ECSA GA4 PROJECT

Quality Assurance 344

Coetzee, JJJ
22619755

Contents

Table of Figures

List of Abbreviations

UCL – Upper Control Limit

LCL – Lower Control Limit

NA – Not Available

USL – Upper Service Level

LSL – Lower Service Level

ANOVA – Analysis of Variance

Introduction

Sales data from an online trading company has been collected. Statistical analysis will be done on the collected data in order to achieve a good overview of what is going on in the company. A conclusion about the state of the company will be made after all results have been discussed.

Part 1: Data Wrangling

Upon inspection of the data, it was found that there are instances in the original dataset which could lead to errors being shown. This is due to the values for the Price feature being either missing or it is having a negative value. Price was the only feature in the dataset that had values which could be considered as invalid.

In total 22 invalid data values were extracted from the original dataset as shown below:

	id	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
12345	1	12345	18973	93	Gifts	*NA*	2026	6	11	15.5	Website
16320	2	16320	44142	82	Household	-588.8	2023	10	2	48.0	EMail
16321	3	16321	81959	43	Technology	*NA*	2029	9	6	22.0	Recommended
19540	4	19540	65689	96	Sweets	-588.8	2028	4	7	3.0	Random
19541	5	19541	71169	42	Technology	*NA*	2025	1	19	20.5	Recommended
19998	6	19998	68743	45	Household	-588.8	2024	7	16	45.5	Recommended
19999	7	19999	67228	89	Gifts	*NA*	2026	2	4	15.0	Recommended
23456	8	23456	88622	71	Food	*NA*	2027	4	18	2.5	Random
34567	9	34567	18748	48	Clothing	*NA*	2021	4	9	8.0	Recommended
45678	10	45678	89095	65	Sweets	*NA*	2029	11	6	2.0	Recommended
54321	11	54321	62209	34	Clothing	*NA*	2021	3	24	9.5	Recommended
56789	12	56789	63849	51	Gifts	*NA*	2024	5	3	10.5	Website
65432	13	65432	51904	31	Gifts	*NA*	2027	7	24	14.5	Recommended
76543	14	76543	79732	71	Food	*NA*	2028	9	24	2.5	Recommended
87654	15	87654	40983	33	Food	*NA*	2024	8	27	2.0	Recommended
98765	16	98765	64288	25	Clothing	*NA*	2021	1	24	8.5	Browsing
144443	17	144443	37737	81	Food	-588.8	2022	12	10	2.5	Recommended
144444	18	144444	70761	70	Food	*NA*	2027	9	28	2.5	Recommended
155554	19	155554	36599	29	Luxury	-588.8	2026	4	14	3.5	Recommended
155555	20	155555	33583	56	Gifts	*NA*	2022	12	9	10.0	Recommended
166666	21	166666	60188	37	Technology	*NA*	2024	10	9	21.5	Website
177777	22	177777	68698	30	Food	*NA*	2023	8	14	2.5	Recommended

Figure 1: Invalid Data

Two observations can be made from this invalid data set. Firstly, it can be seen that there are 17 NAs and 5 negative values for the invalid data. Secondly, whenever the value is negative it is equal to -588.8 which cannot be explained.

Part 2: Descriptive Statistics

An effective way of familiarising yourself is by looking at the features individually. Upon inspection of the dataset, it was found that the features “Class”, “Age”, “Price” and “Delivery.Time” could provide the most insight on the dataset and show some of the tendencies in the data.

Age

Minimum	1st Quarter	Median	Mean	3rd Quarter	Maximum
18.00	38.00	53.00	54.57	70.00	108.00

Table 1: Age Statistics

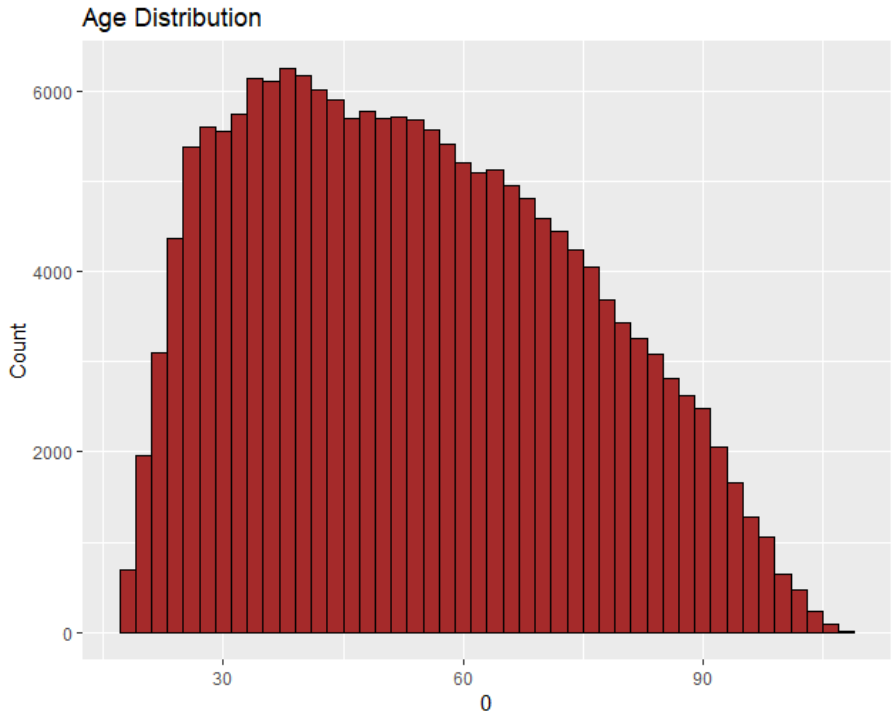


Figure 2: Age Distribution Plot

As expected, Age is normally distributed. This indicates a high volume of customers between ages 38 and 70, after which the sales per age group drastically decreases per age. This can be backed by the fact that there are fewer people that reach those high ages.

The distribution also shows that the minimum age recorded is 18, which could indicate that sales are only being made to customers over the age of 18 years.

