Quality Assurance
ECSA Report

Prepared For: Stellenbosch University Industrial Engineering Department, ECSA

Prepared By: Daniel Tukker (23541660)

Date: 21 October 2022

Abstract

This report contains an overview of the dataset salesData. The valid data is extracted and analysed for trends, wherein conclusions are drawn. Statistical analysis of the dataset, including Process Control Charts and MANOVA charts, is completed to further gain insight into the data. Finally, additional problems are complete and final conclusions are drawn.

Table of Contents

List of Figures

List of Tables

1 Introduction

The aim of this project is to successfully pre-process the given dataset and complete all the required tasks set out by the ECSA GA 4 report guideline.

2 Data Wrangling (Part 1)

Analysis of datasets require, in most cases, pre-processing to extract the data that we want in an easy-to-use format. The initial *salesData* dataset had issues with missing and negative values as can be seen below (errorsData contains only instances with errors, i.e negative and NA values).

	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	12345	18973	93	Gifts	NA	2026	6	11	15.5	Website
2	16320	44142	82	Household	-588.8	2023	10	2	48.0	EMail
3	16321	81959	43	Technology	NA	2029	9	6	22.0	Recommended
4	19540	65689	96	Sweets	-588.8	2028	4	7	3.0	Random
5	19541	71169	42	Technology	NA	2025	1	19	20.5	Recommended
6	19998	68743	45	Household	-588.8	2024	7	16	45.5	Recommended

Figure 1: First 6 instances in the errorsData dataset

Price values cannot be negative as the value is an income, and the values also cannot be “NA” values as each sale recorded needs to have a value attached. These inconsistencies were removed and a new dataset, *cleanData*, was produced as seen below.

	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	463	47101	50	Clothing	1030.86	2021	1	1	9.0	Recommended
2	2627	88087	21	Clothing	428.03	2021	1	1	10.0	Recommended
3	3374	25418	68	Household	13184.41	2021	1	1	48.5	Website
4	5288	13566	94	Household	7021.90	2021	1	1	42.0	Recommended
5	8182	84692	35	Clothing	475.18	2021	1	1	9.0	Recommended
6	9272	46305	72	Clothing	580.98	2021	1	1	8.5	Random

Figure 2: First 6 instances in the cleanData dataset

The data is arranged with a new primary key (“rowno” for row numbers) based on the new ordering of the data by date, in *cleanerData* below.

	rowno	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
1	1	463	47101	50	Clothing	1030.86	2021	1	1	9.0	Recommended
2	2	2627	88087	21	Clothing	428.03	2021	1	1	10.0	Recommended
3	3	3374	25418	68	Household	13184.41	2021	1	1	48.5	Website
4	4	5288	13566	94	Household	7021.90	2021	1	1	42.0	Recommended
5	5	8182	84692	35	Clothing	475.18	2021	1	1	9.0	Recommended
6	6	9272	46305	72	Clothing	580.98	2021	1	1	8.5	Random

Figure 3: First 6 instances in the cleanerData dataset

3 Descriptive analytics on valid dataset (Part 2)

The dataset *cleanerData* was used to analyse and calculate various Process Indices which help us to determine variability within particular processes.

