ECSA Project

Data Analysis & Manipulation

Prepared for: Quality Assurance 344

Prepared by: RTC Laing

(Student number - 23946237)

Date: 7 October 2022

Abstract

This report is for the Engineering Counsel of South Africa (ECSA) project produced by ECSA to be completed by all students studying industrial engineering in the country. The project is additionally used for assessment in the Quality Assurance 344 module at Stellenbosch University.

This project requires one to sufficiently perform data analysis and manipulation for ECSA's data analysis and manipulation outcome. In order to do so, R and RStudio was used to perform data wrangling and statistical analysis on given sales data. Statistical process control is also used and with the use of various graphs and tables a conclusion is reached about the data and problems solved.

Table of Contents

Abstract ii
Table of Figures iv
Table of Tables v
Introduction 1
Part One: Data Wrangling 1
Part Two: Descriptive Statistics 2
Five-point summary: 2
Graphical relationships: 3
Process capacity indices: 6
Part Three: Statistical Process Control 7
Six-Sigma values for s-chart 7
Six-Sigma values for x-bar chart 7
Part Four: Optimising the delivery processes 8
Data in statistical control 8
Unstable Data 9
Type I (Manufacturer's) Error 10
Best delivery time 10
Type II (Consumer's) Error 10
Part Five: DOE and MANOVA 11
Part Six: Reliability of the service and products 11
Problem 6 11
Problem 7 12
Problem 27 13
Delivery Process Vehicles Question 14
Question A 14
Question B 14
Conclusion 15
References 16

Table of Figures

Figure 1: The invalid data from the given set 1
Figure 2: The valid data from the given set 2
Figure 3: Box plot depicting the age of people making purchases of different items 3
Figure 4: Jitter plot depicting the delivery times for each item sold, sorted by type 4
Figure 5: Cumulative bar plot depicting amount of money spent per year, sorted by type 4
Figure 6: Jitter plot showing each items price range 5
Figure 7: Multiple bar graph depicting the average monthly amounts spent on each item 5
Figure 8: Cumulative bar plot depicting the different reasons for purchase's number of sales 6
Figure 9 : Technology graph 8
Figure 10: Clothing graph 8
Figure 11: Luxury graph 9
Figure 12: Gifts graph 9
Figure 13: Graph indicating the recommended delivery time to reduce costs for technology items . 10
Figure 14: The Taguchi loss function for the Cool Food, Inc textbook problem 6 12
Figure 15: The Taguchi loss function for the Cool Food, Inc textbook problem 6 13
Figure 16: Production system diagram from the textbook problem 13

Table of Tables

Table 1: Five-point summary of age 2
Table 2: Five-point summary of price 2
Table 3: Five-point summary of delivery time 3
Table 4: Six-Sigma values used for graphs in part four 7
Table 5: Type I error likelihood for A and B 10
Table 6: Response found for age and price 11
Table 7: Response found for delivery time and price 11
Table 8: Table of vehicle reliability for 0 or 1 failures 14
Table 9: Table of driver reliability for 0 or 1 failures 14

Introduction

The received client data found in the SalesTable2022 csv file contains various different variables and values. These are namely the client ID, age, class, price, year, month, day, delivery and reason why brought for each purchase made. Using these variables allows one to explore the trends and abnormalities of this company's clients and perform some statistical analysis.

The Taguchi loss function is also explored through some example questions from the prescribed textbook and the DOE and MANOVA tests are performed on select variables from the given data.

Part One: Data Wrangling

Once the SalesTable2022 data had been received and opened in R, after a summary function was performed it was clear that it needed to be refined slightly. It was separated into valid and invalid data based on NA values and negative numbers found. With those rows removed, what remained was the valid data. The invalid data is shown below:

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

Figure 1: The invalid data from the given set

Any row with a single NA value was replaced with a full NA column, while the instances with a negative price are shown in full. There were 22 invalid cases in total.

The remaining 179978 valid entries were kept separate in a different table. Both of these data frames were given an additional index as well. The second dataset for the valid data is shown below:

	ValidIndex	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	1	1	19966	54	Sweets	246.21	2021	7	3	1.5	Recommended
2	2	2	34006	36	Household	1708.21	2026	4	1	58.5	Website
3	3	3	62566	41	Gifts	4050.53	2027	8	10	15.5	Recommended
4	4	4	70731	48	Technology	41843.21	2029	10	22	27.0	Recommended
5	5	5	92178	76	Household	19215.01	2027	11	26	61.5	Recommended
6	6	6	50586	78	Gifts	4929.82	2027	4	24	14.5	Random
7	7	7	73419	35	Luxury	108953.53	2029	11	13	4.0	Recommended
8	8	8	32624	58	Sweets	389.62	2025	7	2	2.0	Recommended
9	9	9	51401	82	Gifts	3312.11	2025	12	18	12.0	Recommended
10	10	10	96430	24	Sweets	176.52	2027	11	4	3.0	Recommended
11	11	11	87530	33	Technology	8515.63	2026	7	15	21.0	Browsing
12	12	12	14607	64	Gifts	3538.66	2026	5	13	13.5	Recommended
13	13	13	24299	52	Technology	27641.97	2024	5	29	17.0	Browsing
14	14	14	77795	92	Food	556.83	2025	6	3	3.0	Random
15	15	15	62567	73	Clothing	347.99	2024	3	29	8.5	Website
16	16	16	14839	47	Technology	54650.41	2027	12	30	18.5	Recommended
17	17	17	96208	44	Technology	14739.09	2028	3	17	13.0	Recommended
18	18	18	39674	69	Technology	22315.17	2026	8	20	20.5	Recommended
19	19	19	98694	74	Sweets	546.48	2025	5	9	2.0	Recommended
20	20	20	99187	54	Luxury	81620.21	2027	9	14	3.0	Recommended
21	21	21	59365	72	Gifts	3314.76	2028	4	30	13.0	Recommended
22	22	22	37221	24	Sweets	220.91	2021	3	8	3.0	Recommended

Figure 2: The valid data from the given set

Part Two: Descriptive Statistics

Five-point summary:

Looking at the valid dataset only, using the summary() function the following five-point summaries were found for various categories:

Table 1: Five-point summary of age

AGE	
Minimum	18 years
1st Quartile	38 years
Median	53 years
Mean	54.57 years
3rd Quartile	70 years
Maximum	108 years

Table 2: Five-point summary of price

PRICE	
Minimum	R35.65
1st Quartile	R400
Median	R2259.63
Mean	R12 294.10
3rd Quartile	R15 270.97
Maximum	R116 618.97

Table 3: Five-point summary of delivery time

DELIVERY TIME	
Minimum	0.5 days
1st Quartile	3 days
Median	10 days
Mean	14.5 days
3rd Quartile	18.5 days
Maximum	75 days

Graphical relationships:

The following figure shows the different five-number summaries of the different classes of items.