Price

Minimum	1st Quarter	Median	Mean	3rd Quarter	Maximum
35.65	482.31	2 259.63	12 294.10	15 270.97	116 618.97

Table 2: Price Statistics

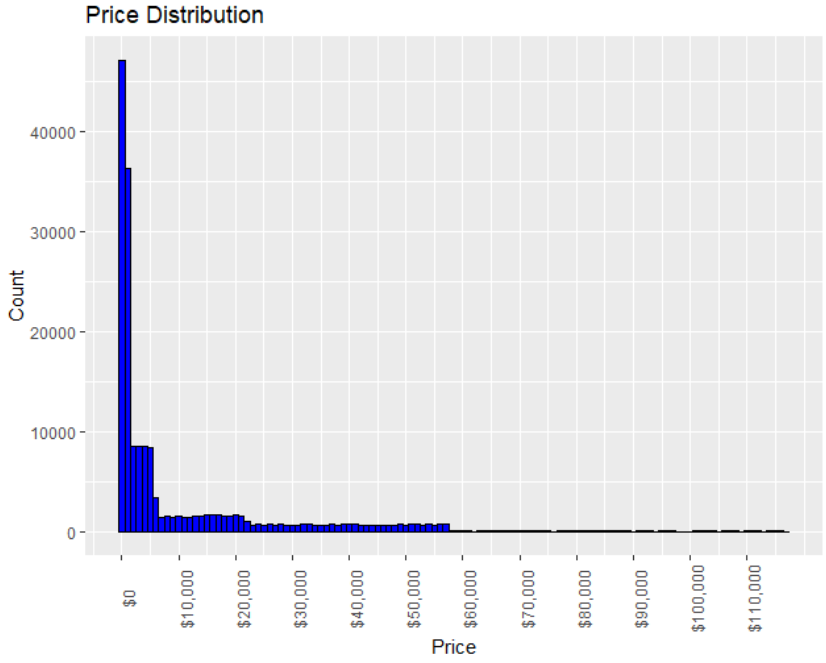


Figure 3: Price Distribution Plot

Analysing this price distribution, the assumption can be made that there are little sales with a price higher than $60 000. This could be due to the fact that all of those sales are for a specific class. As expected, there are no values below $0 as all of the negative and missing values have been removed. From the

To better understand the data, it might be useful to view the data as a distribution of price per class.

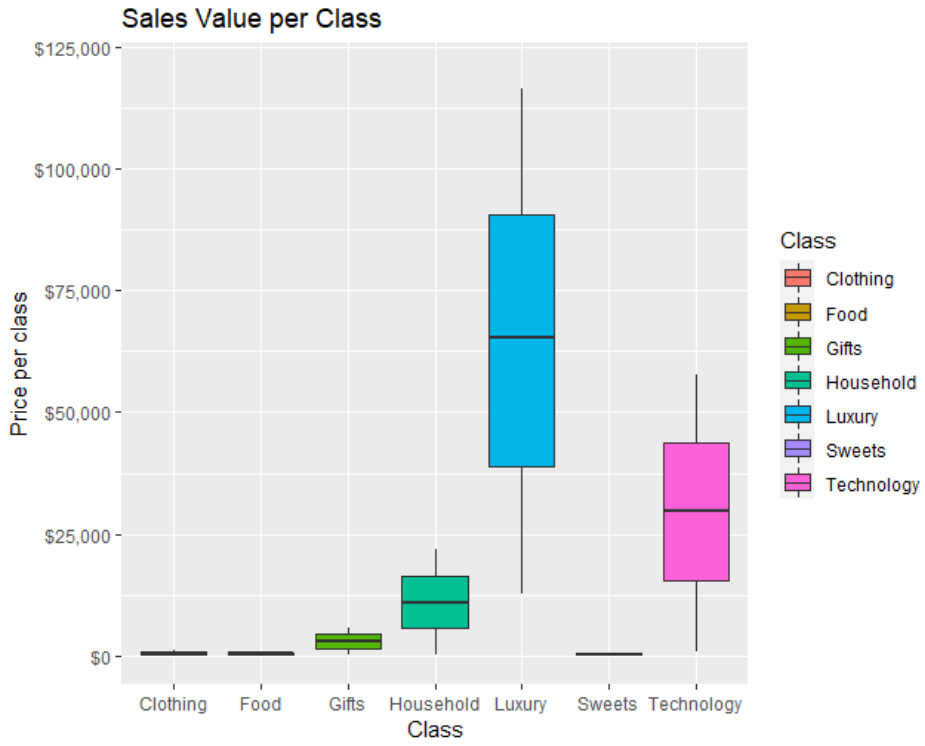


Figure 4: Sales Value per Class Boxplot

From these box plots the comparison can be made for all of the classes, and it is clear that each class only have sales in a specific price range.

Delivery Time

Minimum	1st Quarter	Median	Mean	3rd Quarter	Maximum
0.5	3.0	10.0	14.5	18.5	75.0

Table 3: Delivery Time Statistics

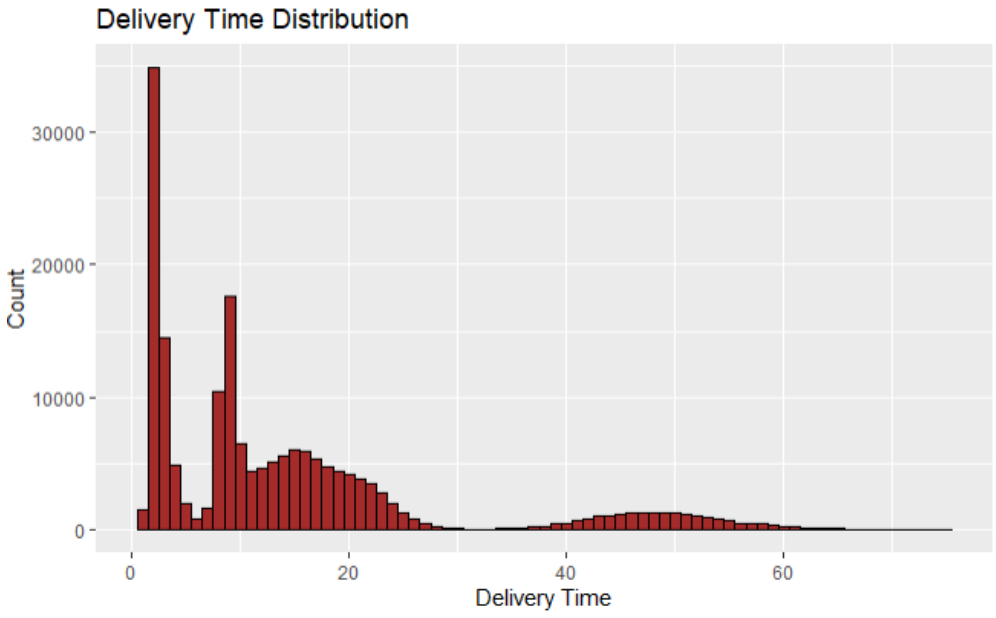


Figure 5: Delivery Time Distribution Plot

It can be seen from the delivery time distribution that delivery time follows a multimodal distribution, with peaks at roundabout 2, 8, 15 and 48 hours.

