A report on quality assurance for a business

by

Jake Carson Oosthuizen
23086297

Prepared for:

The Engineering Council of South Africa

21 October 2022

ABSTRACT

The following report portrays data analysis of an online business. It is required by the Engineering Council of South Africa that all industrial Engineering students at the University of Stellenbosch pass this project to enable graduation. The necessity of this project is to gain knowledge and skills of working and analysing a business's data.

TABLE OF CONTENTS

LIST OF FIGURES

NOMENCLATURE

Symbols	
%	Percentage
Acronyms	
ECSA	The Engineering Council of South Africa
LCL	Lowe Controlling Limit
UCL	Upper Controlling Limit
CL	Centre Line

1 INTRODUCTION

Data analysis of a business is essential to understanding and optimizing the business process. The business works from online orders and then transports products to its customers. The business sells 7 different types of products. Historical data of the business from 2021 to 2029 has been provided. All calculations were made using a programming language called RStudio.

This report starts with data wangling, where valid data and invalid data is separated. Thereafter, the valid data is organized and descriptive statistics on the valid data is provided. Following this, X&S charts for delivery time are calculated to establish controlling limits for each feature of the company. Type I and type II errors are explained and estimated. Minimizing delivery cost calculations were done. DOE and MANOVA test were done on the data for further analysis. Thereafter, reliability of service product calculations was done.

This report ends with a conclusion and list of references.

2 PART 1: DATA WANGLING

2.1 Valid data

Valid data contains all the data after missing values and negative numeric characters have been removed from the given dataset. The dataset contains 179 978 instances that are valid. The original dataset contained 180 000 instances, meaning, that 22 instances was not valid. In figure 1 below it can be observed that all the negative and missing values have been removed from the original dataset. The figure only displays the first 15 instances for convenience measures due to the large dataset.

	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

Figure 1: Valid dataset

2.2 Invalid data

Invalid data contains all the instances that are negative or missing characters. In the original data of 180 000 instances, only 22 instances qualified as invalid data. Due to so little instances of invalid data, measures to remove all invalid data was taken. The consequences of removing these instances are not very high since the percentage of missing values are only 0.0122%. Removing the instances will display more accurate results, rather than replacing missing values with the mean or mode of instances. In figure 2 below it can be observed that all the invalid instances either contain negative or missing values.

	Z	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	2	16320	44142	82	Household	-588.8	2023	10	2	48.0	EMail
3	3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
4	4	19540	65689	96	Sweets	-588.8	2028	4	7	3.0	Random
5	5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6	6	19998	68743	45	Household	-588.8	2024	7	16	45.5	Recommended
7	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	8	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9	9	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10	10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	11	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
12	12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
13	13	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	14	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
15	15	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
16	16	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
17	17	144443	37737	81	Food	-588.8	2022	12	10	2.5	Recommended
18	18	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
19	19	155554	36599	29	Luxury	-588.8	2026	4	14	3.5	Recommended
20	20	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
21	21	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
22	22	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Figure 2: Invalid dataset

2.3 Valid dataset compared to original dataset

The first instance of a missing value is found at instance 12 345. The following instances in figure 3 all contained missing values. It is evident from figure 3 below that the original dataset contained 17 missing values and we have observed from before that there was 22 invalid data instances, meaning that the original dataset contained 5 negative numeric characters.

```
> which(!complete.cases(data))
 [1]  12345  16321  19541  19999  23456  34567  45678  54321  56789  65432  76543  87654
[13]  98765 144444 155555 166666 177777
```

Figure 3: Instances containing negative values

In figure 4 below it can be observed that the instance of missing values were removed. Figure 4 contains the original dataset from instance 12344 to 112347. In instance 12345 in the price column, it can be distinguished that there is a missing value present (NA).

	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
	<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	12344	90260	34	Luxury	42892.	2025	8	4	4	Recommended
2	12345	18973	93	Gifts	NA	2026	6	11	15.5	Website
3	12346	92286	32	Technology	38167.	2028	7	6	19.5	Website
4	12347	89263	44	Clothing	892.	2021	7	2	8.5	Recommended

Figure 4: Not wangled data

After data wangling has taken place, it can be detected that in figure 5 the missing value have been removed. Figure 5 displays the same instances as figure 4 but the missing value have been removed.

	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
	<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	12344	90260	34	Luxury	42892.	2025	8	4	4	Recommended
2	12346	92286	32	Technology	38167.	2028	7	6	19.5	Website
3	12347	89263	44	Clothing	892.	2021	7	2	8.5	Recommended

Figure 5: Wangled data

3 PART 2: DESCRIPTIVE STATISTICS

3.1 Statistical table from the valid dataset

Figure 6 below portrays a summary containing the minimum, 1st Quartile, Median, Mean, 3rd Quartile and the maximum of the valid dataset for different classes. The table below is characterizing each feature and illustrates what is expected of each feature in the future. Figure 6 can also portray outliers and invalid data if they have not been removed yet, for example, the maximum age of a person who participated in the dataset is 108 years old, it is unusual for people to live to that age, but it is possible. Therefore, the figure is useful when new data is being interpreted.

	Minimum	1st Qaurtile	Median	Mean	3rd Quartile	max
Age	18.000	38.000	53.000	54.570	70.000	108.000
Price	35.650	482.310	2259.630	12294.100	15270.970	116618.970
Year	2021.000	2022.000	2025.000	2025.000	2027.000	2029.000
Month	1.000	4.000	7.000	6.521	10.000	12.000
Day	1.000	8.000	16.000	15.540	23.000	30.000
Delivery time	0.500	3.000	10.000	14.500	18.500	75.000

Figure 6: Statistical analysis table

3.2 Analysis of graphs

The graph below plots the Price each person paid for a product or service. It is evident that there is a mode of instances, and that the graph is skewed to the right. Due to the graph that is skewed to the right means that the data could contain outliers. Outliers in data cause predictions for future instance to be less accurate.

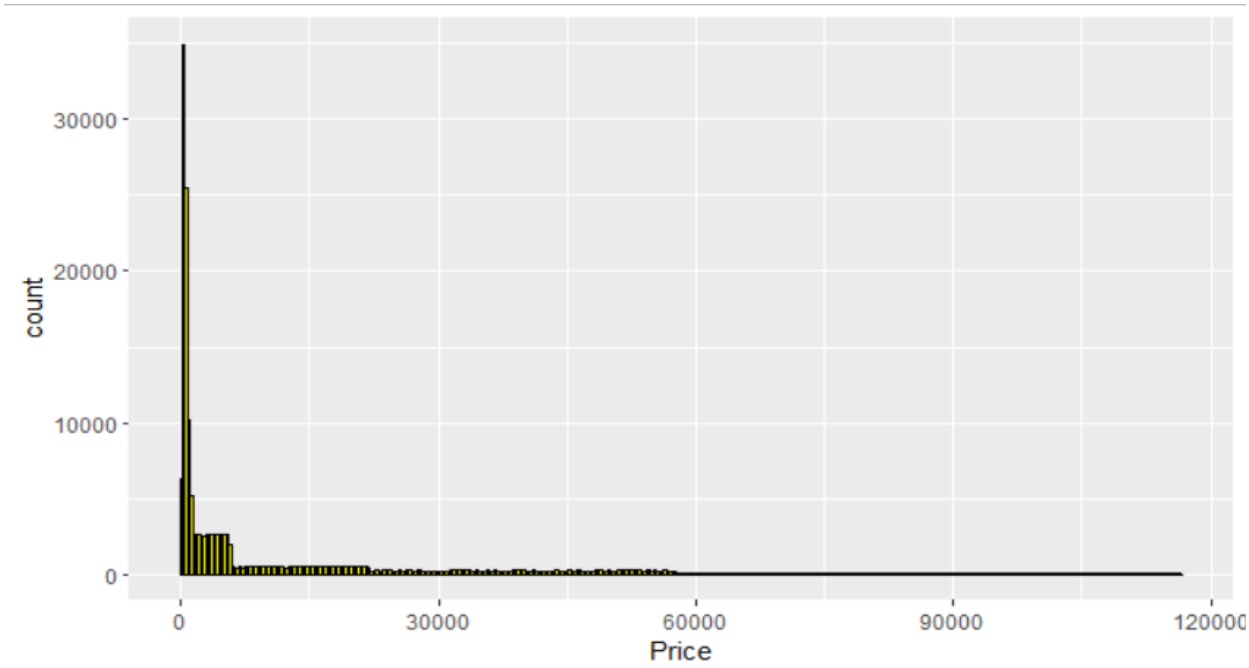

Figure 7: Price graph

Figure 8 below portrays the delivery time of a product to customers. It is evident from figure 8 that the most common delivery times in days are 2 days and 4 days. The graph is normally distributed between 35 and 65 days, with a mean of 50 days.