This graph shows that the highest mean age of the customer is for buying food, which makes sense as it is the most essential item for living. The lowest mean age of the customer is for technology, reflecting the younger generation's affinity for buying and utilising new technology every year.

Sweets had the highest inter-quartile range, showcasing its global appeal to everyone of all ages.

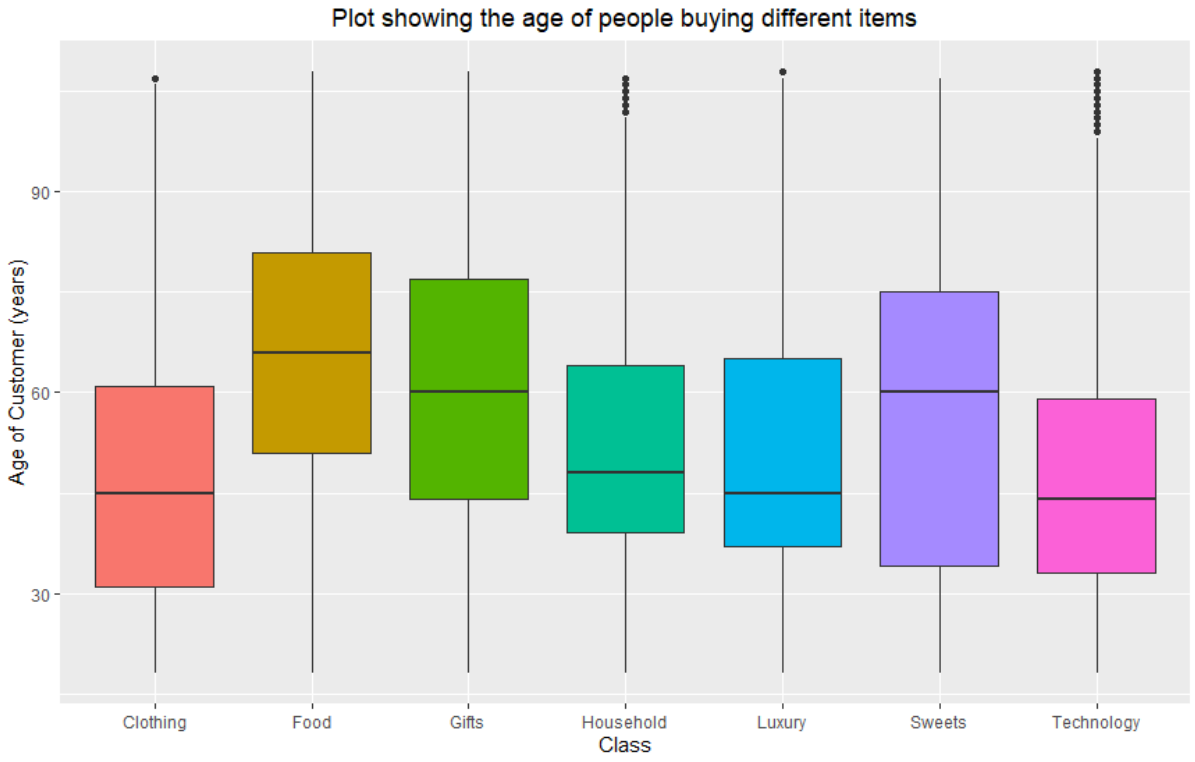

Figure 3: Box plot made in RStudio depicting the age of people making purchases of different items

The following figure shows the delivery times of the different classes of items.

This graph shows that household items not only had the biggest range of delivery times, but also had the highest delivery times by a clear margin. The shorter delivery times were for the two edible items, food and sweets. Technology also has a very wide range of delivery times.