Process Capability Indices for Delivery Time

Assumptions:

- USL = 24 hrs
- LSL = 0

An LSL of 0 is logical as it is possible for a customer to take immediate delivery of a product that they have bought. This would be considered an over-the-counter sale of a product, where the client is immediately handed their product after payment.

First, in order to calculate C_p, C_{pu}, C_{pl}, and C_{pk} the standard deviation and mean of delivery time has to be calculated. The standard deviation of the delivery time is calculated to be 3.50 hours.
Now all of C_p, C_{pu}, C_{pl} and C_{pk} can be calculated using their respective statistical formulas.

```
Cp = ((USL-LSL)/(6*stdev))
Cpu =(USL-mean(Ts$Delivery.time))/(3*stdev)
Cpl =(mean(Ts$Delivery.time)-LSL)/(3*stdev)
Cpk =min(Cpu,Cpl)
```

Figure 6: Process Capability Formulas

The following values were the result of the calculations:

C_p = 1.142
C_{pu} = 0.380
C_{pl} = 1.905
C_{pk} = 0.380

As C_{pl} and C_{pu} values grow larger, they indicate that the process is off centre. All of the Process Capability Indices calculated have very low values showing that the Delivery Times are good as is.

Part 3: Statistical Process Control

First of all, the dataset with only valid entries has to be ordered by means of ascending date as well as having an ascending ID for each date. The data is then grouped by their respective class. In order to be able to build the X-chart and s-chart all of the control limits first have to be calculated for each of the sales classes. This done by using the known statistical formulas for each of the thirty samples consisting of 15 sales each.

X-chart

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	9.390601	9.250401	9.110200	8.970000	8.829800	8.689599	8.549399
Food	2.702226	2.631484	2.560742	2.490000	2.419258	2.348516	2.277774
Gifts	9.451412	9.087978	8.724545	8.361111	7.997678	7.634244	7.270811
Household	50.126859	48.938647	47.750435	46.562222	45.374010	44.185798	42.997585
Luxury	5.468973	5.224501	4.980028	4.735556	4.491083	4.246610	4.002138
Sweets	2.883225	2.748076	2.612927	2.477778	2.342629	2.207479	2.072330
Technology	22.888932	22.050770	21.212607	20.374444	19.536282	18.698119	17.859957

Figure 7: X-Chart

S-Chart

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	0.8665596	0.7614552	0.6563509	0.5512465	0.4461422	0.3410379	0.2359335
Food	0.4372466	0.3842133	0.3311800	0.2781467	0.2251134	0.1720801	0.1190468
Gifts	2.2463333	1.9738773	1.7014213	1.4289652	1.1565092	0.8840532	0.6115971
Household	7.3441801	6.4534101	5.5626402	4.6718703	3.7811003	2.8903304	1.9995605
Luxury	1.5110518	1.3277775	1.1445032	0.9612289	0.7779546	0.5946803	0.4114060
Sweets	0.8353391	0.7340215	0.6327039	0.5313862	0.4300686	0.3287509	0.2274333
Technology	5.1805697	4.5522224	3.9238751	3.2955278	2.6671805	2.0388332	1.4104859

Figure 8: S-Chart

From both the X-chart and s-chart values shown above it is clear that the delivery times for all classes except for “Household” and “Technology”. These two classes have very large control limits compared to the other five classes.

First 30 Samples

Along with the Statistical Control Charts the first thirty samples is also plotted for viewing. The solid red lines represent the upper and lower control limits, and the green dotted lines represent the U1Sigma, U2Sigma, L1Sigma and L2Sigma values for each of the respective product classes.

It is important to also consider that any sample which lies closer to the LCL should be good, as a shorter delivery time is desired by customers most of the time.

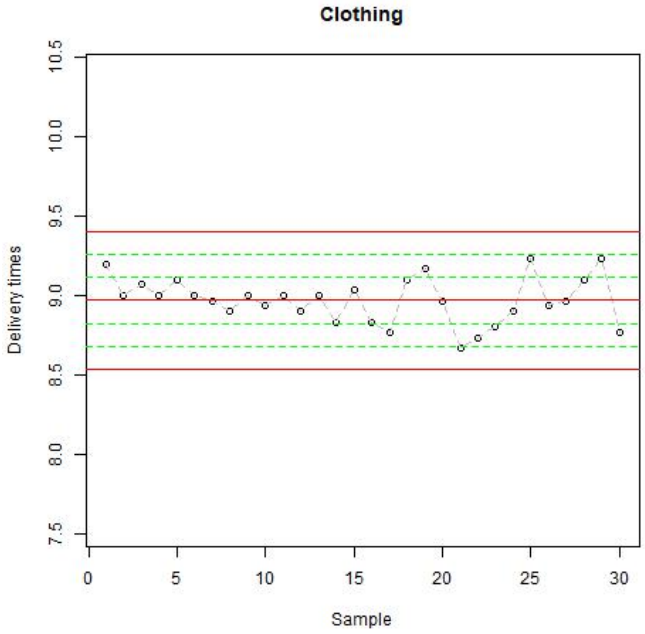


Figure 9: Clothing Sample Control Chart

The control chart for Clothing shows that most of the sample data lie between U2Sigma and L2Sigma control limits. There is however one sample getting closer to the LCL.

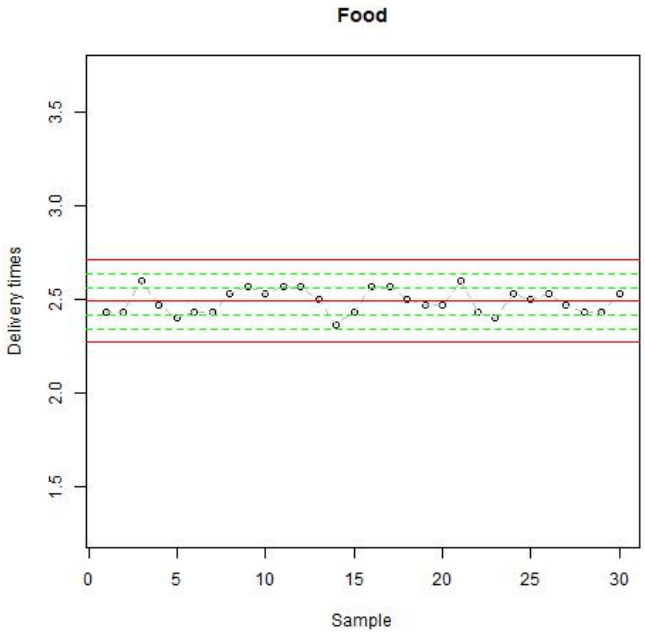


Figure 10: Food Sample Control Chart

The control chart for Food shows that all of the samples lie in between the U2Sigma and L2Sigma control limit. This can be explained by food being perishable, and some food require refrigeration to be fresh upon delivery.