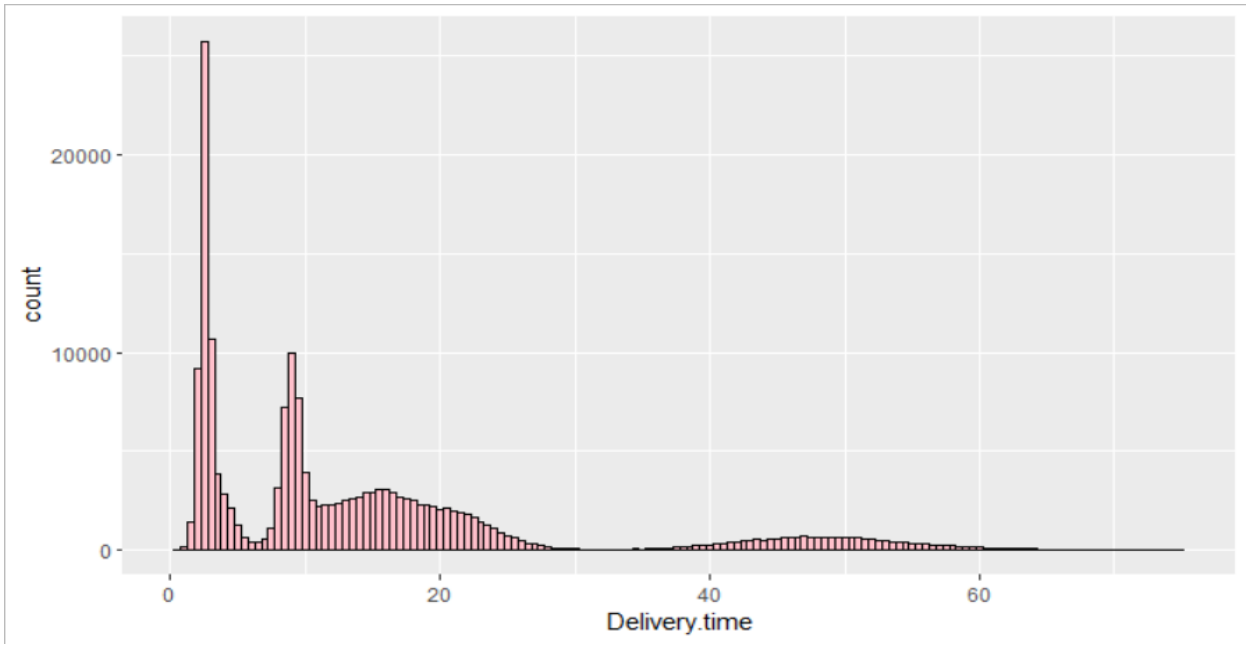

Figure 8: Delivery time in days

Figure 9 below displays amount of each instance relative to the year it happened in, it is clear from the graph below that it is uniformly distributed, meaning that there is no correlation between the year of the instance and a target feature.

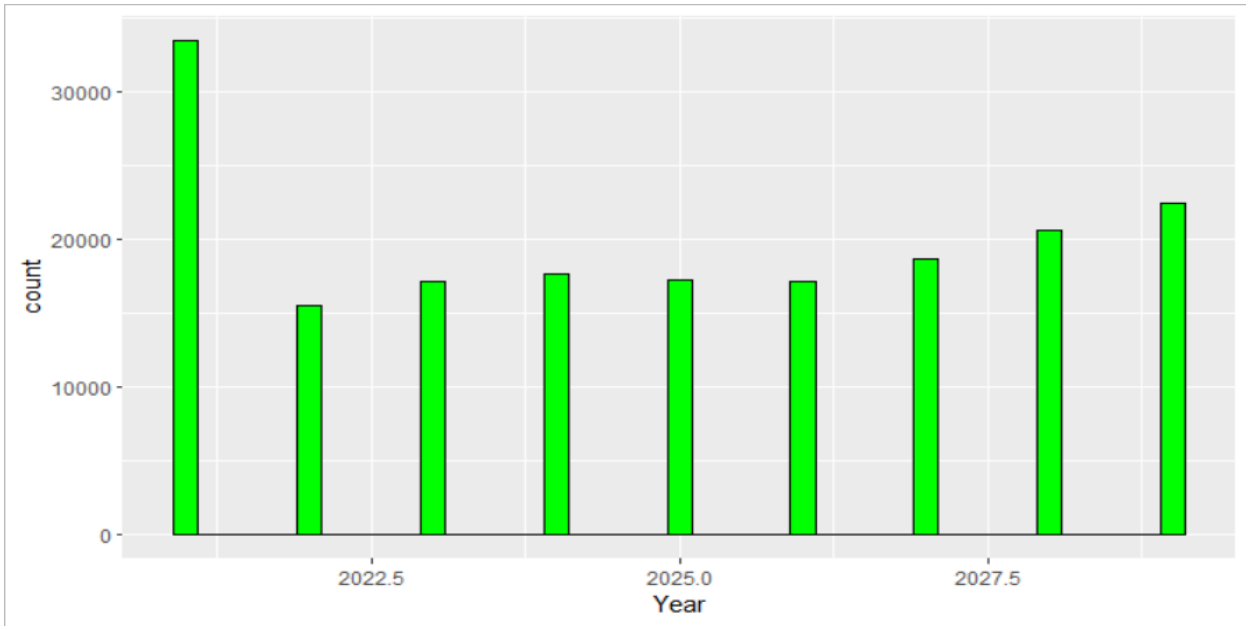

Figure 9: Uniformly distributed year graph

Figure 10 below displays the quantity of each instance relative to the day it happened in, it is clear from the graph below that it is uniformly distributed, meaning that there is no correlation between the day that the instance happened and the target feature.

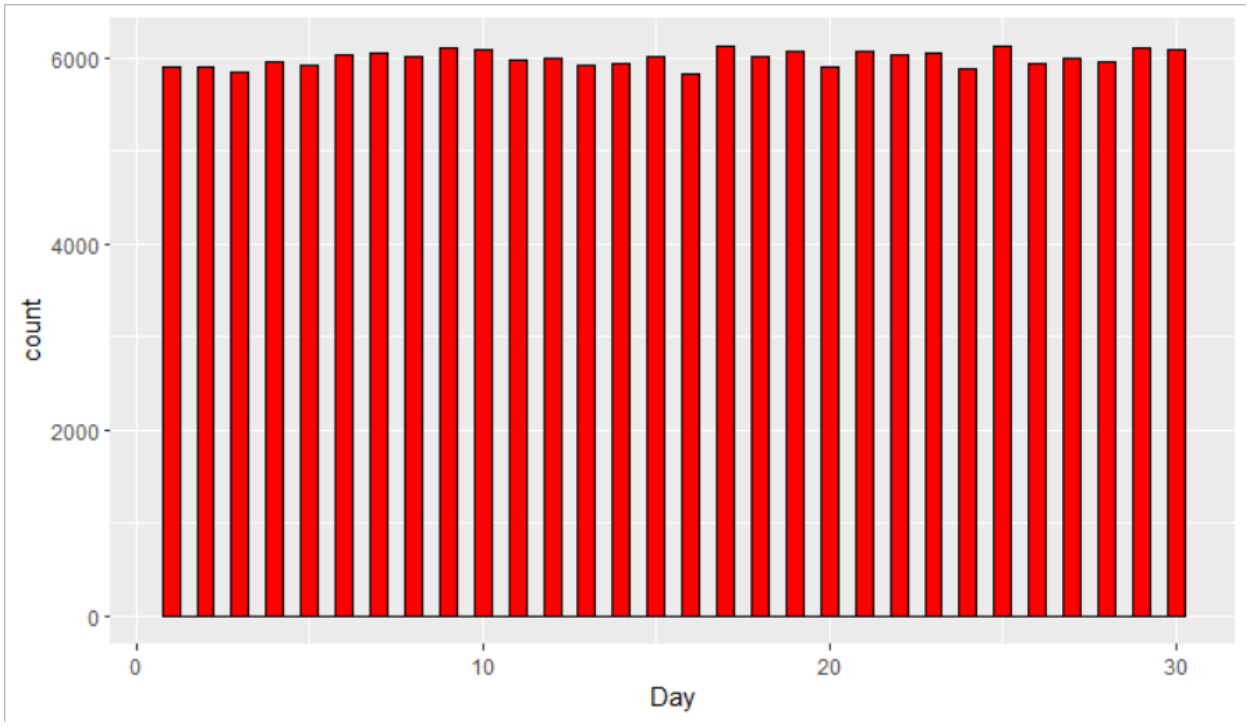

Figure 10: Uniformly distributed graph of the day that and instance happened

Figure 11 below displays the quantity of each instance relative to the month it happened in, it is clear from the graph below that it is uniformly distributed, meaning that there is no correlation between the month that the instance happened and the target feature. The graph does not portray any hidden patters or information in the data.