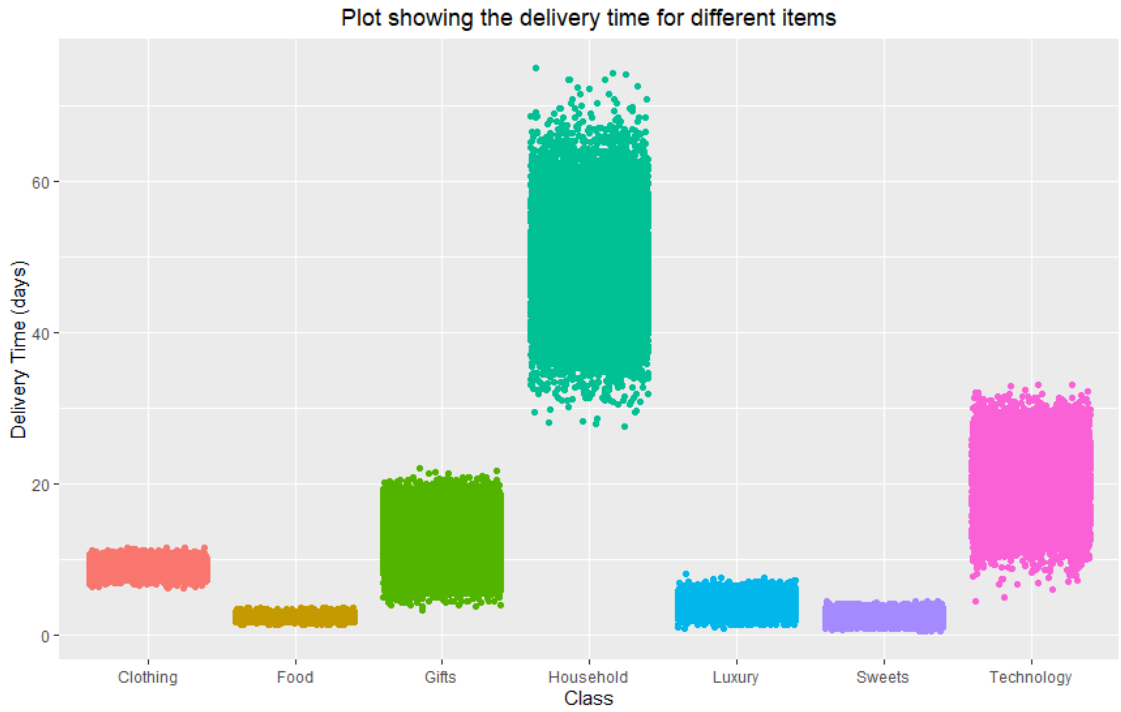

Figure 4: Jitter plot made in RStudio depicting the delivery times for each item sold, sorted by type

The following figure shows the prices of the different items and total cost spent over the years.

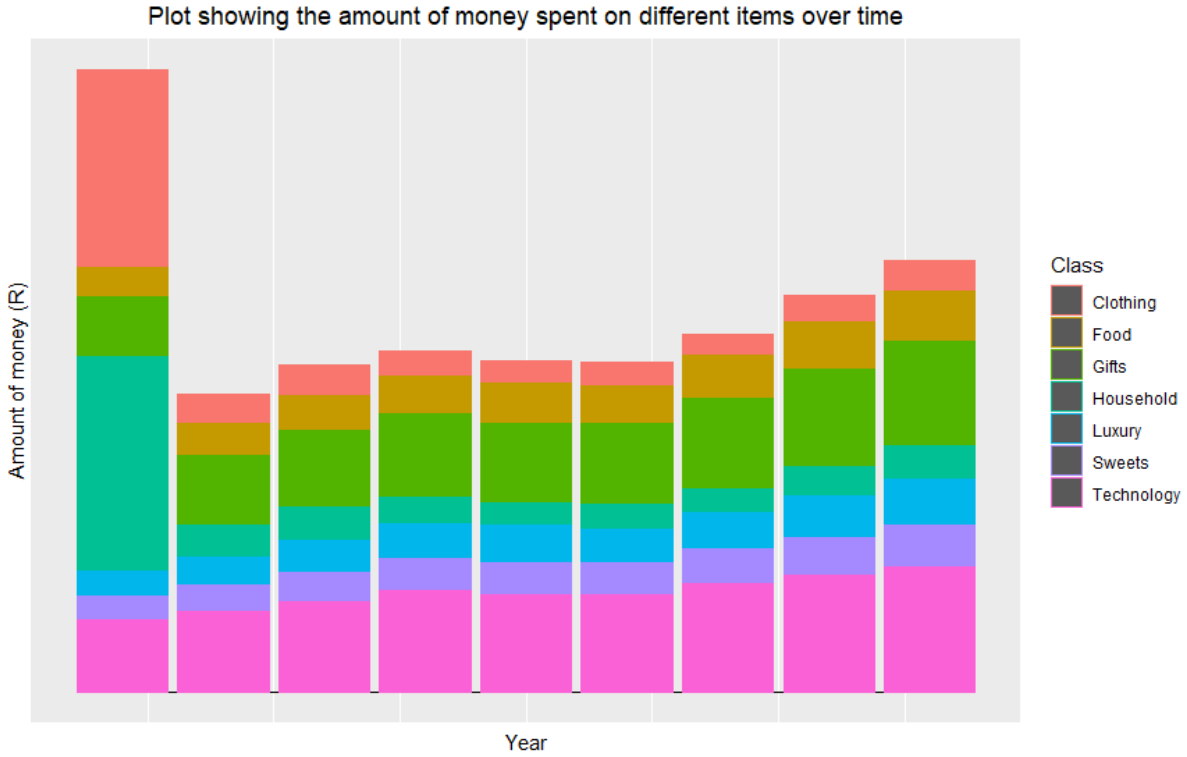

Figure 5: Cumulative bar plot made in RStudio depicting amount of money spent per year, sorted by type

The following figure shows the price ranges of the different items by class.

It shows that the amount of money spent on luxury goods has a very large range of prices, it also has the most expensive amount of all of them. Sweets had the lowest average price, this is an unsurprising result due to their cheap production and lack of outliers when it comes to pricing.

The clothing data in the given set is surprisingly low when it comes to cost, shown in other plots above and again here, with its small range.

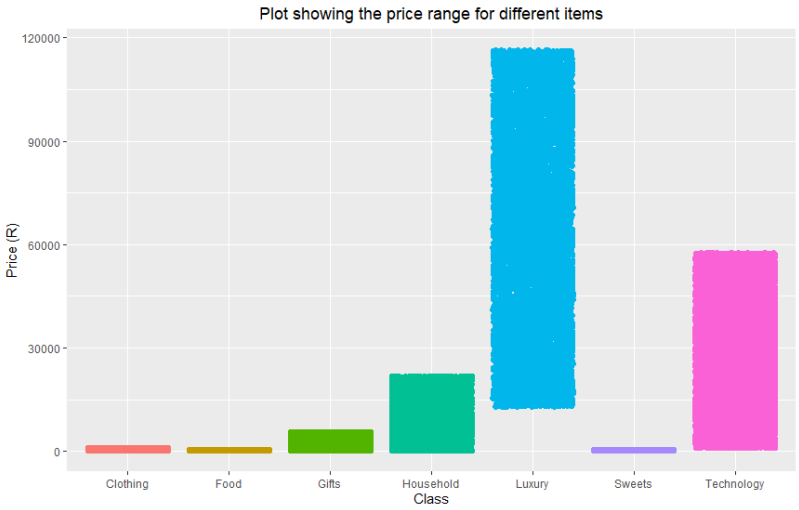

Figure 6: Jitter plot made in RStudio showing each items price range

The penultimate figure based on the sales data shows the average amount of purchases spent on the different classes of items.