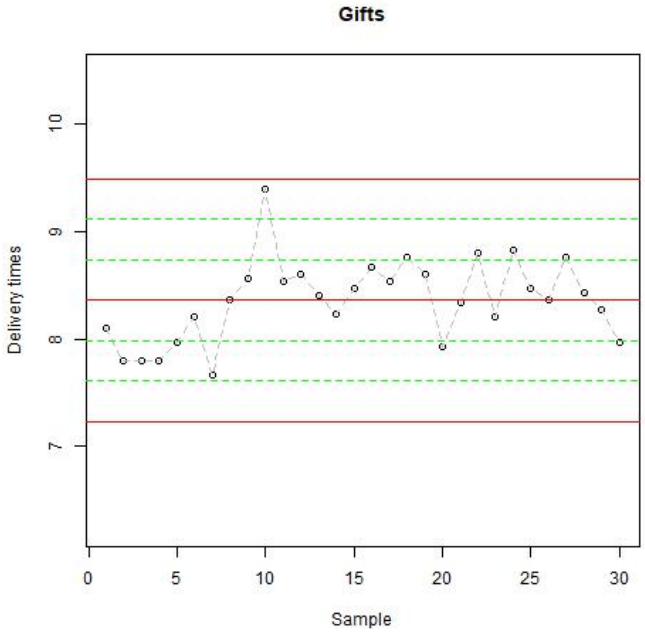


Figure 11: Gifts Sample Control Chart

The control chart for Gifts shows that the samples for Gifts all lie in between the U2Sigma and L2Sigma control limits with exception of only one sample being between the UCL and U2Sigma.

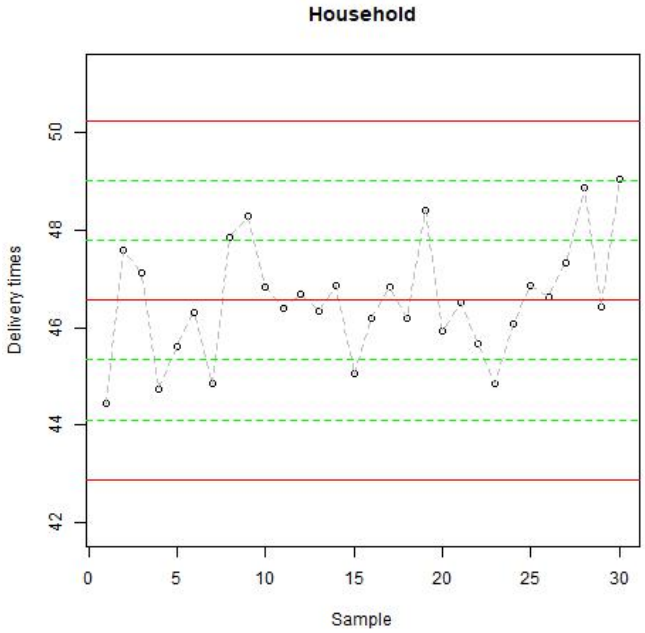


Figure 12: Household Sample Control Chart

Looking at the control chart for Household items it can be seen that the CL for Household items take a long time to be delivered. Delivery times are around 46 ours with about 2 hours of play to the upper and lower boundaries. There is also a wide variation of 7 hours between the lower and upper boundaries.

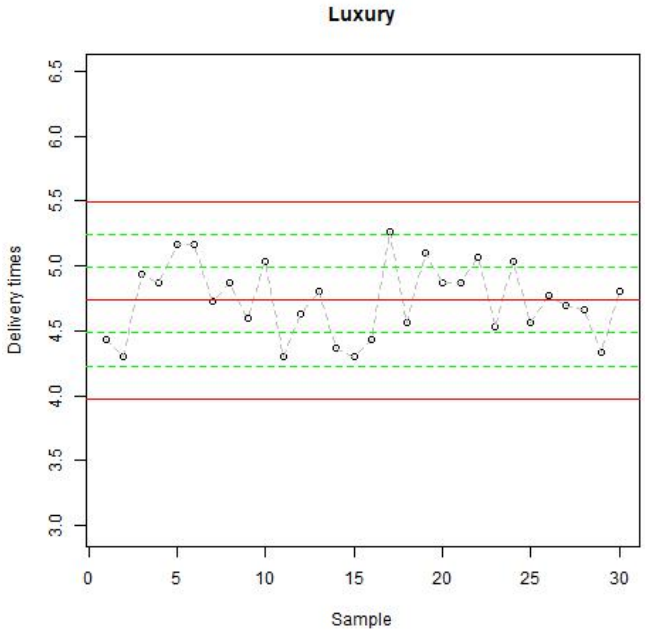


Figure 13: Luxury Sample Control Chart

From the control chart for Luxury items, it can be seen that Luxury items are being delivered within 4 – 5.5 hours. Most of the samples lie in between the U2Sigma and L2Sigma control limits with one sample closer to the UCL and one sample being closer to the LCL

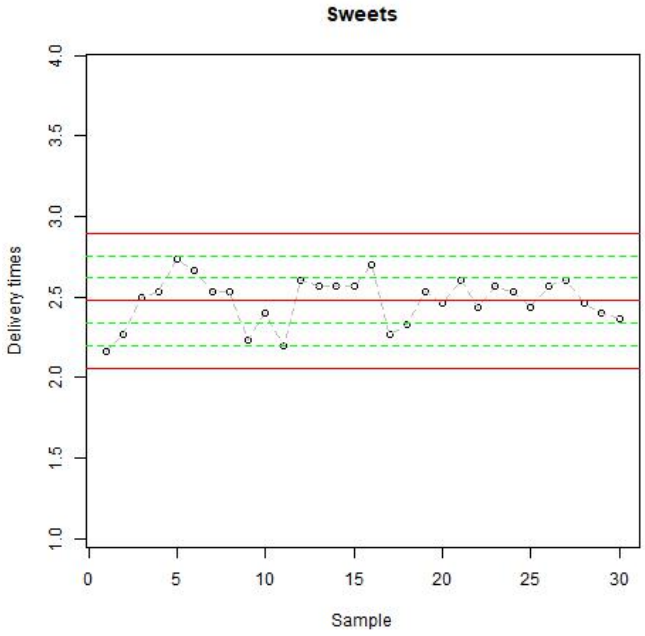


Figure 14: Sweets Sample Control Chart

It can be seen from the Sweets control chart that sweets are being delivered within 2 – 3 hours. Most of the samples lie in between the U2Sigma and L2Sigma control limits with exception of two samples closer to the LCL.