3.1 Graphical analysis of the dataset

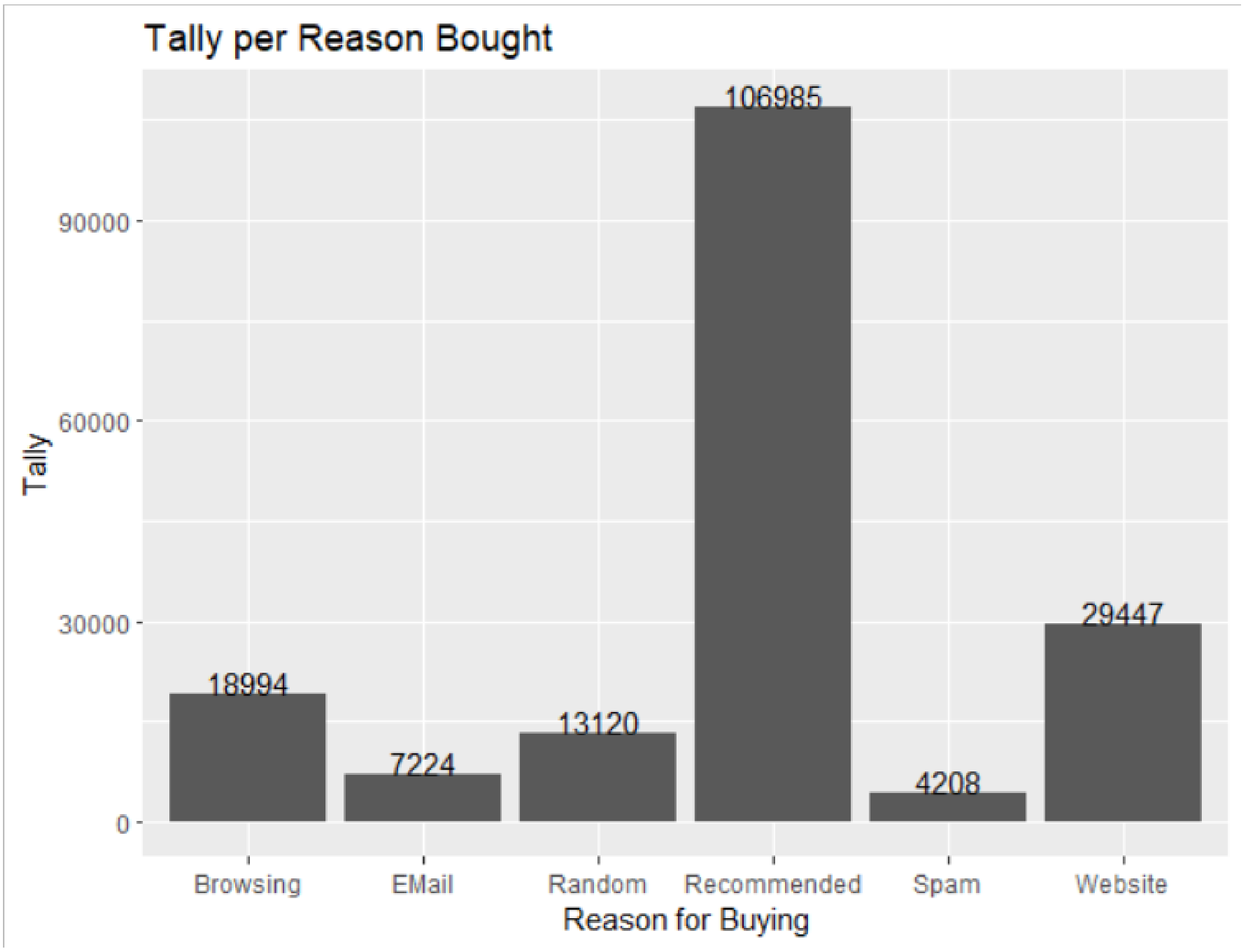

Figure 4: Bar graph of quantities purchased per reason bought

This figure demonstrates the power of word of mouth and experience, as the recommended columns is the vast majority of sales made. The actual website and via browsing would be the second and third most common methods due to online advertising and advertisement personalisation due to website cookies.