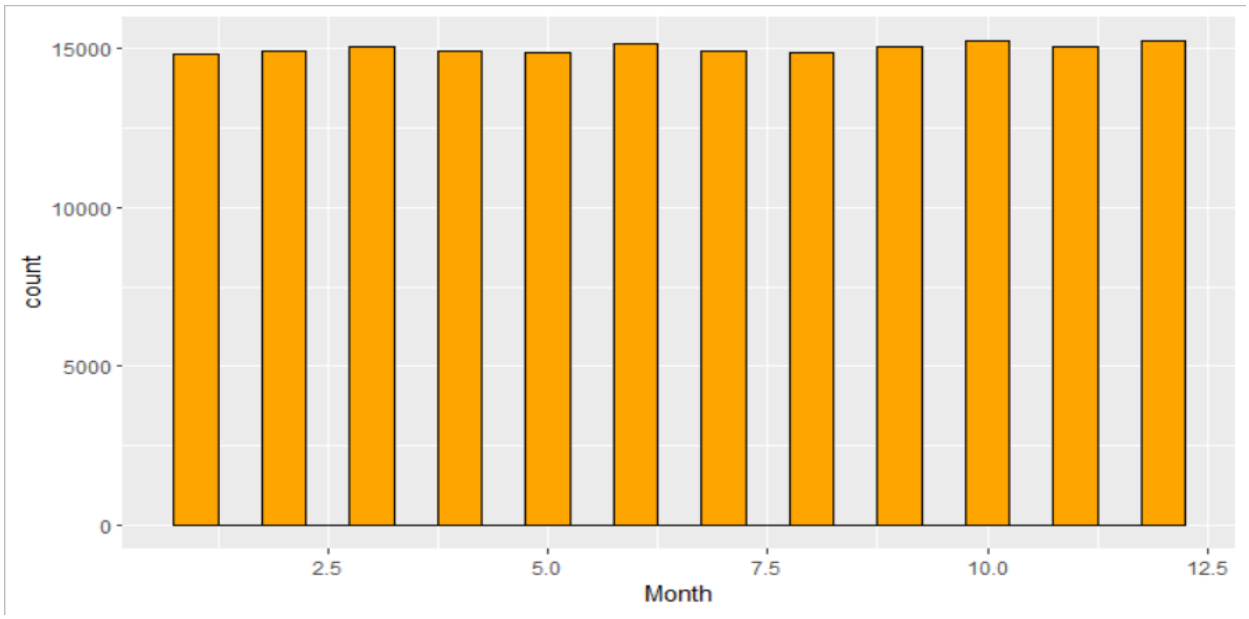

Figure 11: Uniformly distributed graph of the month that an instance occurred in

Figure 12 below portrays a graph from each age of a person who participated in the data, the graph is displayed as a unimodal or right-tailed graph. The mode age of people who participated in the data is roughly 40 years old. It is observed that an increase in age results in less frequent participation in the data.

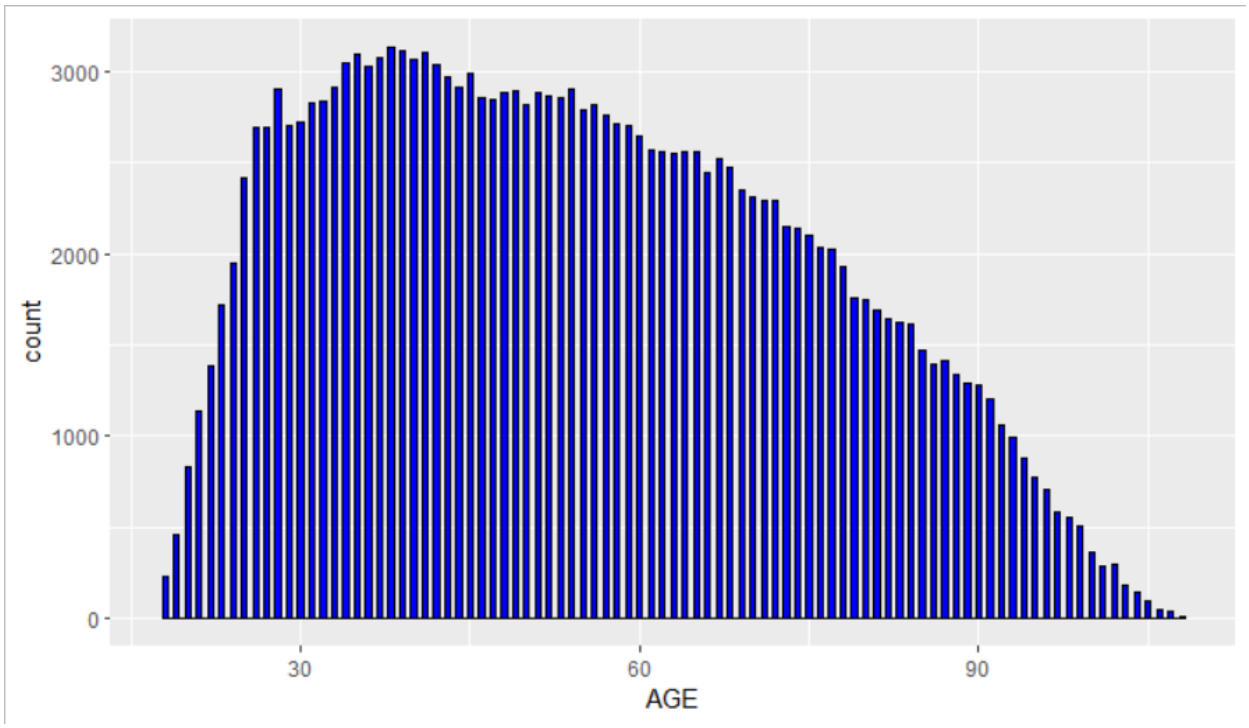

Figure 12: unimodal distribution of the age of people who participates in the data

3.3 Process Capability indices

Figure 13 below clearly indicated that the data contained in the ranges displays a normal distributed graph with a mean of 20 hours. The two red line indicate the given USL=24 hours and LSL=0 hours.

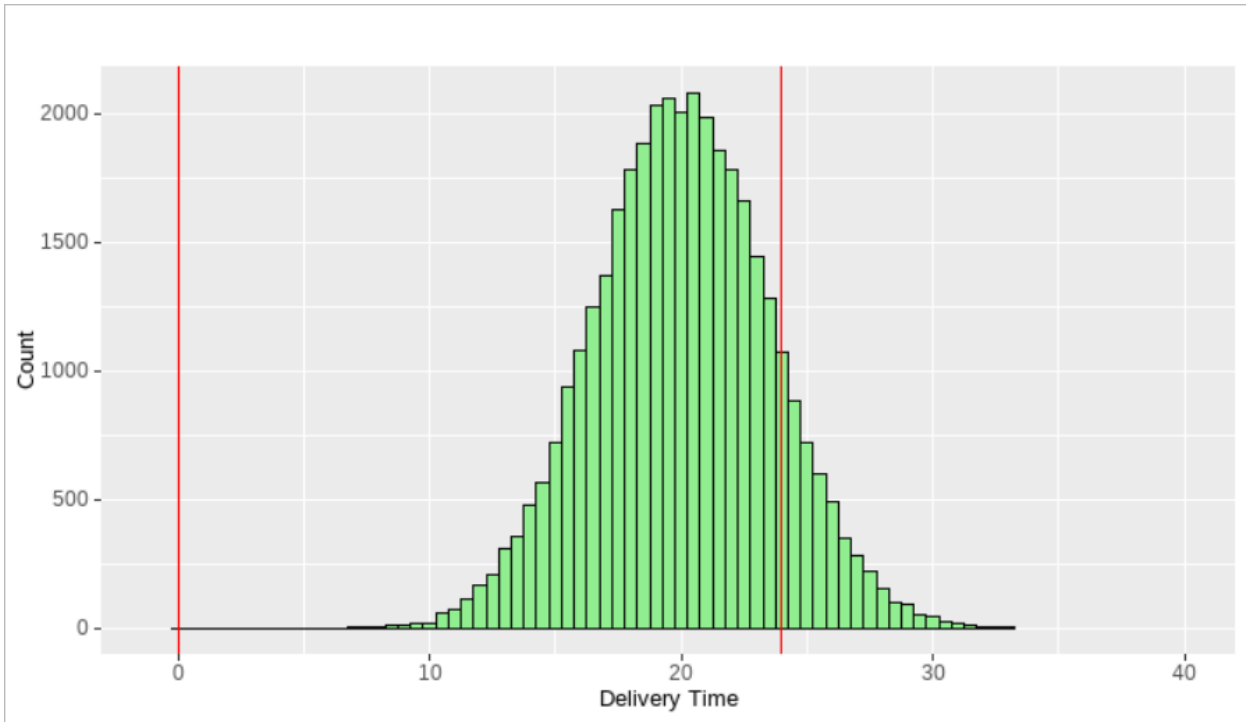

Figure 13: Normal distributed graph of delivery time

3.4 Process Capability calculations

The table below displays the Cp, Cpu, Cpl and Cpk od the delivery of items to customers. It can be observed from the table below that there is a large difference between the Cpl and Cpu. The figure above displays that the delivery time is not centered between the LCL and UCL (red lines in the graph above). The delivery times are closer to the upper specification limit and further away from the lower specification limits.

Cp	1.142207
Cpu	0.3796933
Cpl	1.90472
Cpk	0.3796933

4 PART 3: STATISTICAL PROCESS CONTROL

X&S-charts provide a good and efficient overview to see which measure are out of control, to detect where improvements could be made. 7 different types of items are sold namely technology, clothing, households and luxury items, food, gifts and sweets. The starting date of the items sold is the 1st of January 2021. The figure below displays formulas used to calculate the different sigma controlling limits.

```
U1Sigma <- CL + (UCL - CL)/3
U2Sigma <- CL + (UCL - CL)/3*2
L1Sigma <- CL - (CL - LCL)/3
L2Sigma <- CL - (CL - LCL)/3*2
```

Figure 14: Controlling limit formulas

4.1 X-chart

Figure 15 below displays the X-chart control limits. The chart portrays the first 30 samples. Each sample contains 15 sales, therefore, the total population contains 450 instances. The valid dataset was arranged in order of date before calculations were made. The order is from the oldest date to the latest date, in order of year, month then day. All numeric characters are rounded off to 3 decimal digits.

