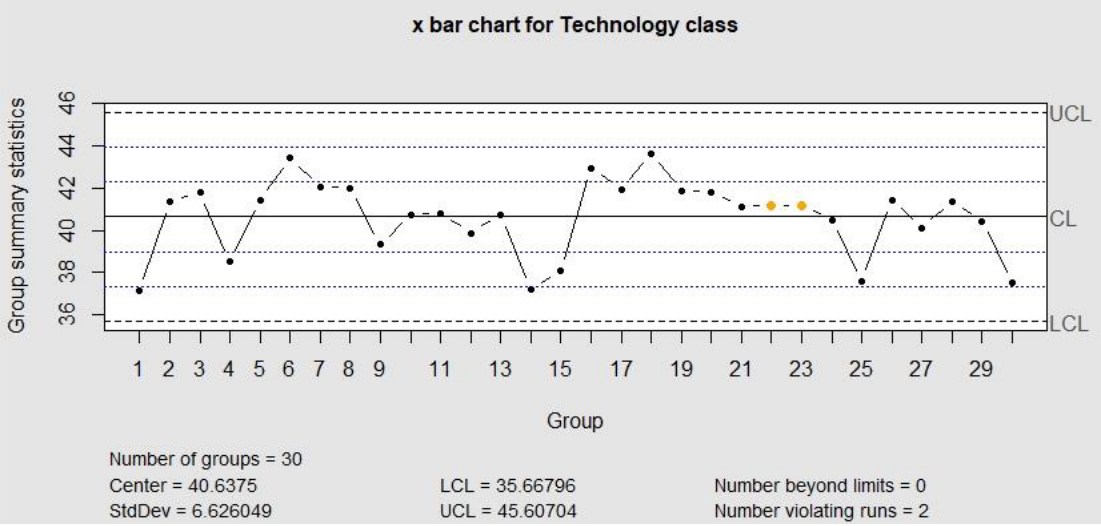
Sweets class charts





Technology class charts





Part 4: Optimising the delivery processes

The focus of this part of the report is using the information found in part 3 and discussing how to enhance the delivery process.

4.1 Samples that give indications of being out of control

To identify samples that are out of control, two different criteria will be applied, namely:

A. When sample means fall outside of the control range.
B. The number of successive sample standard deviations within the sigma-control ranges (-0.4 to +0.5).

4.1.1 Samples that were outside of the control limits (condition A)

All samples that were outside of the control limits (as determined by the control limits in part 3) were recorded and saved in a separate file for each class of objects.

4.1.2 Sample standard deviations that occur most frequently (condition B)

The standard deviation for the samples of each class required to be documented in order to use this condition. The most consecutive values between -0.4 and +0.5 sigma-control limits were discovered after the values had been evaluated.

4.2 Type 1 (Manufacturer's) error

There is always a chance of making a mistake due to the constant uncertainties in statistics. False positive conclusions, also known as type one error, occur while evaluating a hypothesis. In this situation, the null hypothesis (H0) is "the process is in control and centered on the centerline derived using the first 30 data," therefore if this hypothesis were rejected even though it were true, a type 1 error would result. The significance level () is a hypothesis test's type 1 mistake.

4.3 Delivery time optimization

The delivery timeframes for the technology item will be optimized in this area. Currently, an item's late delivery costs R321 per day. The reduction of this cost is the aim of process optimization, but slashing delivery times also has a cost. A one-day mean delivery time reduction costs R1.5 per item. An optimized process will look for the lowest overall cost, which is the sum of the costs associated with reducing the average number of days late and the costs associated with late items. To ensure that there are no negative delivery times, the mean can only be reduced to the number of times the minimum delivery time.

The initial item-late cost is R 8 874 045, and the revised total cost is R 8 356 929 after the mean delivery time has been lowered from 40 to 34 days. Profit rises by R 517 116 as a result of this.

4.4 Type 2 (Consumer's) error

The likelihood of failing to reject a false hypothesis is the type two error for the technological process. Not rejecting is the statistical equivalent of "accepting." A type two error is when the conclusion is drawn that there was an effect when there wasn't. For condition A, the type two error is 0.00385, or 0.3%. The likelihood that the error will occur is quite low.

Part 5: DOE and MANOVA

Multiple continuous dependent variables are combined into a weighted linear combination, or composite variable, as part of the MANOVA study. The MANOVA compares the newly created combination's differences from one another after that. In essence, MANOVA examines whether the independent grouping variable contributes to the statistical significance of the variation in the dependent variable.

the summary of the MANOVA that compares the averages of each class's delivery times. At a significance level of 0.09588, or 9.58%, the p-values for the delivery times of the each class are statically important.

the MANOVA summary that contrasts the ages of each consumer who bought a product from a given class. At a significance threshold of 0.1012, or 10.2%, the p-values for the delivery times of the each class are statistically relevant.

Part 6: Reliability of the service and products

A refrigerator part for a company, Lafrigeradora, has a specification of 0.040 ± 0.35 cm. It costs $30 to scrap a part that does not meet the specifications. Determine the Taguchi loss function for this situation. $L(x)=k(x-N)$^2
Where k = 24489.8 and N = 0.035, results in the function

Conclusion

The business has a few areas to improve on to increase the business' income. The business is doing a lot of things right as many of the new costumers are through recommendations, thus showing high customer satisfaction. Improving on the quality of processes will ensure that this is maintained, and that customers stay loyal which will lead to the customer base potentially increasing exponentially. The delivery times for household deliveries should be seen to as over the years it has deteriorated, ensuring that the times stop increasing is an urgent priority. The delivery times for the gift class specifically has been increasing at a worrying rate and should be attended to with urgency as well as it will lead to a loss of sales from the time wasted. A large increase of almost half a million rands in profit can be made by reducing the average delivery times for technology class deliveries by 6 days. This was determined and calculated with the optimum reduction in average delivery times, with given costs. And this can be applied to each of the other classes.

References

Bhandari, P. (2021). *Type 1 and type 2 errors.* [Online]. Available: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/ [2021, 1 Nov]

Evans, J. and Lindsay, W. (2010). *Manageing for Quality and Performance Excellence*. 10th ed.x

MANOVA Test in R: Multivariate Analysis of Variance [Online]. Available: http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance [2021, 23 Oct]

MANOVA [Online]. Available: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/manova/ [2021, 21 Oct]

MANOVA Test in R: Multivariate Analysis of Variance. [Online]. Available: https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/anova/how-to/general-manova/interpret-the-results/key-results/ [2021, 7 Nov]

Scrucca, L. (2017). *A quick tour of qcc*. [Online]. Available: https://cran.r-project.org/web/packages/qcc/vignettes/qcc_a_quick_tour.html [2021, 12 Oct]

https://www.whatissixsigma.net/taguchi-loss-function/