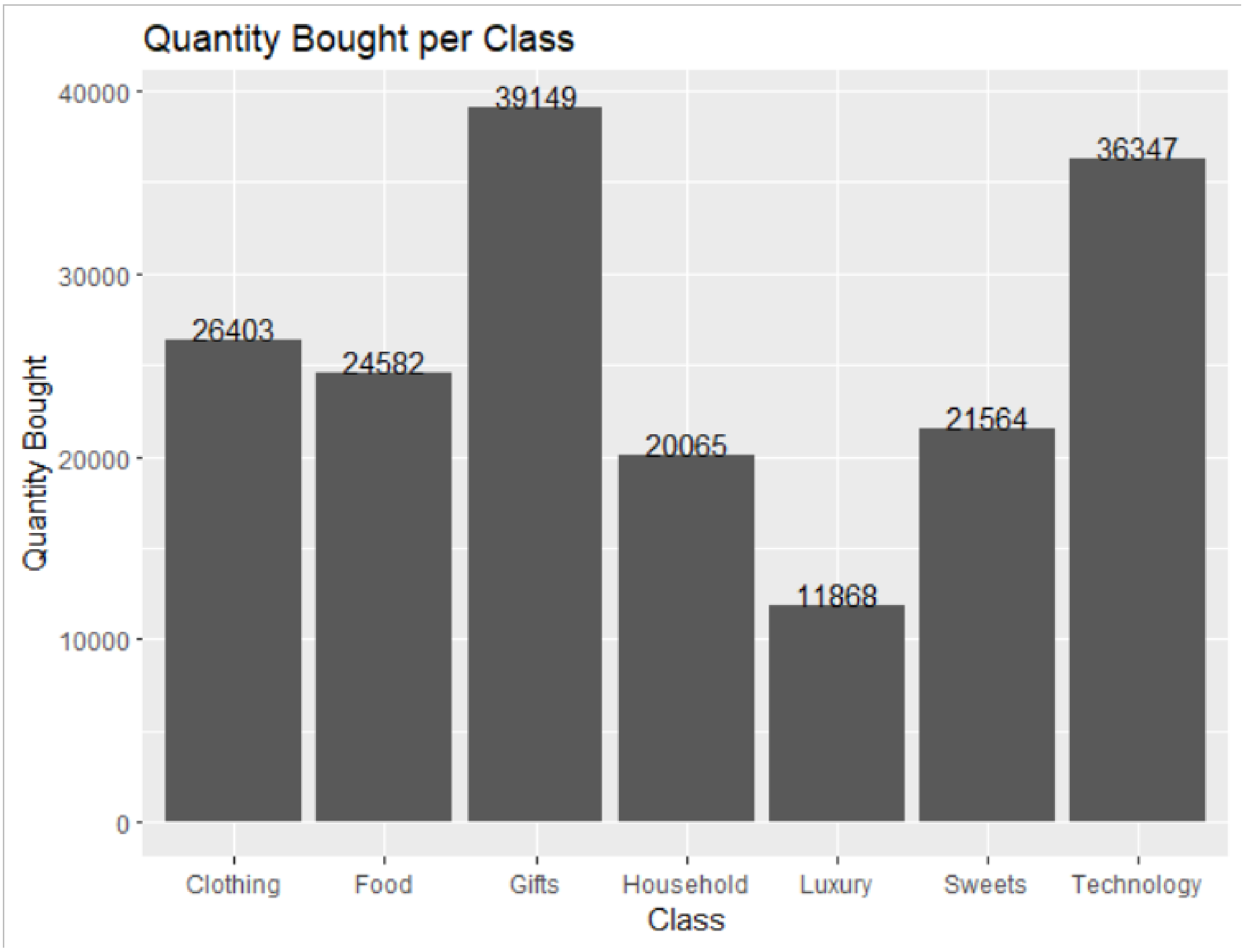

Figure 5: Bar graph for quantity bought per class

An initial observation of the above graph shows that luxury has the smallest quantity, and this is because there is generally a minority of people in the general population who can afford to live with luxuries. Food and clothing are necessities and are among the top 4 most purchased. Most people nowadays have some form of technology, which is the reason for the high quantity of technology purchased.