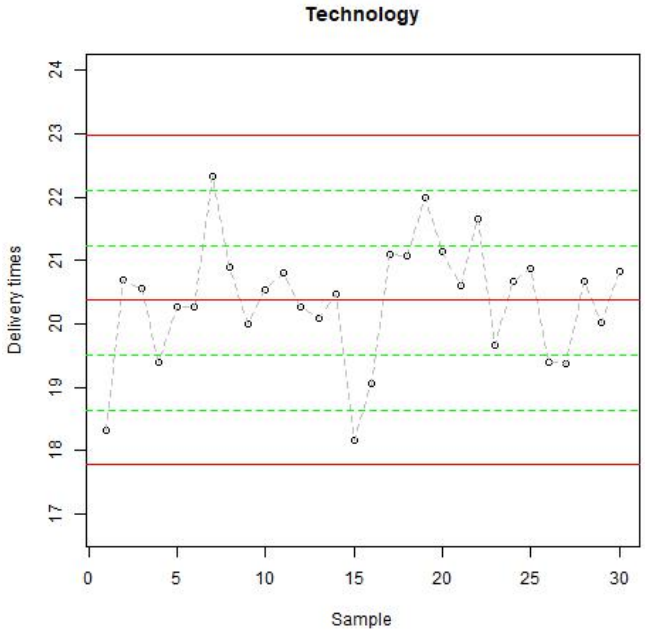
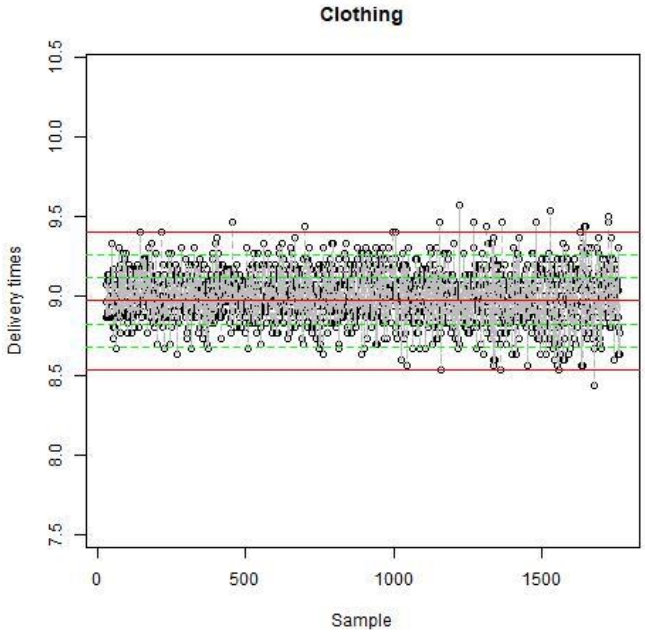


Figure 15: Technology Sample Control Chart

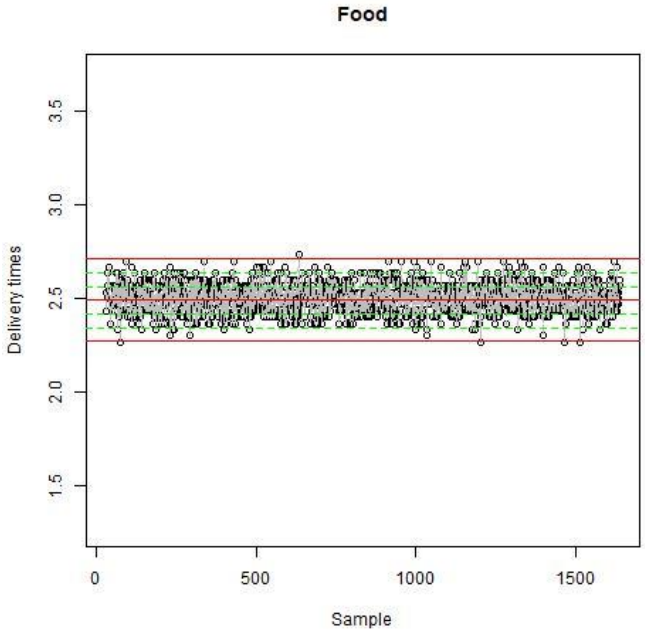
From the Technology control chart, it can be seen that the delivery times lie in between 18 - 23 hours. Most of the samples lie in between the U2Sigma and L2Sigma control limits with exception of two samples closer to the LCL and one sample closer to the UCL

Rest of Valid Data Control Chart



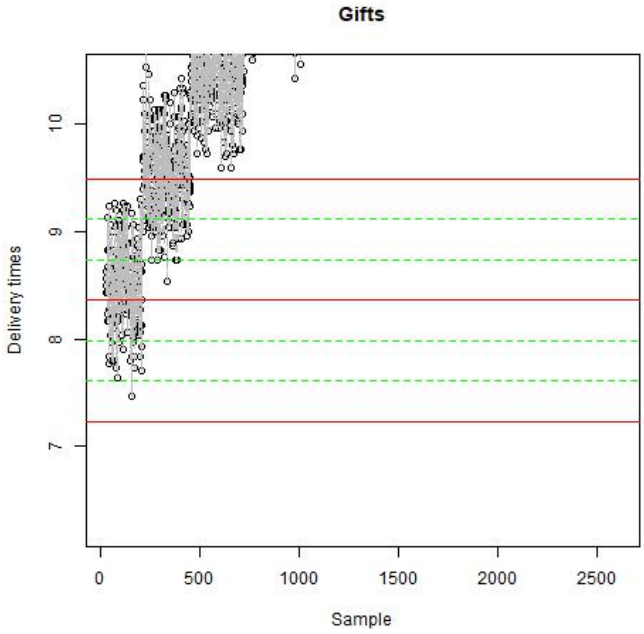
Looking at the rest of the samples for Clothing it can be seen that most of the samples lie in-between the UCL and LCL with only a few outside of these boundaries. Most of the out of boundary samples are above the UCL.