It shows that luxury items have the lowest average sales per month due to their low regularity of sales and high cost, while technology and gifts have the highest as their demand is highest.

Figure 7: Multiple bar graph made in RStudio depicting the average monthly amounts spent on each item

The final figure based on the data given shows the number of purchases made for each reason and sorted further by class.

This shows that the most dominant factor in a person purchasing any of the different class items is based on a recommendation. This is a strong indicator at the power word of mouth has in the marketplace even today. It also shows that spam and e-mails are not big factors and the numbers for this are the lowest by a wide margin.

Website and browsing have similar total purchase numbers and class distribution which show how the two are very closely linked.

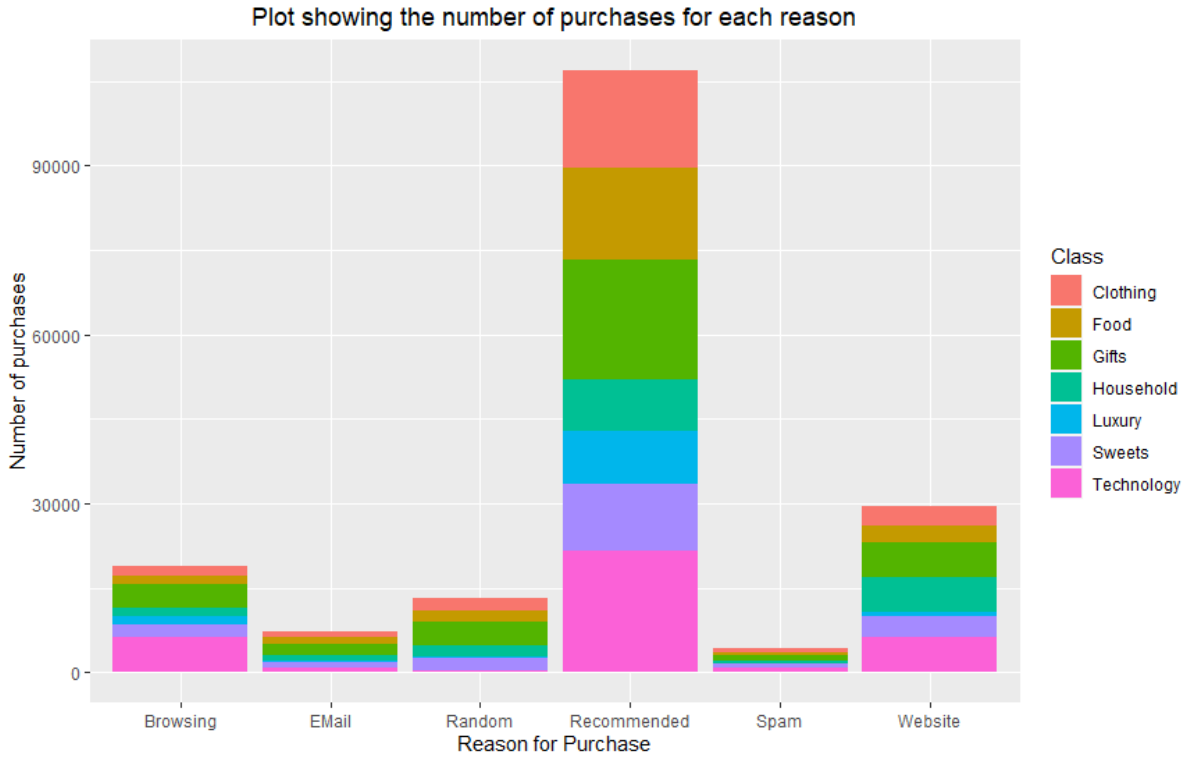

Figure 8: Cumulative bar plot made in RStudio depicting the different reasons for purchase's number of sales

Process capacity indices:

The four different process capacity indices were found in order to reflect a processes ability to produce within specified limits, the upper and lower limits of 24 and 0 were used in order to return the following values:

Cp = 1.142207

Cpu = 0.3796933

Cpl = 1.90472

Cpk = 0.3796933

Part Three: Statistical Process Control

Six-Sigma values for s-chart

To construct s-charts, the following points of data are used:

Table 4: Six-Sigma values used for graphs in part four

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	5.09202	4.95482	4.13713	0.73646	2.50175	1.68406	1.54686
Clothing	0.83613	0.81359	0.67933	0.08448	0.41079	0.27653	0.254
Household	7.18186	6.98834	5.83506	0.93158	3.52851	2.37523	2.18171
Luxury	1.47185	1.47185	1.19583	0.1918	0.72313	0.48678	0.44712
Food	0.42808	0.41655	0.3478	0.04761	0.21032	0.14158	0.13004
Gifts	2.2049	2.14549	1.79142	0.26237	1.08329	0.72922	0.66981
Sweets	0.81476	0.79281	0.66197	0.0833	0.4003	0.26946	0.24751

Six-Sigma values for x-bar chart

To construct the continued x-bar charts shown in part four, the following points of data were used:

Table 5: Six-Sigma values used for graphs in part four

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Technology	22.78013	21.96244	21.14475	20.32706	19.0937	18.69168	17.87399
Clothing	9.37104	9.23677	9.1025	8.96823	8.83397	8.6997	8.56543
Household	49.94807	48.79479	47.64151	46.48824	45.33496	44.18168	43.0284
Luxury	5.44906	5.21271	4.97635	4.74	4.50365	4.26729	4.03094
Food	2.69446	2.62572	2.55698	2.48824	2.41949	2.35075	2.28201
Gifts	9.44691	9.09284	8.73877	8.38471	8.03064	7.67657	7.3225
Sweets	2.87486	2.74403	2.61319	2.48235	2.35152	2.22068	2.08984

Part Four: Optimising the delivery processes

Data in statistical control

Using the values shown previously, all of the data was plotted and analysed.