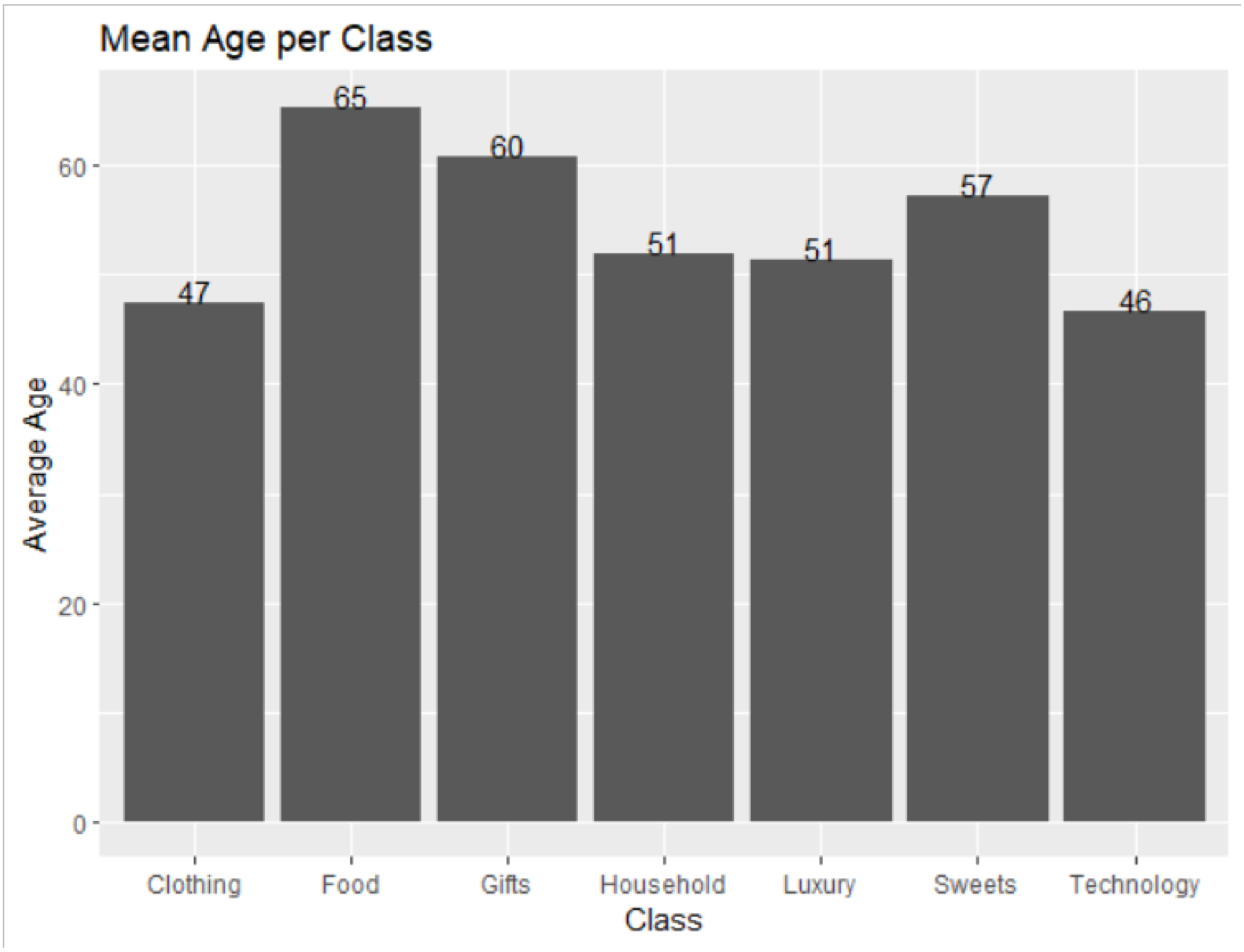

Figure 6: Bar Graph of mean age per class

It is to be expected that the mean age of the technology class would be the lowest as most elderly people grew up without smartphones, new laptops, and other types of technology, whereas the younger generations make use of these on a day-to-day basis for work and play. Food would naturally be high as younger generations are dependent on the older generations for survival until they start working. Clothing is also a class that would tend to be lower, as younger people are much more likely to follow fashion trends.