Figure 16: Clothing Control Chart



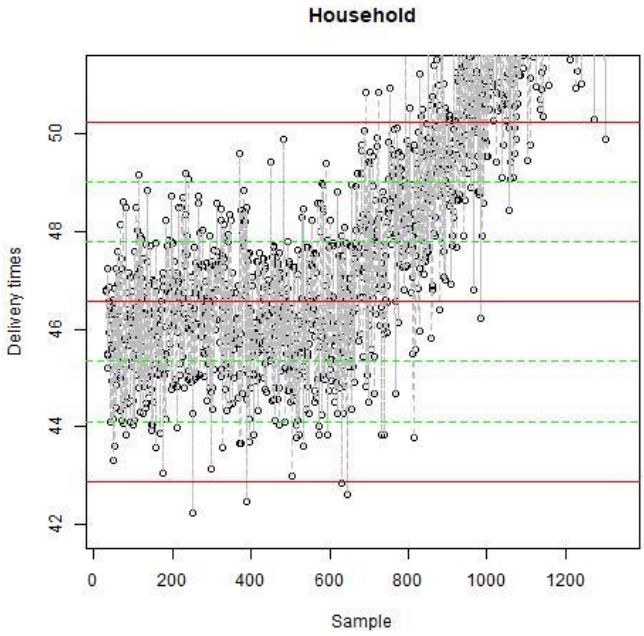
Of the rest of the samples for Food only 5 samples are out of boundaries. Most of the out of boundaries samples are below the LCL, indicating a shorter delivery time.

Figure 17: Food Control Chart



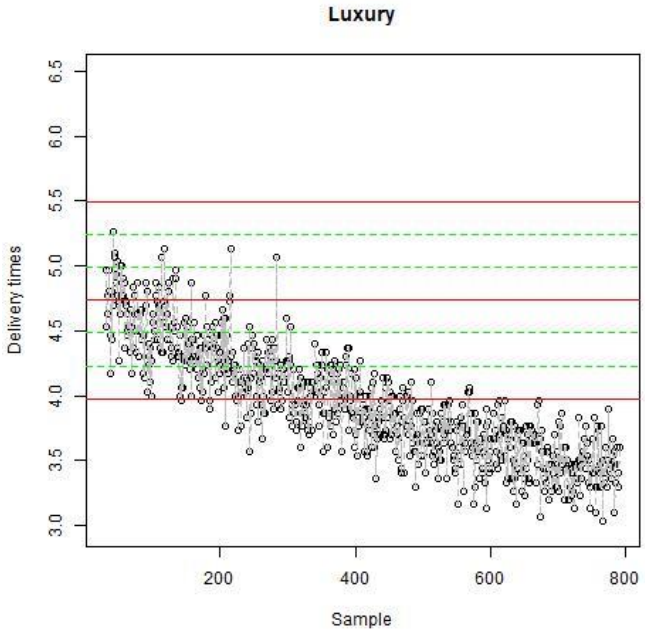
Gifts have a very high number of samples that lie above the UCL. This indicates that there is a big problem with the delivery of gifts.

Figure 18: Gifts Control Chart



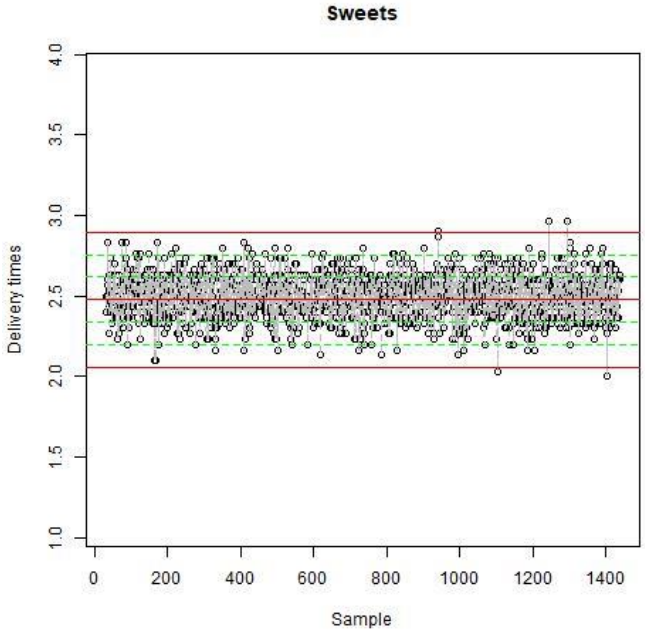
The largest portion of the household samples lie in between the control limits. However, there is also a many samples that lie above the UCL. This could also indicate a problem with delivery of household items.

Figure 19: Household Control Chart



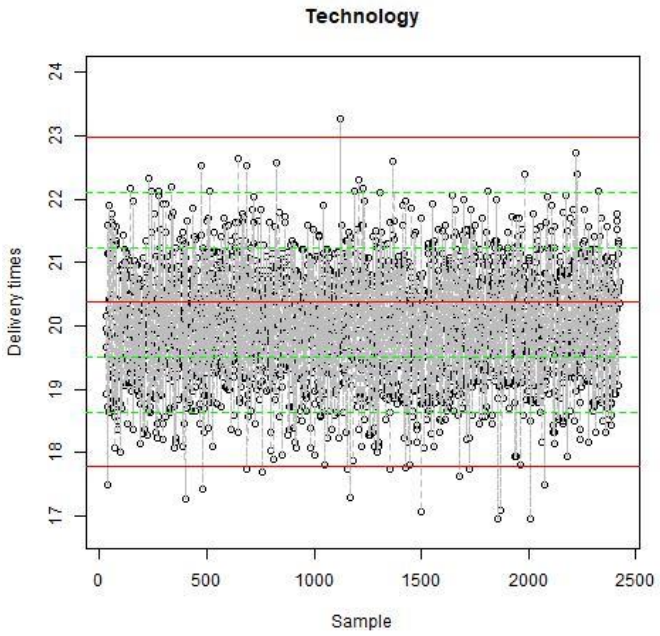
All of the samples for Luxury lie under the UCL. This is a very good observation which is also supported by almost half of the samples being below the LCL. This indicates that many of the samples are being delivered much faster than the lower control limit.

Figure 20: Luxury Control Chart



Almost all of the Sweets samples lie in between the UCL and LCL with only exception to 5 samples, of which 3 are slightly above the UCL and 2 samples being below the LCL

Figure 21: Sweets Control Chart



Although Technology has higher control limits comparing to the rest of the classes, all the samples are below the UCL with exception to only one sample being slightly above the UCL.