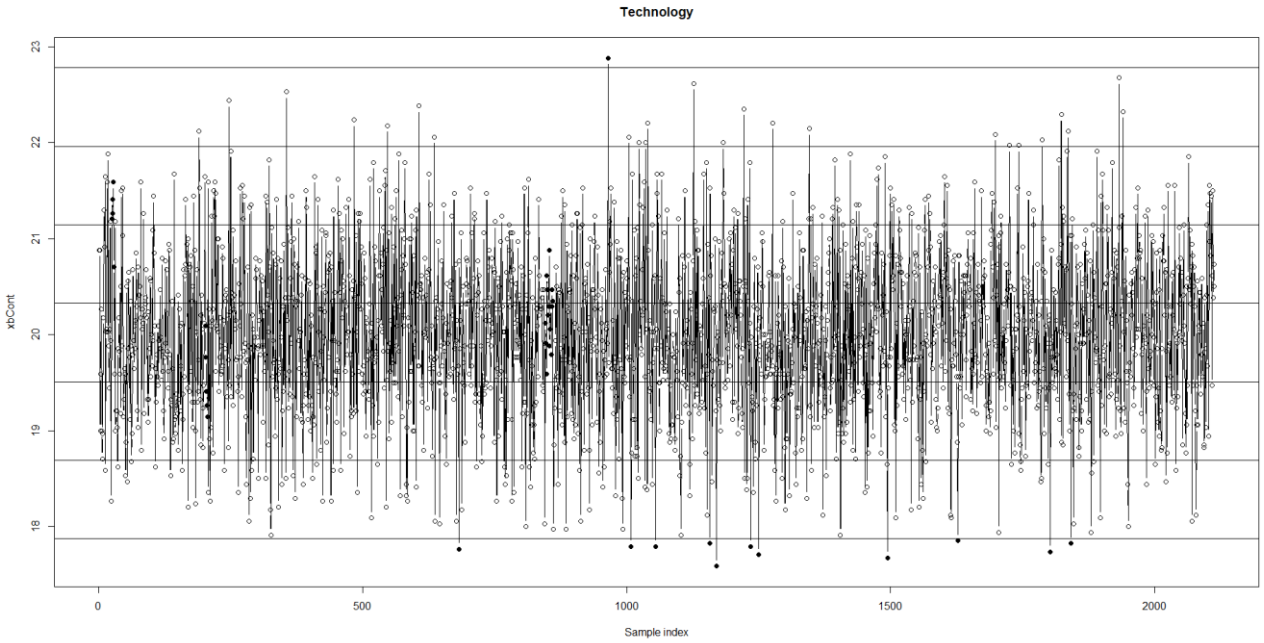

Figure 9 : Technology graph

With both the technology and clothing graphs shown, the data is in statistical control. This is shown as both graphs have only a handful of outliers outside of the six-sigma range, indicated with the black filled dots. In both figures we realise there are a similar number of points above and below the centre line, indicating both samples follow a normal distribution. Most of the points in both clothing and technology fall within the range showing a high level of control in the data given.

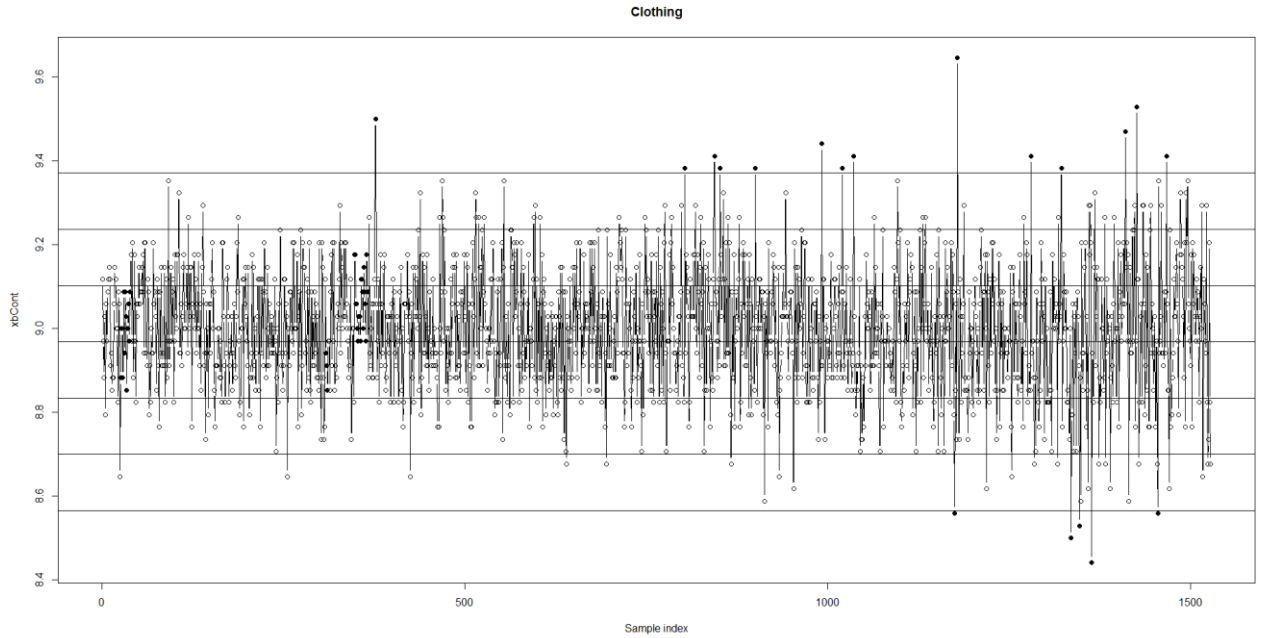

Figure 10: Clothing graph

Unstable Data

The following two plots showed the most distinct patterns of the seven classes.