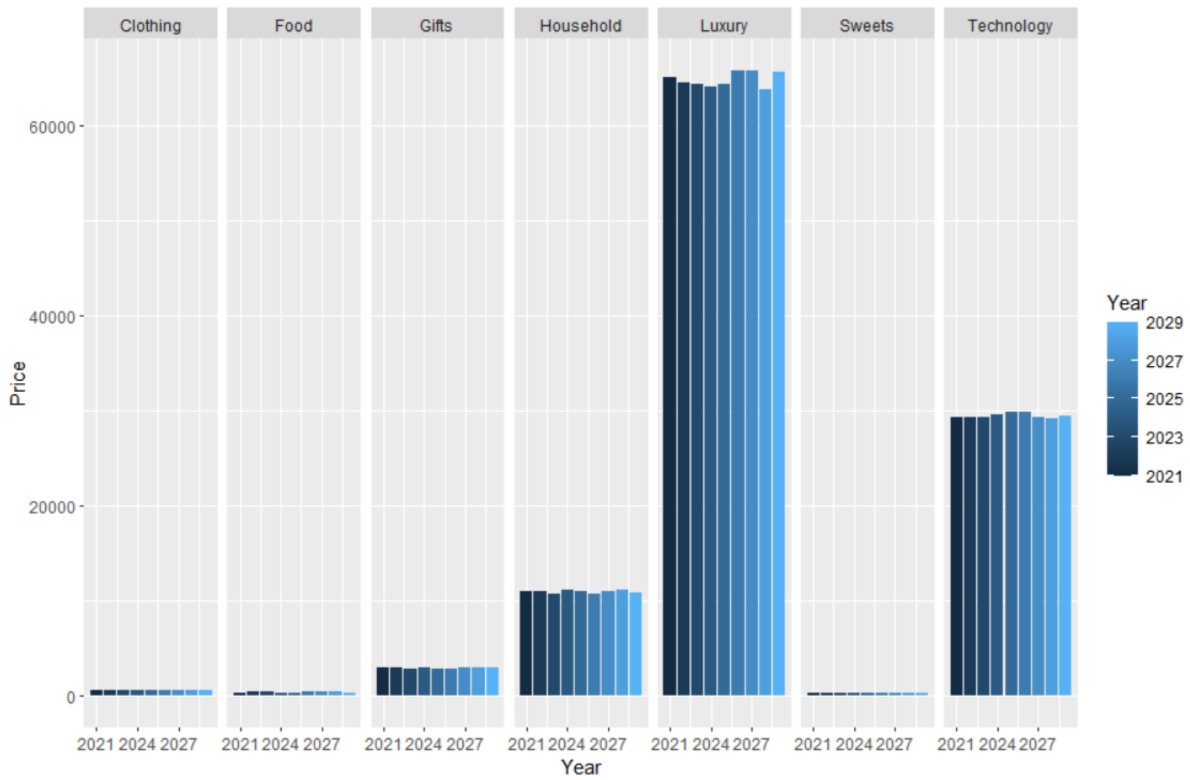

Figure 7: Facet grid (per class) of bar graphs indicating mean price per year

The above graph shows that the price for luxury items well exceeding any other class. The fluctuation between classes, year on year, was minimal. Technology costs are also high due to the premium companies place on new devices that are in high demand.

3.2 Potential capability (C_p)

The first Process Index is C_p, which is potential capability, measuring how well a certain process can perform assuming a normal distribution for the process, centre between the lower and upper specification limits. It was found that the C_p value was 1.142207. This shows that the distribution of the process is within the specifications of the process.

3.3 Potential capability, upper (C_{pu})

This Process Index is C_{pu}, which is potential capability but with specifications consisting of only an upper limit. It was found that the C_{pu} value was 0.3796933. This shows that the process is unable to meet its given requirements. This is due to the upper limit being little over a single standard deviation from the mean.

3.4 Potential capability, lower (C_{pl})

This Process Index is C_{pl}, which is potential capability but with specifications consisting of only a lower limit. It was found that the C_{pl} value was 1.90472. This shows that the process is able to meet its given requirements. This is due to the lower limit being much more than 3 standard deviations from the mean.