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	23.032	220146.000	21.260	20.374	19.489	18.603	17.717
Clothing	9.410	9.263	9.112	8.970	8.823	8.677	8.530
Household	0.852	49.072	47.817	0.542	45.307	44.052	0.232
Luxury	5.507	5.250	4.993	4.736	4.478	4.221	3.964
Food	2.714	2.639	2.565	2.490	2.415	2.341	2.266
Gifts	9.507	9.125	8.743	8.361	7.979	7.597	7.215
Sweets	2.906	2.763	2.620	2.478	2.335	2.193	2.050

Figure 15: X-chart

4.2 S-chart

Figure 16 below illustrates the S-chart control limits. The chart portrays the first 30 samples. Each sample contains 15 sales, therefore, the total population contains 450 instances of sales for each item. The valid dataset was arranged in order of date before calculations were made. The order is from the oldest date to the latest date, in order of the year followed by the month and then the day. All numeric characters are rounded off to 3 decimal digits.

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.2950	4.6530	4.0110	3.3680	2.7270	2.0860	1.4440
Clothing	0.8770	0.7700	0.6640	0.5580	0.4513	0.3450	0.2390
Household	7.5020	6.5920	5.6820	4.7720	3.8620	2.9530	2.0430
Luxury	1.5370	1.3500	1.1640	0.9780	0.7910	0.6040	0.4180
Food	0.4450	0.3920	0.3380	0.2840	0.2300	0.1760	0.1220
Gifts	2.2840	2.0070	1.7300	1.4530	1.1760	0.1760	0.6220
Sweets	0.8520	0.7490	0.6450	0.5420	0.4390	0.3350	0.2320

Figure 16: S-chart

4.3 Initial control limits of the first 30 samples

The subsequent graphs represents the first 30 samples of each item sold by the business. Included in the graphs are the center limit (CL), upper center limit (UCL), and the lower center limit (LCL). If all the points in the graph is located within the ULC and LCL then the class in is control. It is observed that all the classes are in control except for the sweets class. Sample number 18 of the sweets class standard deviation in the S-chart is located above the UCL and therefore it is out of control and this sample needs to be removed.