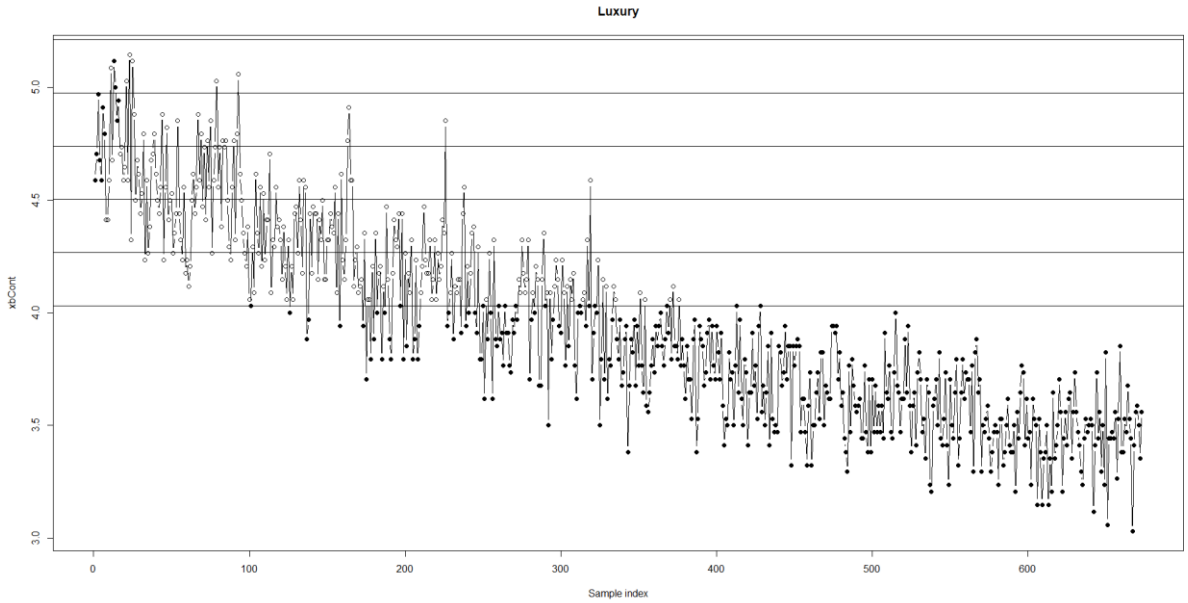

Figure 11: Luxury graph

The luxury chart is unstable and out of control and unstable due to a large amount of data points that lie below the -3 Sigma limit. A decreasing trend is seen as more samples are added, indicating that there is a shift in the mean for the average delivery time. This decreasing patten might be due luxuries being more expensive, where customers expect their purchase to arrive quickly

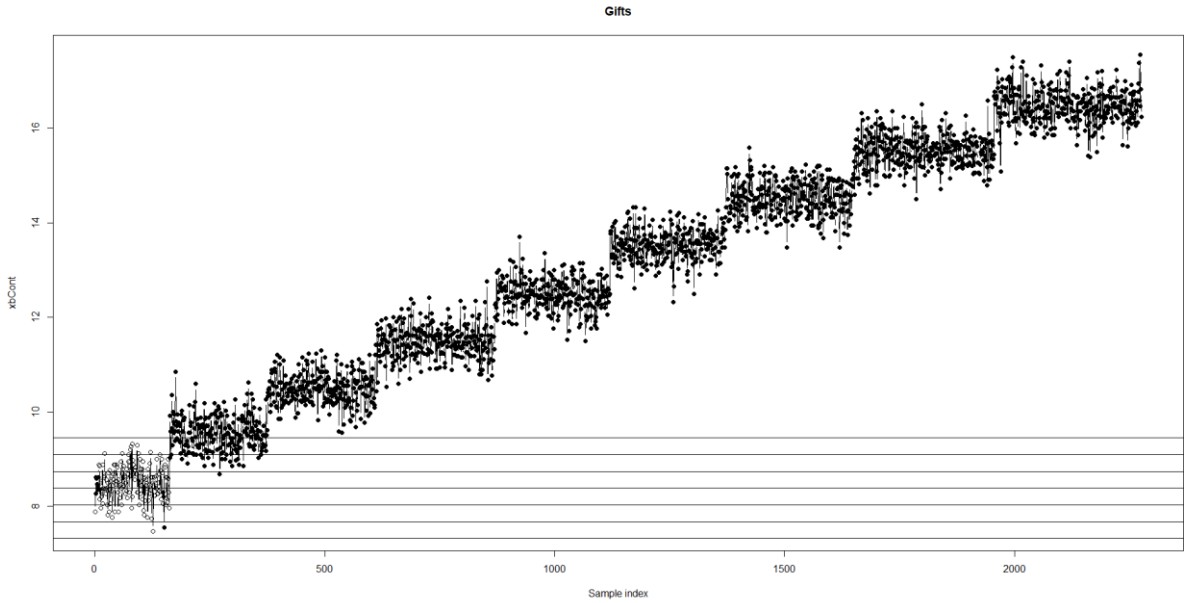

Figure 12: Gifts graph

The gifts class is also unstable, showing an increasing trend in delivery times that causes majority of the sample points to be beyond the 3-sigma limit. While the data points are unstable, they definitely seem to follow a trend which one would assume would continue in such a matter with more data points.

Type I (Manufacturer's) Error

Using the data provided, the following error probabilities were calculated.

Table 5: Type I error likelihood for A and B

	A	B
Type I error probability	0.002699	0.00319

A manufacturer's or type I error is explained as when a process is deemed to be seemingly out of control yet it is actually in control and working. This error incurs costs to the manufacturer for this error hence the name. Control charts are presumed to follow the normal distribution, thus for a process to be deemed out of control it indicates there is a point beyond the 3-sigma limit. To find the type I error likelihoods, the pnorm() function in R was used, as it assumes a normal distribution. The probabilities found were very small which was to be expected for manufacturer's error.