Figure 22: Technology Control Chart

Part 4: Optimising the Delivery Processes

Sample Means outside of Outer Control Limits

Class	1st	2nd	3rd	3rd last	2nd last	Last	TOTAL
Clothing	9.400000	9.400000	9.466667	8.433333	9.466667	9.500000	22
Food	2.266667	2.733333	2.266667				5
Gifts	10.233333	9.666667	9.700000	16.56667	16.30000	16.03333	2296
Household	42.233333	42.46667	42.83333	57.36667	54.56667	55.80000	410
Luxury	4.000000	4.000000	3.966667	3.400000	3.300000	3.600000	455
Sweets	2.900000	2.033333	2.966667				5
Technology	17.50000	17.26667	17.43333	17.80000	16.96667	17.50000	20

Table 4: Sample Means outside of Outer Control Limits

Length and index of samples within Control Limits

Class	Most Consecutive Samples	Ending Sample Number
Clothing	406	529
Food	125	783
Gifts	225	644
Household	71	568
Luxury	497	791
Sweets	636	787
Technology	224	1171

Table 5: Length and Index of Samples

4.2 Probability of making a Type I error`

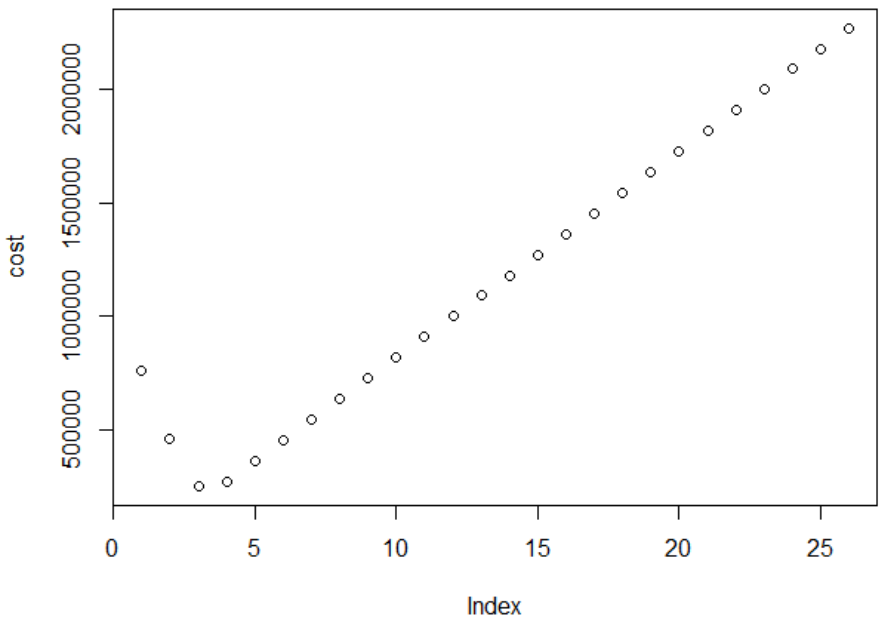
The probability of making a Type I error is to be calculated using the R function “pnorm” as it will use a normal distribution to calculate the value. The calculation is done using a value of 3, as it represents the upper control limit, and then the value is multiplied by two due to the symmetry of a normal distribution.

```
> typ1 <- (1 - pnorm(3))*2
> pot1 <- typ1*100
> pot1
[1] 0.2699796
```

Figure 23: Type I Error Calculation

This indicates that the chance of making a Type I error is 0.2699796 %.

4.3 Delivery time centring



In order to have the lowest cost it is calculated to centre the delivery around 3 hours. This will incur a cost of R250 002.50.

Part 5: MANOVA

When analysing the previous parts done it was decided upon to have the hypothesis as follows:

Ho: The “Class” feature of a sale has no impact on the features “Price” and “Delivery Time”.

Ha: The “Class” feature has an impact on the features “Price” and “Delivery time”.

```
> summary(manova1)
             Df Pillai approx F num Df den Df    Pr(>F)
Class         6 1.6797   157291     12 359942 < 2.2e-16 ***
Residuals 179971
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 24: MANOVA

```
> summary.aov(manova1)
 Response Delivery.time :
              Df   Sum Sq Mean Sq F value    Pr(>F)
Class          6 33458565 5576427  629429 < 2.2e-16 ***
Residuals   179971  1594452       9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response Price :
              Df     Sum Sq    Mean Sq F value    Pr(>F)
Class          6 5.7168e+13 9.5281e+12   80258 < 2.2e-16 ***
Residuals   179971 2.1366e+13 1.1872e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 25: MANOVA Summary

Assuming that Class has no impact on the Delivery Time and Price, we expect the Price and Delivery Time per class to be random, and not have any tendencies. Comparing this expectation to the real findings in the two plots below, we find that it is not the case. Therefore, the Null Hypothesis is rejected due to the significant impact class has on the delivery time and price of a product.