4.3.1 Technology

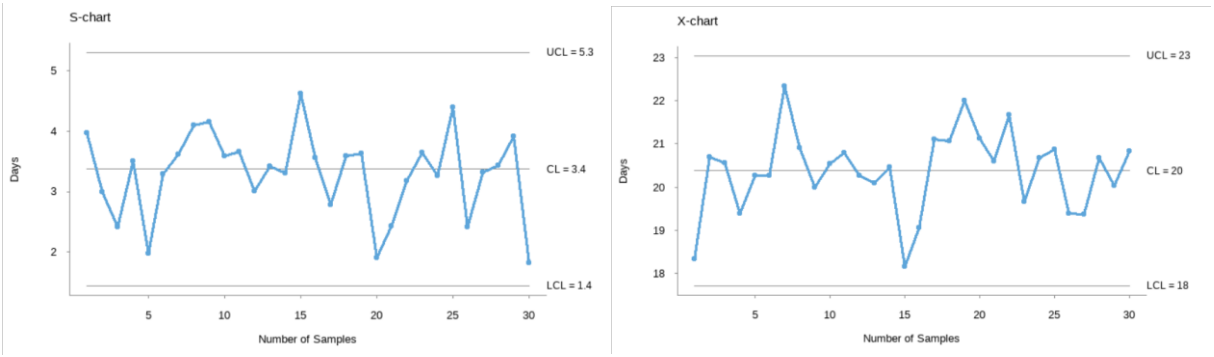

Figure 17: S&X-chart for technology items

4.3.2 Clothing

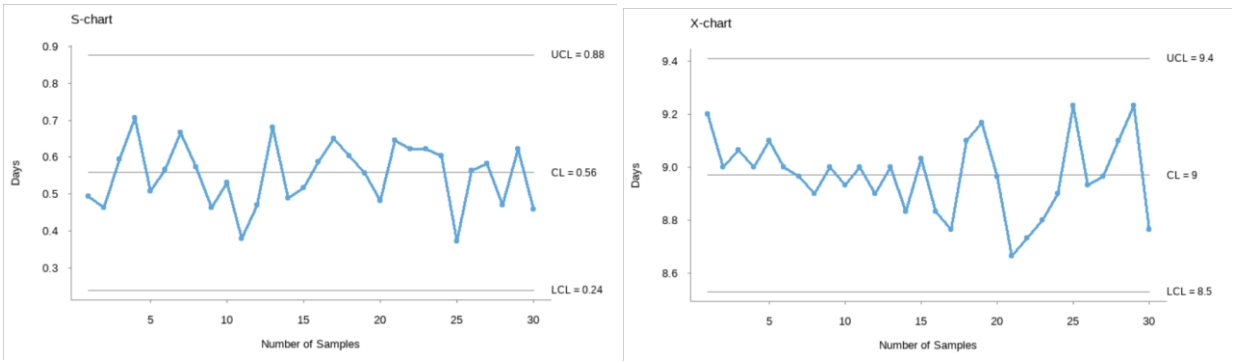

Figure 18: S&X-chart for clothing items

4.3.3 Household

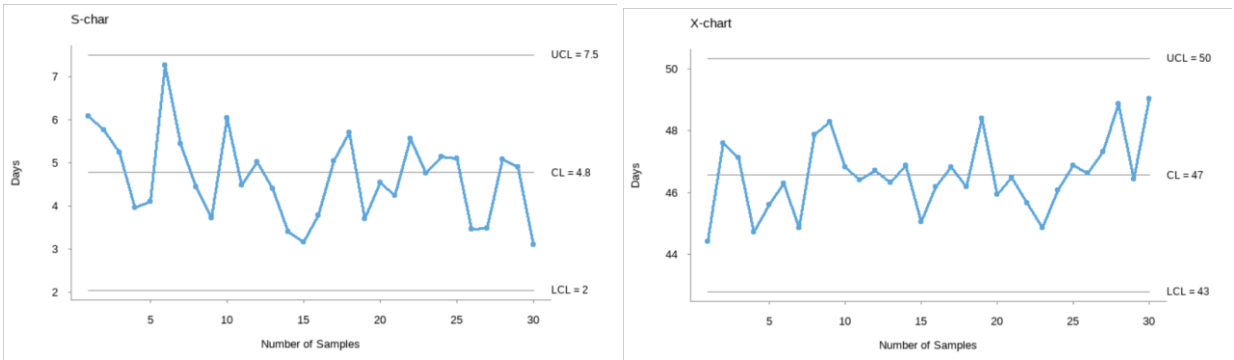

Figure 19: S&X-chart for household items

4.3.4 Luxury

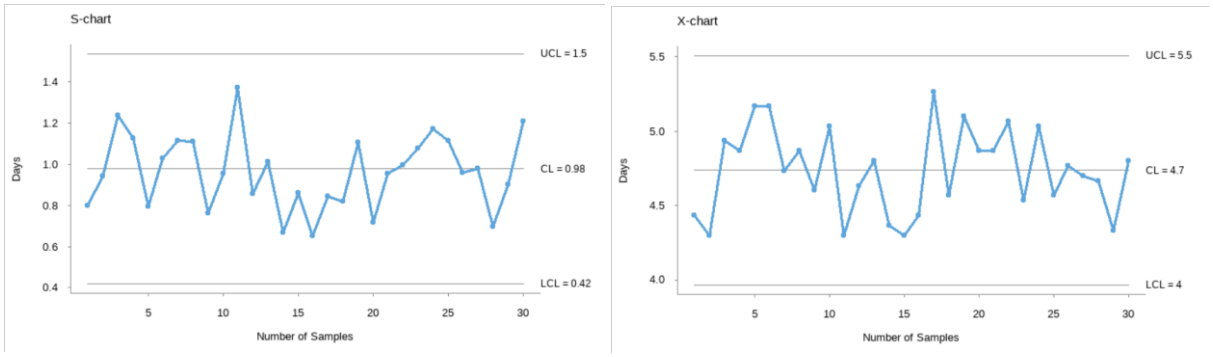

Figure 20: S&X-chart for luxury items

4.3.5 Food

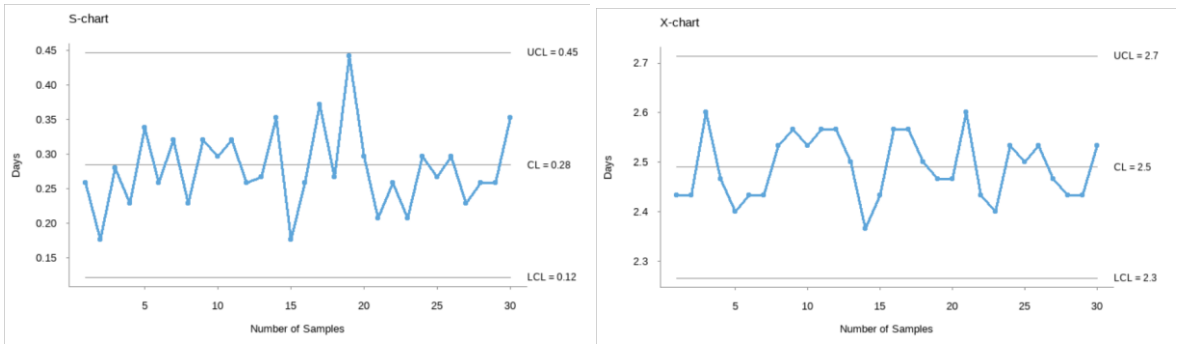

Figure 21: S&X-chart for food items

4.3.6 Gifts

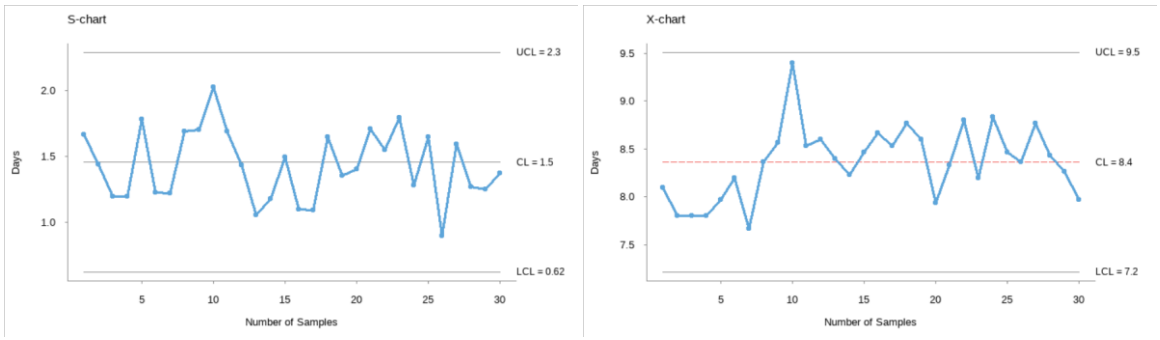

Figure 22: S&X-chart for gifts items

4.3.7 Sweets

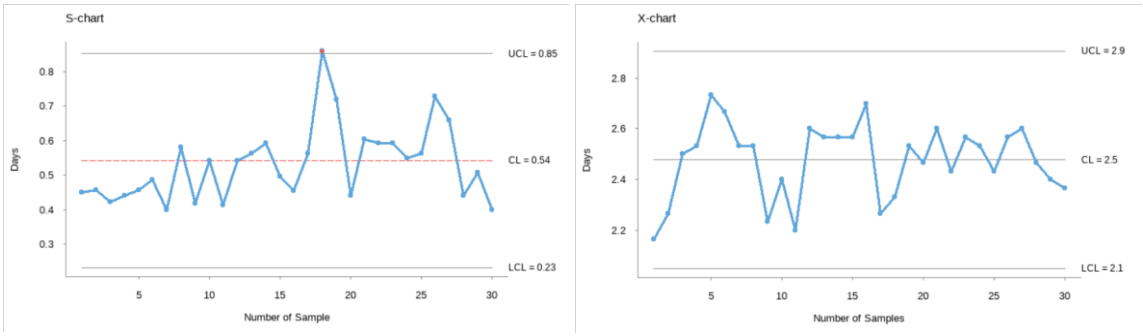

Figure 23: S&X-chart for sweets items

4.4 Control limits of all the data

The following data consists of S&X-charts of all the data following the first 30 samples. All the samples contain 15 sales for each class. The graphs portrays if the delivery times of each item is within or outside the controlling limits.

4.4.1 Technology

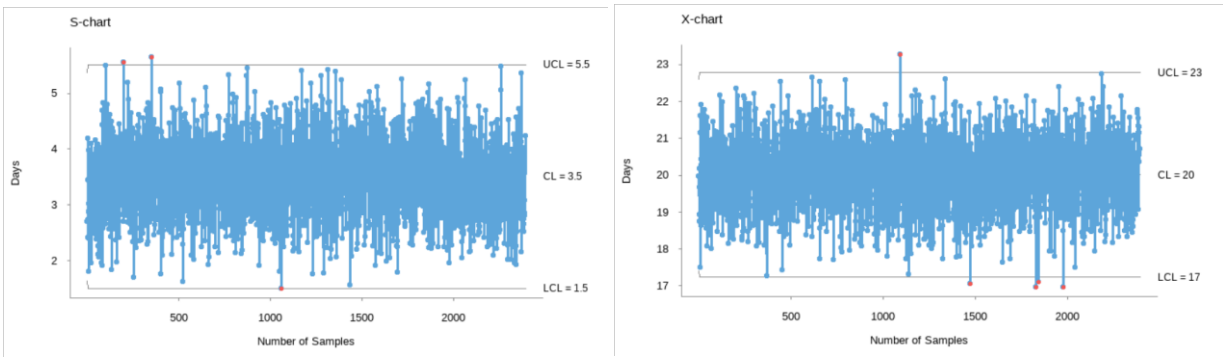

Figure 24: S&X-chart for technology items

The majority of the samples are located within the controlling limits. There are very few instances that are outside the controlling limits. It is accepted that technology is under control and that the x-bar is appropriate.

4.4.2 Clothing

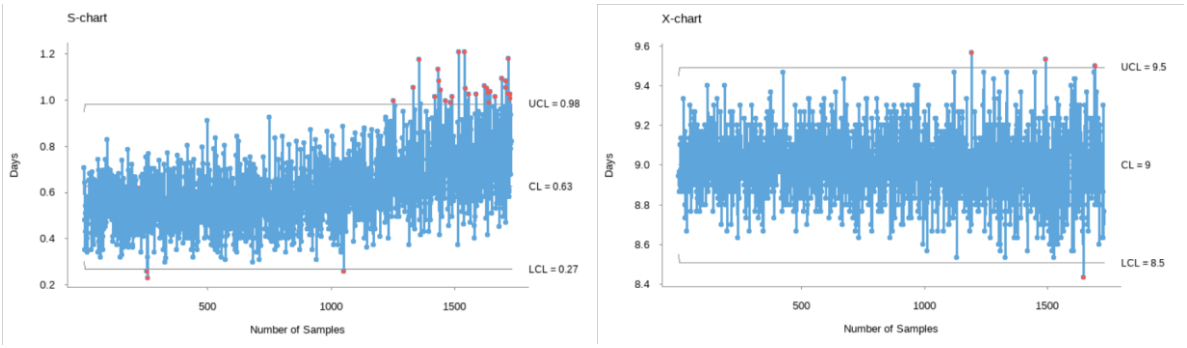

Figure 25: S&X-chart for clothing items

Most of the samples are located within the controlling limits. There are some samples that lay outside the controlling limits in the S-bar and they are close to one another. Seasonal sales could be a measure of these samples being outside the controlling limits. The X-bar has corrected this error and the X-bar chart is accepted as being appropriate.

4.4.3 Household

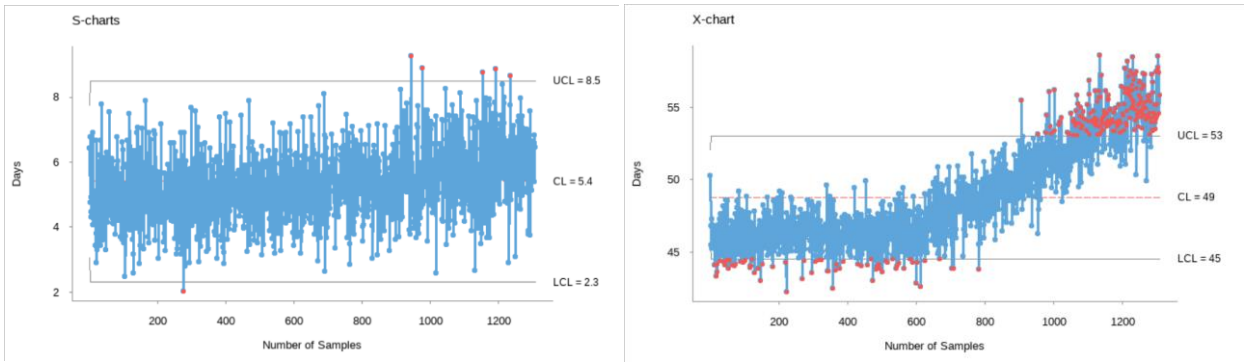

Figure 26: S&X-chart for household items

Delivery times for household items are relatively stable and then it suddenly increases drastically. The delivery times for household items are therefore not in control and investigation needs to be appointed to why delivery times increase. The increase of delivery times of household items could relate to an increase of orders which should be delivered at the same time, causing an increase in delivery times.