Best delivery time

By using the brute-force method an answer for finding the optimal delivery time in hours is found. This is where the centre line should be placed that will result in the lowest cost. The following graph indicated this turning point or "sweet spot" in red where the graph is at a minimum before it starts to increase again. Therefore, one must centre the line on 3 hours for maximum profit, shown below:

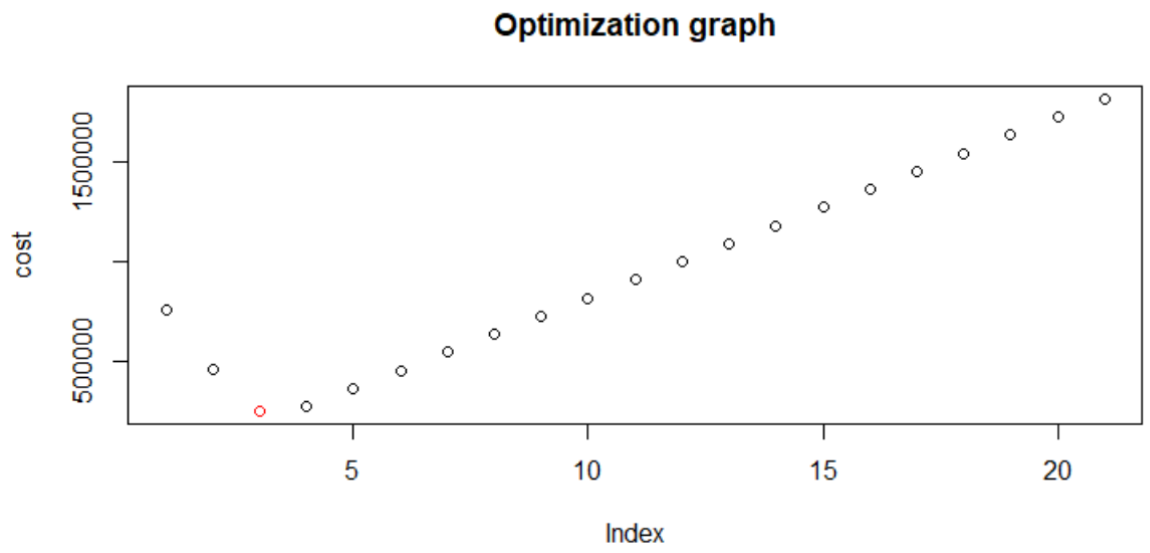

Figure 13: Graph indicating the recommended delivery time to reduce costs for technology items

It is clear from this that after 3 hours, the cost of delivery continues to grow, this is again proved using the min() function. The cost found for three hours was R250 002.50

Type II (Consumer's) Error

Using the delivery process average of 23 hours, the consumer's error calculated was 0.394005, this can also be shown as a 39.4% chance, which is far greater than the manufacturer's error.

Part Five: DOE and MANOVA

The following tables represent values found using MANOVA.

Table 6: Response found for age and price

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
Price	1	1 485 839	1 485 893	3646.7	< 2.2 e-16
Residuals	179976	73 331 231	407		

Table 7: Response found for delivery time and price

	Df	Sum Sq	Mean Sq	F value	Pr(> F)
Price	1	304 294	304 294	1576	< 2.2 e-16
Residuals	179976	34 748 723	193		

For a MANOVA test to be conducted both an independent and dependent variable must be chosen. In this case the independent variable was price. The two independent variables were delivery time and age. A significant value of 0.001 was assumed to find the values shown in tables 6 & 7.

The test is done in order to test whether the dependent variable selected has a significant influence on the independents. The returned P values are both less than 2.2×10^{-16} which is an incredibly small value, essentially zero. This is significantly smaller than 0.001 for both independent variables. This indicates that both the age and delivery time differ greatly depending on the price of the product.

Part Six: Reliability of the service and products

Problem 6

A blueprint specification for the thickness of a refrigerator part at Cool Food, Inc. is 0.06 ± 0.04 centimetres (cm). It costs $45 to scrap a part that is outside the specifications. Determine the Taguchi loss function for this situation.

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

$$\text{Substitute } x = 0.04\text{: } \$45 = k\,(0.04)^2$$

$$\therefore k = 28\,125$$

$$\therefore L(x) = 28\,125\,(x - T)^2$$

The Taguchi loss function found above is then graphed in the figure on the following page.

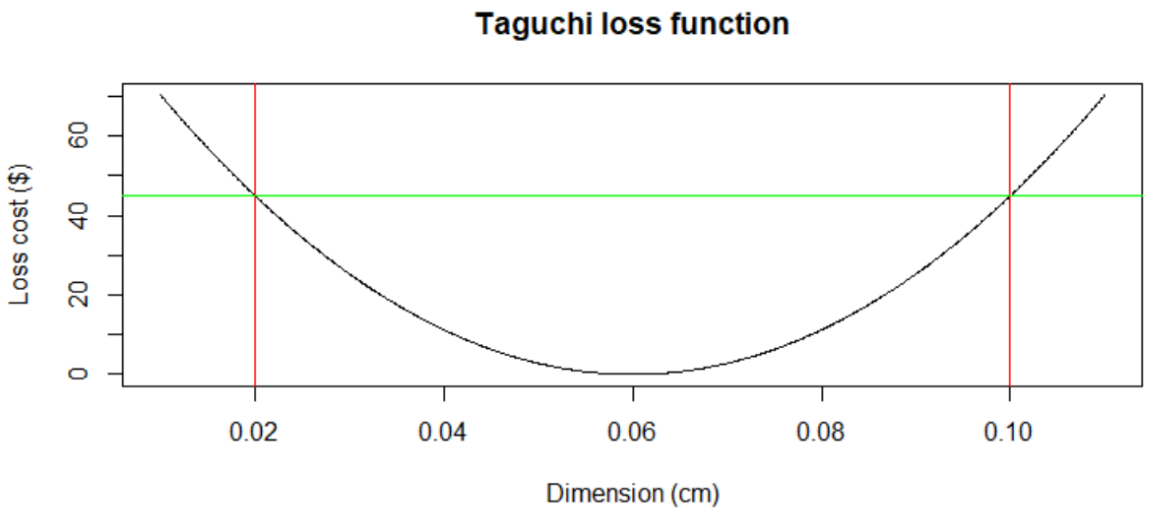

Figure 14: The Taguchi loss function for the Cool Food, Inc textbook problem 6

In the figure above one can see these costs follow the quadratic form, as they become more expensive the further you get away from 0.06 cm, while at 0.06 cm there is no cost. The red lines are situated at ± 0.04 cm and the horizontal line in green is the cost of scrap.