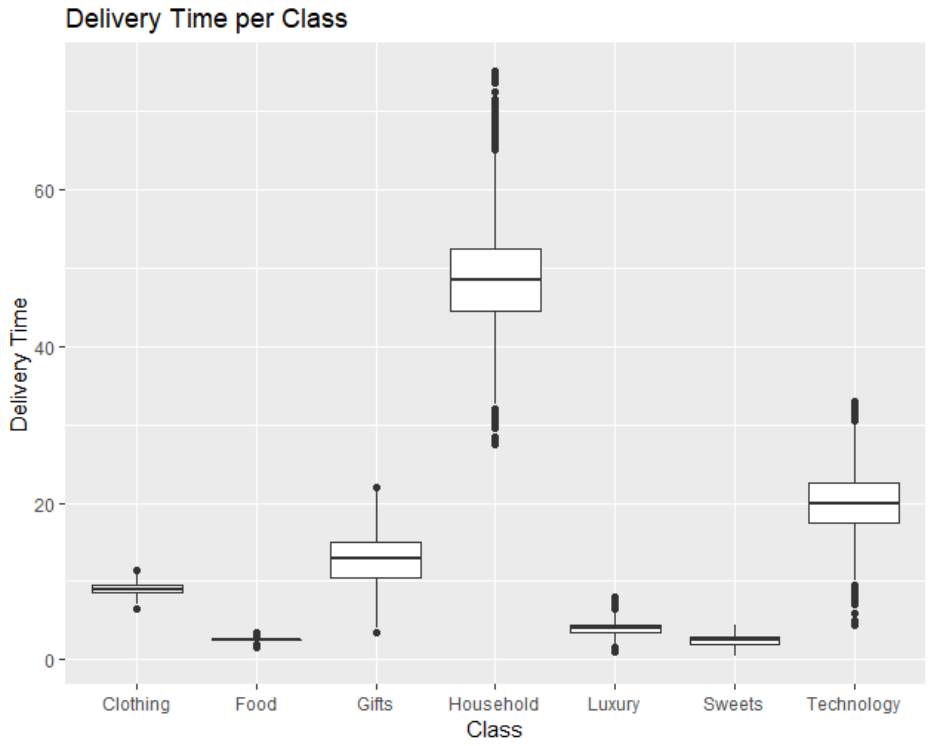


Figure 26: Delivery Time per Class

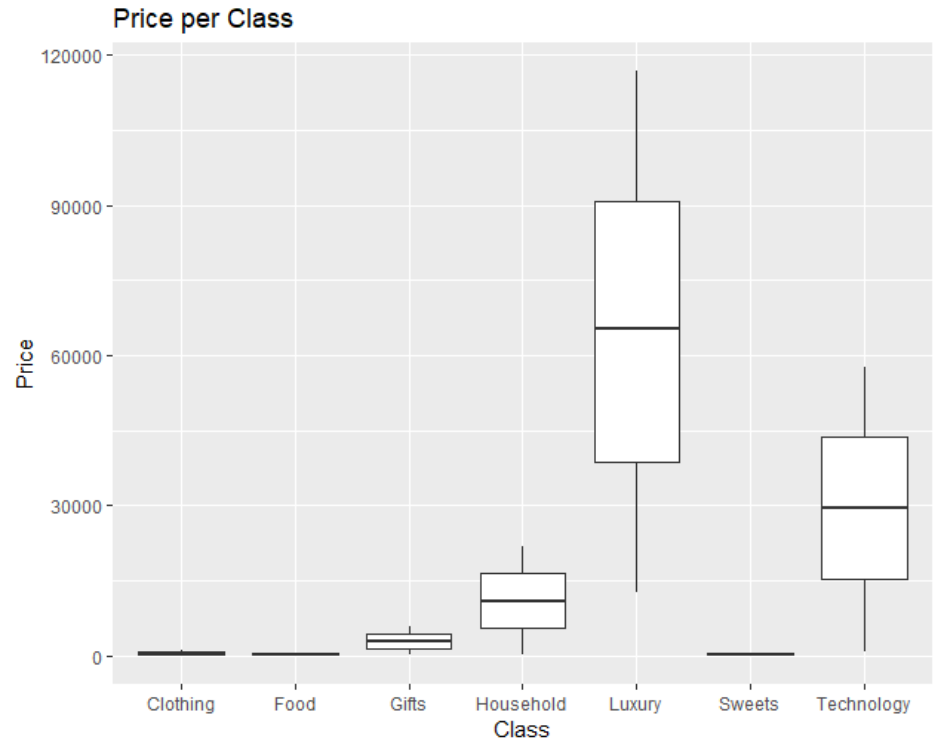


Figure 27: Price per Class

Part 6: Reliability of the Service and Products

6.1

Question 6

We know the following about the Loss Function.

- L = $45 per unit
- (y-m) = 0.04

$L = k \times (y - m)^2$

By substituting the known variables into the equation, the k value can then be calculated:

$$L = k \times (y - m)^2$$

$$45 = k \times (0.04)^2$$

$$k = \frac{45}{(0.04)^2}$$

$$k = 28125$$

Therefore, the Taguchi Loss Function is as follows:

$L = 28125 \times (y - m)^2$

Question 7a

We know the following about the Loss Function.

- L = $35 per unit
- (y-m) = 0.04

$L = k \times (y - m)^2$

By substituting the known variables into the equation, the k value can then be calculated:

$$L = k \times (y - m)^2$$

$$35 = k \times (0.04)^2$$

$$k = \frac{35}{(0.04)^2}$$

$$k = 21875$$

Therefore, the Taguchi Loss Function is as follows:

$L = 21875 \times (y - m)^2$

Question 7b

$L = 21875 \times (y - m)^2$

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

$L = \$15.95\ per\ unit$

6.2

Question 27a

The system reliability with one machine at each stage can be calculated as follows.

$$System\ Reliability = (Reliability\ A) \times (Reliability\ B) \times (Reliability\ C)$$

$$System\ Reliability = 0.85 \times 0.92 \times 0.90$$

$$System\ Reliability = 0.7038$$

Therefore, the system with only one machine at each stage is 70.38% reliable.

Question 27b

The reliability for each of the sets of two machines in parallel first have to be calculated.

$$Reliability\ A = 1 - (1 - Reliability\ of\ Machine\ A)^2$$

$$Reliability\ A = 1 - (1 - 0.85)^2$$

$$Reliability\ A = 0.9775$$

$$Reliability\ B = 1 - (1 - Reliability\ of\ Machine\ B)^2$$

$$Reliability\ B = 1 - (1 - 0.92)^2$$

$$Reliability\ B = 0.9936$$

$$Reliability\ C = 1 - (1 - Reliability\ of\ Machine\ C)^2$$

$$Reliability\ C = 1 - (1 - 0.90)^2$$

$$Reliability\ C = 0.99$$

The System Reliability can therefore be calculated as follows:

$$System\ Reliability = (Reliability\ A) \times (Reliability\ B) \times (Reliability\ C)$$

$$System\ Reliability = 0.9775 \times 0.9936 \times 0.9900$$

$$System\ Reliability = 0.9615$$

The System with two machines in parallel at each of the stages is 96.15% reliable.

This is a 25.77% reliability increase from the system with only one machine in each of the stages.

6.3

First the probabilities for a failure for either the vehicles or the employees was calculated. Using this the system was calculated to be reliable for 237 out of 365 days per year. With the addition of only one vehicle, resulting in a total of 21 vehicles, the system was calculated to be reliable for 328 out of the 365 days per year.

Conclusion

The company is currently in a relatively okay state with its delivery system. Small improvements could be made to further improve the state of the company. The implementation of various control charts could provide excellent feedback on how the company is doing and sow possible room for improvement. There are also various ways of increasing value for the company that should be explored.

References

Anonamous. (2015). *MANOVA*. Herwin van cran.r-project: https://cran.r-project.org/web/packages/MANOVA.RM/vignettes/Introduction_to_MANOVA.RM.html

[Accessed 15/10/2022]

Hernandez, F. (2015). *Data Analysis with R - Exercises*. Herwin van Github: http://fch808.github.io/Data-Analysis-with-R-Exercises.html

[Accessed 04/10/2022]

Hessing, T. (2014). *X Bar R Control Charts*. Herwin van Six Sigma Study Guide: https://sixsigmastudyguide.com/x-bar-r-control-charts/

[Accessed 12/10/2022]

Hessing, T. (2014). *X Bar S Control Chart*. Herwin van Six Sigma Study Guide: https://sixsigmastudyguide.com/wp-content/uploads/2014/06/XbarS_chart_def.png

[Accessed 13/10/2022]