4.4.4 Luxury

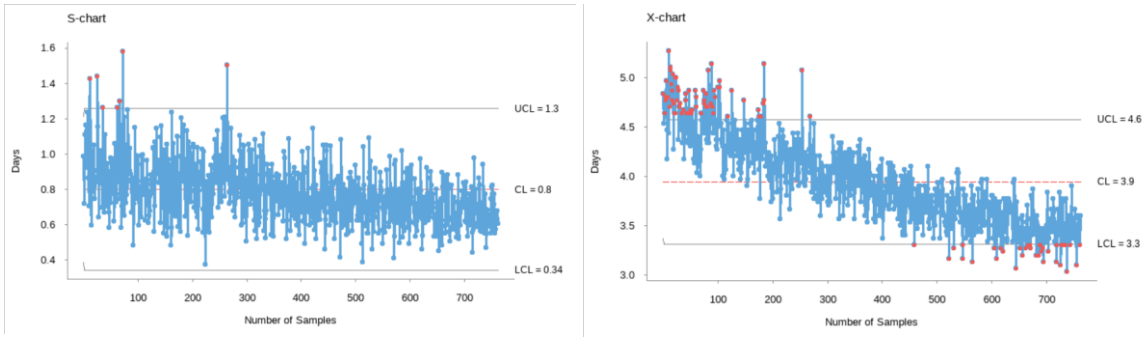

Figure 27: S&X-chart for luxury items

Delivery times of luxury items decrease constantly from the start to the finish. A reason for delivery times to decrease could be that luxury items should be delivered faster to increase sales and generate higher profits. The S-bar however portrays that only 6 samples are outside the controlling limits. Therefore, the S-bar can be accepted as being under control.

4.4.5 Food

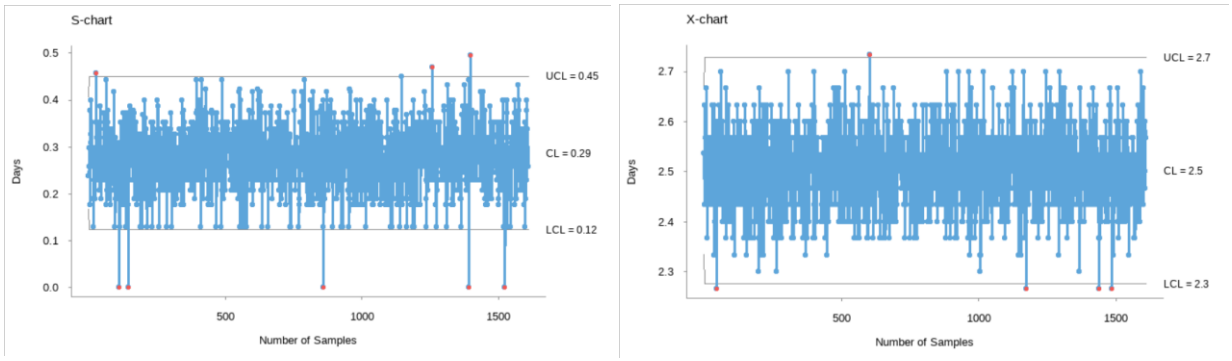

Figure 28: S&X-chart for food items

The S-bar and X-bar for food only has a few instances that are out of control. It can be concluded that both the S-bar and X-bar is under control.

4.4.6 Gifts

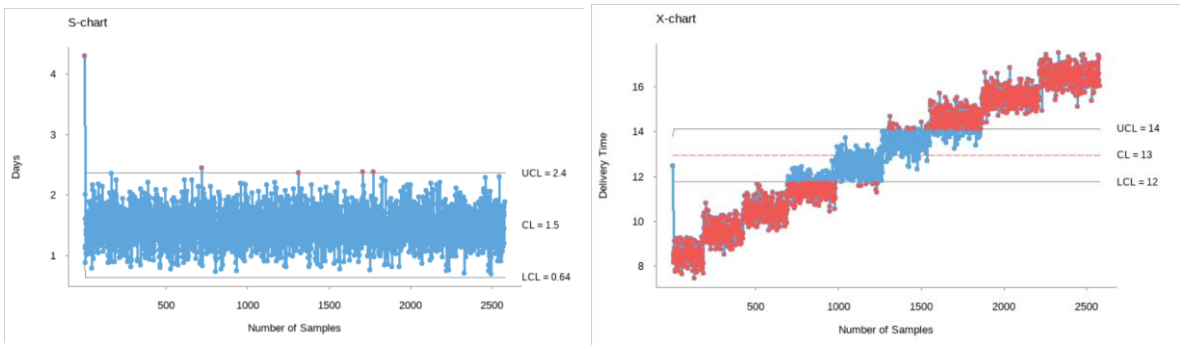

Figure 29: S&X-chart for gift items

The delivery times for gifts increase constantly and the reason for that needs to be investigated. The delivery times are not stable at any point for all the samples. A reason for an increase in delivery times could be that the demand for gifts increases overtime. The S-bar in under control because it only consists of 5 samples that are located outside the controlling limits.

4.4.7 Sweets

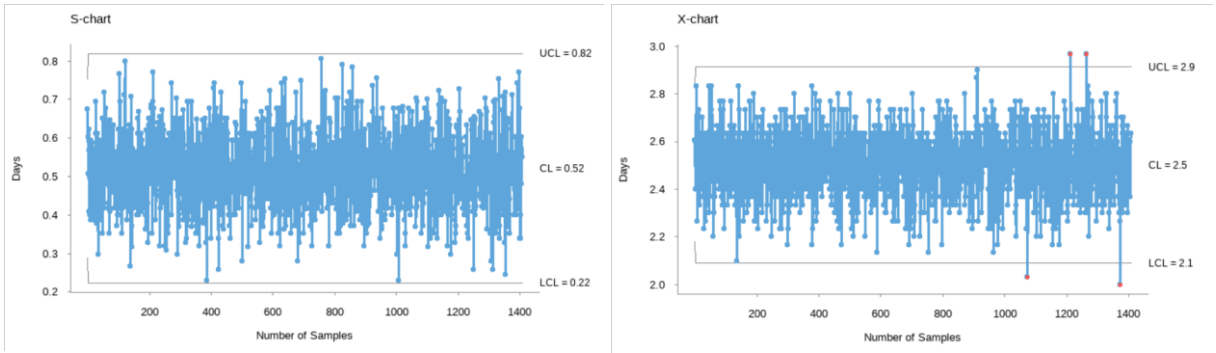

Figure 30: S&X-chart for sweets items

The S-bar for sweets is completely under control since no samples are located outside the controlling limits. The delivery times for gifts only have 4 samples that are outside the controlling limits, 4 of the 5 samples have happened recently and investigation needs to be appointed to prevent this issue for future deliveries.

5 PART 4: OPTIMIZING DELIVERY PROCESSES

5.1 Out-of-control samples

Figure 22 below portrays the total of samples that are located outside the controlling limits. The figure is a good visual to identify which classes need to be investigation for improvement. Furthermore, it also displays the first 3 and last 3 samples that are located outside the controlling limits. It is evident from the figure below that household, luxury and gift items have the greatest quantity of out-of-control limits. Technology, clothing, food and sweet of the other hand only has a few samples that are outside of the controlling limits, subsequently meaning that they are in control.

	Total	1st	2nd	3rd	3rd Last	2nd Last	Last
Technology	12	37	398	483	1872	2009	2071
Clothing	14	455	702	1152	1677	1723	1724
Household	3934	252	387	643	1335	1336	1337
Luxury	414	184	207	227	789	790	791
Food	1	633	633	633	633	633	633
Gifts	2282	213	216	218	2607	2608	2609
Sweets	4	1104	1243	1294	1243	1294	1403

Figure 31: Out of control samples

5.2 Most consecutive samples between -0.3 and 0.4 sigma-control limits

The figure below portrays the most consecutive instances for each class that is situated between the -0.3 and +0.4 sigma-control limits. It also portrays the last sample in the given range.

	Consecutive	Last position
Technology	6	776
Clothing	4	223
Household	3	45
Luxury	4	63
Food	7	94
Gifts	7	477
Sweets	3	56

Figure 32: Consecutive instances

5.3 Estimation of making a type I error

The following Null hypothesis was given: the process is in control and centered on the centerline calculated using the first 30 samples. The alternative hypothesis was given as follows: that the process in not in control and has moves from the centerline or has increased or decreased in variation. A type I error is classified when the process is classified as out of control relative to the data, but patterns

incorrectly justified that the process is in control. The figure below illustrates how a type I error works in a table format.

	Process is in control	**Process is out of control**
SPC portrays that the process is out of control	Type I error from the manufacturer	Correct indication to improve process
SPC portrays that the process is in control	Correct indication, do not do anything	Type II error from the customer

It can be concluded that a type I error can be classified as the manufacturers error.

5.4 Optimize delivery process

The following information was given:

For this part, you should use the individual delivery times and not the samples. Use all the data that is available. If you lose R329/item-late-hour in lost sales if you deliver technology items slower than 26 hours, and it costs you R2.5/item/hour to reduce the average time by one hour, on how many hours should you center the delivery process for best profit?

How to solve this question:

To start the calculations to minimize the delivery cost, it is required to calculate the current additional cost. It was found that the average delivery time should decrease by 14 days to require an average delivery time of 31 days. This generated the cost to be minimized and the profit to be maximized fro the business.

5.5 Type II error

A type II error is established when a process is in control and stay within its controlling limits, but the process is actually out of control. Type II errors occur when no hidden patterns in the data is established even though there should have been. Sample space is an important feature to have correct to avoid type II errors.

6 PART 5: DOE AND MANOVA

MANOVA is a powerful test to indicate which variables in your dataset correlates with other variables in your dataset. The variables in our dataset will be appointed to the 7 different features of our dataset, namely: technology, clothing, luxury, food, gifts, sweets and household items.