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 scrap, they found a way to reduce the scrap cost to $35 per part.

a) Determine the Taguchi loss function for this situation.

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

$$\text{Substitute } x = 0.04\text{: } \$35 = k\,(0.04)^2$$

$$\therefore k = 21\,875$$

$$\therefore L(x) = 28\,125\,(x - T)^2$$

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

$$\text{Substitute } x = 0.027\text{: } 28\,125\,(0.027)^2$$

$$\therefore L = \$15.95$$

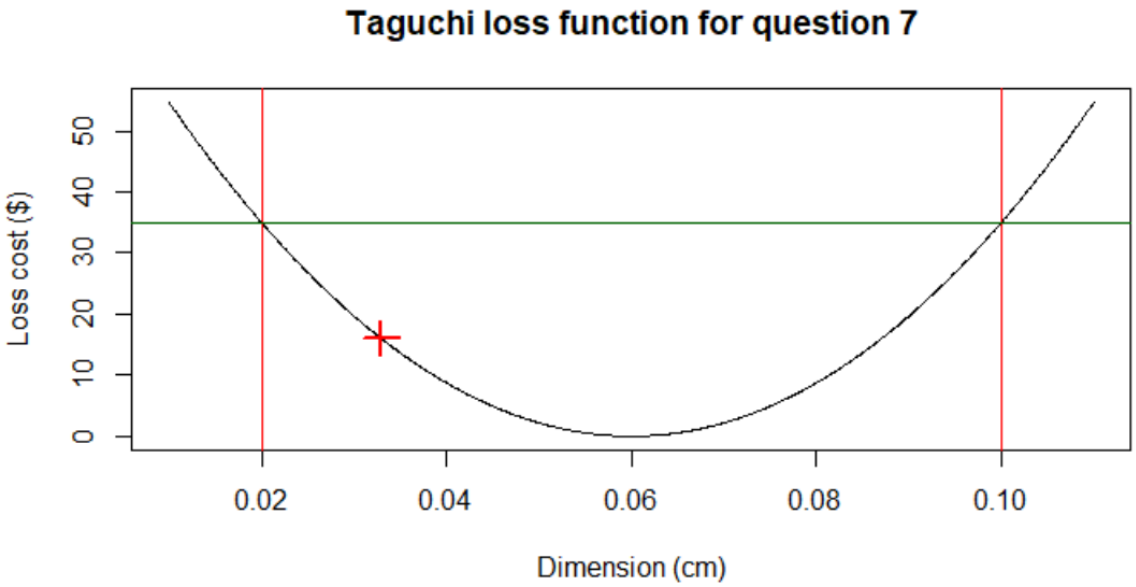

Figure 15: The Taguchi loss function for the Cool Food, Inc textbook problem 6

Looking at the figure above it once again follows the quadratic form, as the cost increases increasingly moving further away from 0.06 cm. The + symbol represents the L value found earlier.

Problem 27

Magnaplex, Inc. has a complex manufacturing process, with three operations that are performed in series. Because of the nature of the process, machines frequently fall out of adjustment and must be repaired. To keep the system going, two identical machines are used at each stage; thus, if one fails, the other can be used while the first is repaired (see accompanying figure).

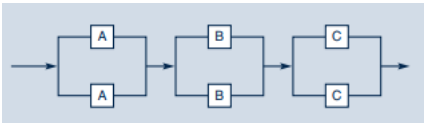

Figure 16: Production system diagram from the textbook problem

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

Reliability = 0.85 × 0.92 × 0.90

= 0.7038

= 70.38%

∴ This is the system's reliability when there is a single machine at each stage (or connected in series)

When each stage has two identical machines (ie. connected in parallel), the reliability then becomes:

Reliability = (1 - (1 - 0.85) × (1 - 0.85)) × (1 - (1 - 0.92) × (1 - 0.92)) × (1 - (1 - 0.9) × (0.1))

= 0.9615316

= 96.15%

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

Reliability increased by 25.77%, which decreases the amount of waste by a large margin. This is a significant change, which could result in more accurate predictions for future sales forecasting. Provided Magnaplex can find additional budget for backup machines, it would result in much higher profits in the future. They should most certainly look into connecting all of their machines in parallel.

Delivery Process Vehicles Question

For the delivery process, there are 20 delivery vehicles available, of which 19 is required to be operating at any time to give reliable service. During the past 1560 days, the number of days that there was 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.

a) Estimate on how many days per year we should expect reliable delivery times, given the information above.

b) If we increased our number of vehicles by one to 21, how many days per year we should expect reliable delivery times?

Question A

Table 8: Table of vehicle reliability for 0 or 1 failures

	Probability
0 vehicle failures	0.8615411
1 vehicle failure	0.1288543
TOTAL	0.9903954

Table 9: Table of driver reliability for 0 or 1 failures

	Probability
0 driver failures	0.9344269
1 driver failure	0.06347701
TOTAL	0.9979039

Multiplying these total probabilities with 365 to represent the number of days in one year returned:

361.4943 days of reliable vehicles and 364.2349 days of reliable drivers

If the number of days calculated for both of these values was rounded down to the nearest integer, then this reflects 361 and 364 days of reliable deliveries for a vehicles and drivers respectively in a year consisting of 365 days total.

Question B

Using the same logic as above and modifying the code slightly, increasing the number of vehicles by one thus means there will be 364.9965 days of reliable drivers.

Rounding this number to the nearest integer would then indicate that every day of the year would be available for reliable drivers with the final value equal to 365 days.

Conclusion

Throughout this report, the given data was separated into valid and invalid datasets, using the valid set to create value and interpret it properly. This was done by creating numerous graphs and charts, using DOE and MANOVA tests, while looking at the different error probabilities as well. Looking at these forms of analysis and the problems solved in the final part of the report it is clear that a greater understanding of data analytics has been achieved and further familiarity working in RStudio and with the R language has been gained. The management decisions suggested reflect this and the final exercises gave a further understanding and familiarity for potential future situations.

References

Evans, J. and Lindsay, W. (2020) Managing for Quality and Performance Excellence. 11th edn. Boston, MA, USA: Cengage Learning.