The hypothesis for this test means that at least one of our features has a correlation with another feature. The null hypothesis for this test means that one feature in our dataset has no correlation with any of the other feature in our dataset. Information gained from these two hypotheses provides hidden patterns in our dataset and therefore it is easier to distinguish problems in our company to avoid in the future.

6.1 Age compared to the reason for purchasing a product

The features that are going to be inspected is the correlation between the age of a person who bought an item and the reason for buying that item. Research shows that as the age of a person increases the less technological active they are (Smith, 2020). In figure 22 below it can be observed that mean age is a person who bought an item due to using email in slightly older than someone who bought an item due to visiting the company's website. Therefore, these two features are compared to one another to analyse if the age has an impact on the reason for purchase of items.

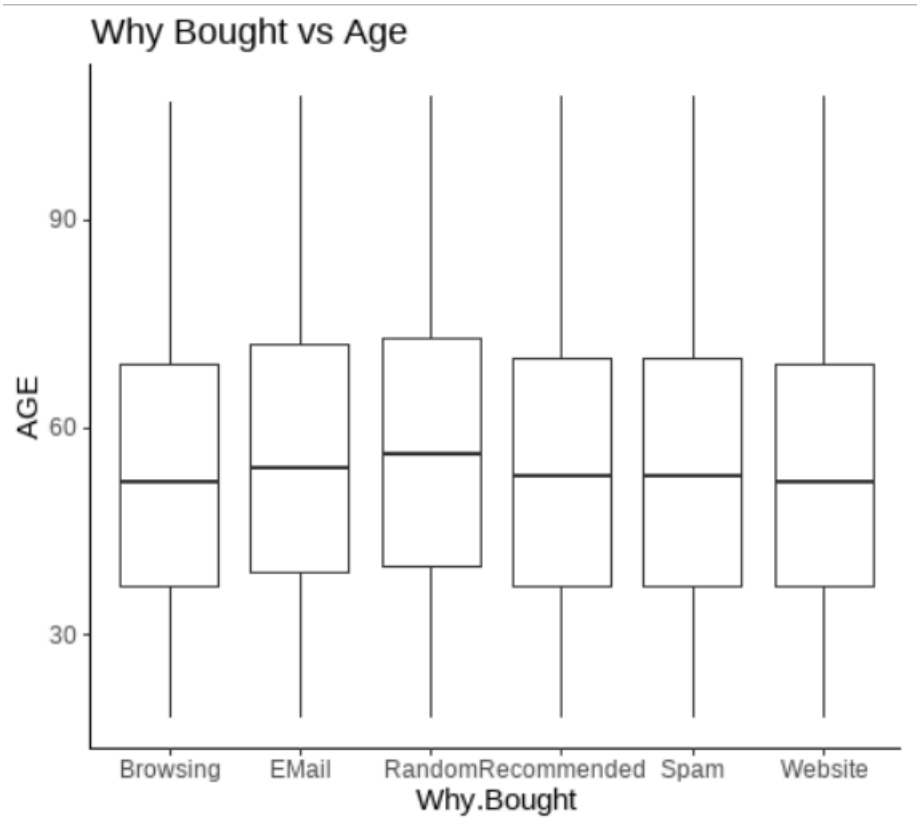

Figure 33: Age of a person vs why they bought it

A P value of 0.05 is chosen for the MANOVA test since it is the most common P value used worldwide.

A MANOVA test in RStudio between the age of a person and the reason they bought an item is going to be analysed. The table below provides a summary of the outcomes calculated in RStudio.

Independent variable	Reason for purchase
Dependant variable	Age and why bought
NULL hypothesis	The age of a person has no influence on the reason for purchasing a product
Alternative hypothesis	At least one age has an influence on the reason of purchase of a product

The P values calculated for the test equals 2.2e-16.

The results of the P value generated will result in rejecting the NULL hypothesis since the p value that is generated is less than the chose P value and accepting the alternative hypothesis. Therefore, it is accepted that the age of a person does influence the reason for purchasing a product.

6.2 Why bought compared to the price of the item

The features that are going to be inspected is the correlation between the price of an item an item and the reason for buying an item. In the figure below it can be observed that the price differs for reasons that a person bought a product.

Figure 34: Why bought compared to price

The MANOVA test is going to be used to analyze if there is any correlation between the price and the reason for purchasing an item. A P value of 0.05 is being used since it is the most common value used worldwide for an MANOVA test. The table below generates a summary of result conducted in RStudio.

Independent variable	Reason for purchase
Dependant variable	Price of item purchased
NULL hypothesis	The price of an item has no influence on the reason for purchasing a product
Alternative hypothesis	At least one price has an influence of the reason of purchase of a product

The P value generated for this test is equal to 2.2e-16. Since the P value generated is less than the P value chosen initially results in rejecting the NULL hypothesis. Therefore, the alternative hypothesis will be accepted.

7 PART 6: RELIABILITY OF THE SERVICE AND PRODUCTS

7.1 Problem 6

Information given:

Variation of the process is having an allowance of 0.04cm. A variation of more than 0.04cm would cost the business $45 per item.

7.1.1 Calculating variables

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

$45 = K(0.04)^2$

$K = 28125$

Taguchi Loss Function:

$L(x) = 28125(x-0.06)^2$

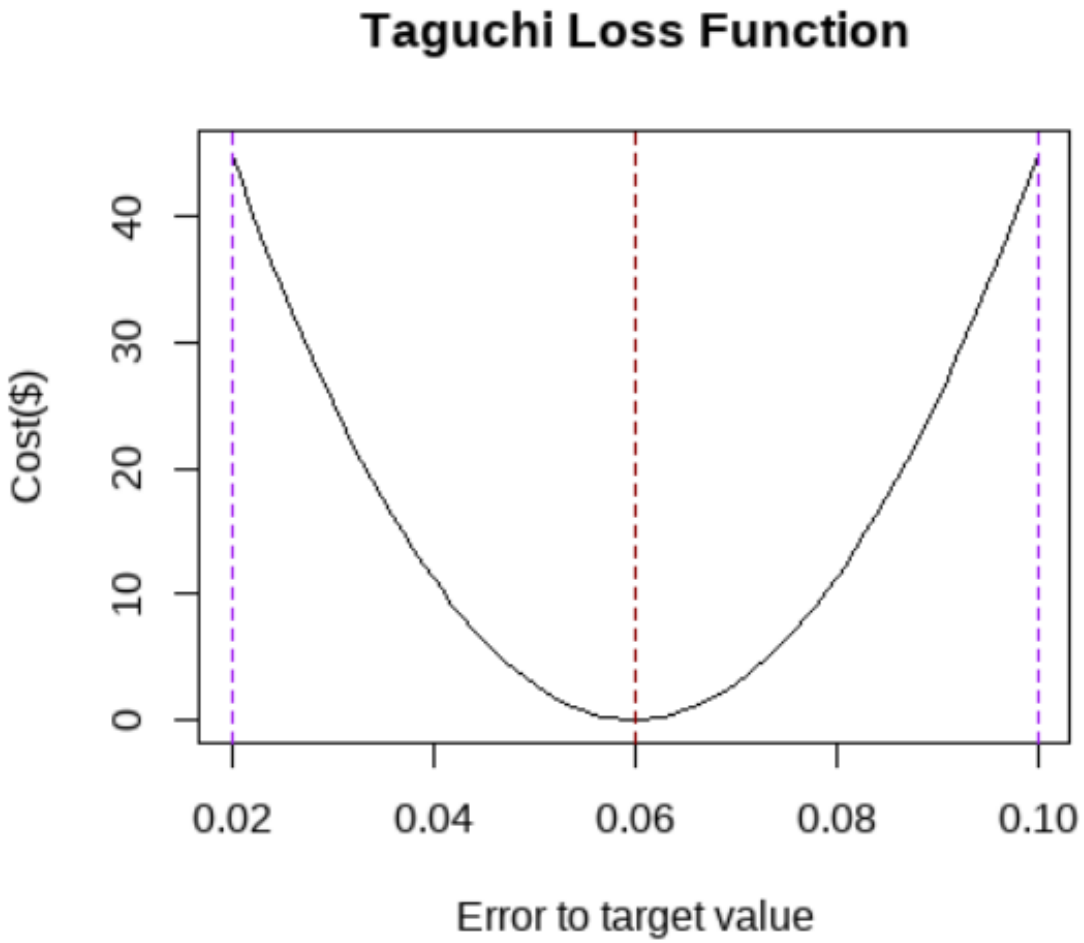

Figure 35: Taguchi Loss Function

The target value or the product is 0.06. The bigger the variance is from the target value will relate to a decrease in the products quality. It will also generate an increase in a loss for the company as the variance increases from 0.06. The company should avoid a high variance from the targe value since it will subsequently yield the efficiency and product unreliability to decrease.

7.2 Problem 7

7.2.1 a)

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

$35 = K(0.04)^2$

$K = 21875$

Taguchi Loss Function:

$L(x) = 21875(x-0.06)^2$

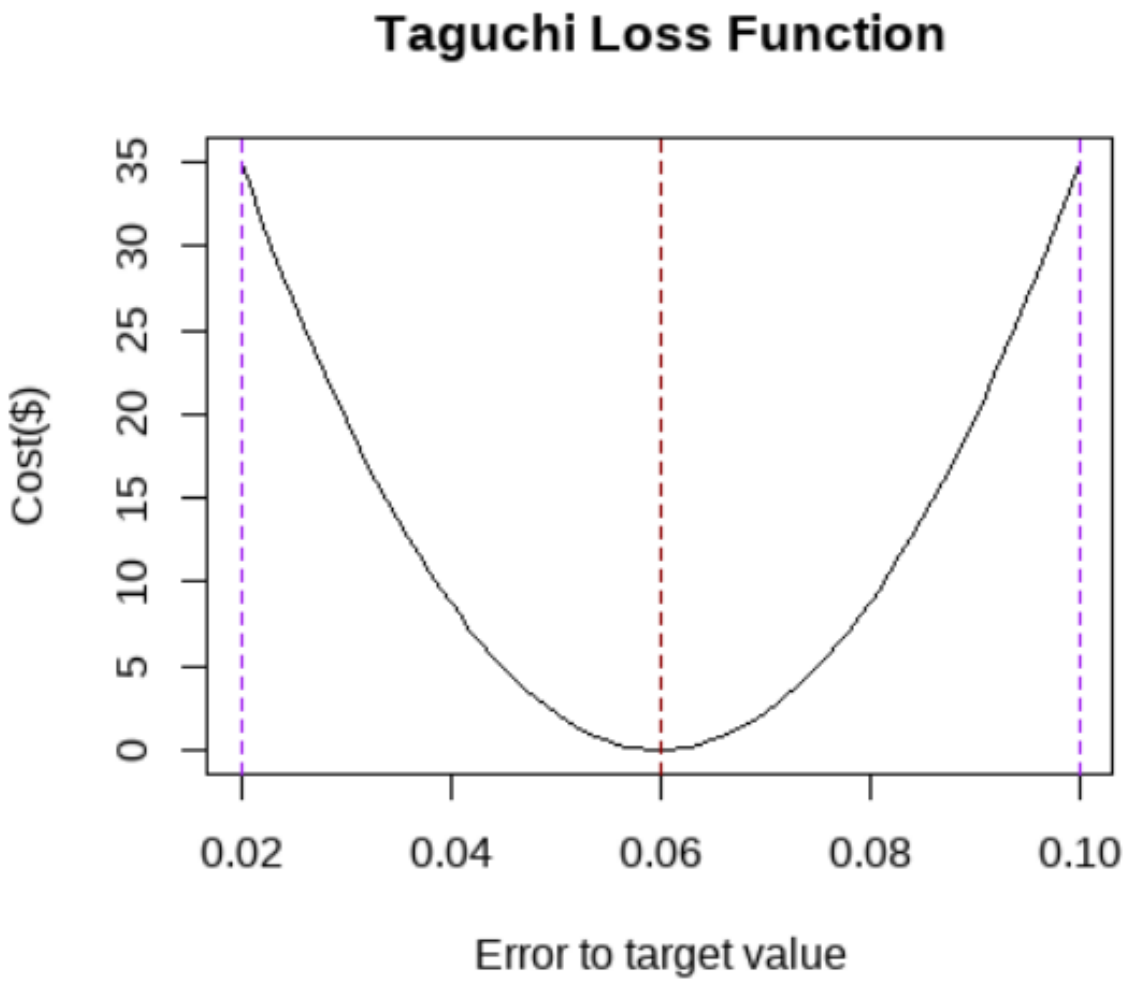

Figure 36: Taguchi Loss Function

The target value or the product is 0.06. The bigger the variance is from the target value will relate to a decrease in the products quality. It will also generate an increase in a loss for the company as the variance increases from 0.06. The company should avoid a high variance from the targe value since it will subsequently yield the efficiency and product unreliability to decrease.

7.2.2 b) If the process deviation from the target can be reduced to 0.027cm

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

$L(0.072) = 15.94687$

Due to the process deviation from the target value that is reduced means that the company will generate a lower loss, it will cost the company $15.95 for each item that is produced outside the error limits. The reduction of the process deviation highlights how important it is to consisted waste reduction when manufacturing products.

7.3 Problem 27

7.3.1 a) The probability when only one machine is allocated to a station

$$\text{Probability of the system working} = 0.85 \times 0.92 \times 0.90$$

$$= 0.7038$$

7.3.2 b) The probability when two machines are allocated to a station

$$\text{Probability of the system working} = (1 - (1 - 0.85)^2) \times (1 - (1 - 0.92)^2) \times (1 - (1 - 0.90)^2)$$

$$= 0.9615$$

By replacing one machine per station by two machines per station generates an increase in probability that the system will work by 25,77%. The reason for the increase in reliability in the system is due to the extra identical machine that can still function even when the other machine has failed. The company will benefit greatly by using a manufacturing system where there are two identical machines as each station. It will increase the throughput quantity in the future for the business.

7.4 Section 6.3

<u>The question that should be answered is stated below:</u>

For the delivery process, there are 21 delivery vehicles available, of which 20 is required to be operating at any time to give reliable service. During the past 1560 days, the number of days that there were only 20 vehicles available was 190 days, only 19 vehicles available was 22 days, only 18 vehicles available was 3 days and 17 vehicles available only once. There are also 21 drivers, who each work an 8-hour shift per day. During the past 1560 days, the number of days that there were only 20 drivers available was 95 days, only 19 drivers available was 6 days and only 18 drivers available, once only. Estimate on how many days

per year we should expect reliable delivery times, given the information above. If we increased our number of vehicles by one to 22, how many days per year we should expect reliable delivery times?

7.4.1 Reliability delivery times

The equation below was used in RStudio to calculate the following answers:

$$P(x) = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}$$

The equation provided above is the binomial equation. The equation calculates the probability of exactly the number of successes of a instance happening. All calculation where done is RStudio and the final values have been rounded to the nearest 3 decimal digits.

Calculations for 21 vehicles:

In order for the company to produce its customers with a reliable delivery process the company will require 20 out of 21 vehicles to be available at all times. Meaning that only one vehicle at most can be absent at any time.

In the calculation it was calculated that if exactly zero vehicles fail in one whole year the probability would be 0.861 or 86.1%. If a calculation is done to calculate the number of days in a year that contains 365 then the quantity of zero vehicles that fain in a year will be: 365*0.861= 314.265 days.

The company however only requires 20 vehicles to be available at a single any time during the year, meaning that the company will still be able to supply reliable services to its customers when one vehicle is not working. When calculations are made to find the reliability of 20 vehicles that are available for delivery throughout the year the probability equals 0.990 or 99.0%. The number of days in the year where 20 vehicles are available is equal to 361.5 days.

For the company to deliver items to customers the vehicles and drivers' reliability should be taken into consideration since each vehicle needs a driver to deliver products. Therefore, since 20 vehicles should be available at all times, then 20 drivers should also be available. There are 21 drivers so only one driver can be absent at a time. The probability that no driver is absent from services equals to 0.934 or 93.4%. Meaning that in one year there will be 341.07 days that all drivers are available.

The company however only needs 20 drivers to be available at a time. The probability that only 1 driver is absent at all times during a year is 0.0635 or 6.35%. Meaning that is one year that contains 365 days there will be one driver that will be absent 23.169 days.

When both the reliability of at least 20 drivers and vehicles are taken into consideration then the probability of that happening is 0.9883 or 98.83% in one year. Subsequently that in ne whole year there will be 360.74 days where the company will product reliable delivery services to its customers.

If the number of days that the company produces reliable service is rounded to the nearest integer than it will results in the company to produce customers with reliable service for 361 days in a year that consists of 365 days.

When the number of vehicles is increased to 22:

The same calculations for this question are made, the only adjustment made to answer this question is that there could be 2 vehicles absent at a time instead of one vehicle. When the same method for calculating the probability of 20 out of 22 vehicles are made, it is concluded that the probability of at least 20 vehicles and drivers being available at any time is 0.9878or 99.95%. Subsequently this will enable the company to produce customers with reliable services for 364.056 days in a year. If the number of days is rounded to the nearest integer, then the company will supply reliable service for 364 days in a year.

8 CONCLUSION

It was established that data wangling is needed to analyse data accurately. Removing missing values and negative values is essential to data wangling. Descriptive statistics was performed on all 7 features of the business. Hidden patterns in the dataset wad found.

Delivery process times was studied and portrayed in a X&S chart. Statistical process control limits where established. Type I and type II errors have been explained. DOE and MANOVA tests were done of the data. Statistical related questions were answered. All calculations were made using a programming language namely RStudio.

9 REFERENCES

Hernandez, F. (2015) *Data analysis with R - exercises* - . Available at: http://fch808.github.io/Data-Analysis-with-R-Exercises.html (Accessed: October 1, 2022).

Krunal, J. (2019) *A guide to control charts*, *iSixSigma*. Available at: https://www.isixsigma.com/tools-templates/control-charts/a-guide-to-control-charts/#:~:text=Control%20limits%20are%20calculated%20by,the%20average)%20for%20the%20LCL (Accessed: October 6, 2022).

MANOVA, D. (2022) *MANOVA test in R: Multivariate analysis of variance*, *STHDA*. Available at: http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance (Accessed: October 3, 2022).

Smith, A. (2020) *Older adults and technology use*, *Pew Research Center: Internet, Science & Tech*. Pew Research Center. Available at: https://www.pewresearch.org/internet/2014/04/03/older-adults-and-technology-use/ (Accessed: October 15, 2022).