3.5 Process capability index (C_{pk})

This Process Index is C_{pk}, which measures both proximity of a process readings mean to the centre of the upper and lower specification limits, and the spread of the process readings. It was found that the C_{pk} value was 0.3796933, as this value is simply the minimum of C_{pu} and C_{pl}. This indicates that the process is unable to meet its given requirements.

3.6 Why does LSL = 0 make sense?

The feature being dealt with is delivery time. This is thus the time from the start of a process until it is completed. Realistically, it is impossible that the process could be completed in a negative number of minutes; although 0 is highly unlikely, it is still the lowest feasible value and is therefore used as the lower specification limit.

4 Statistical Process Control (Part 3)

4.1 $\bar{X}$ Summary Chart and Process Control Chart Initialisation

Class	UCL	U2Sigma	U1Sigma	centreLine	L1Sigma	L2Sigma	LCL
Technology	22.97462	22.10789	21.24117	20.37444	19.50772	18.641	17.77427
Luxury	5.49397	5.24116	4.98836	4.73556	4.48275	4.22995	3.97715
Household	50.24833	49.01963	47.79092	46.56222	45.33352	44.10482	42.87612
Sweets	2.89704	2.75729	2.61753	2.47778	2.33802	2.19827	2.05851
Gifts	9.48856	9.11275	8.73693	8.36111	7.98529	7.60948	7.23366
Food	2.70946	2.63631	2.56315	2.49	2.41685	2.34369	2.27054
Clothing	9.40493	9.25996	9.11498	8.97	8.82502	8.68004	8.53507

Table 1: $\bar{X}$ Summary Table

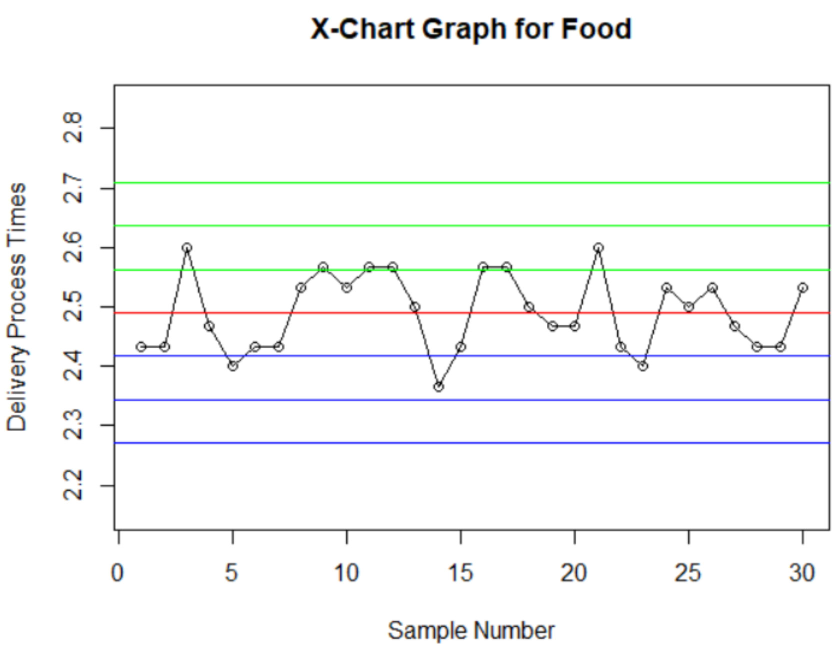

Figure 8: $\bar{X}$ process control chart for all instances in the "Food" class

4.2 S Summary Chart and Process Control Chart Initialisation

Class	UCL	U2Sigma	U1Sigma	centreLine	L1Sigma	L2Sigma	LCL
Technology	5.18057	4.55222	3.92388	3.29553	2.66718	2.03883	1.41049
Luxury	1.51105	1.32778	1.1445	0.96123	0.77795	0.59468	0.41141
Household	7.34418	6.45341	5.56264	4.67187	3.7811	2.89033	1.99956
Sweets	0.83534	0.73402	0.6327	0.53139	0.43007	0.32875	0.22743
Gifts	2.24633	1.97388	1.70142	1.42897	1.15651	0.88405	0.6116
Food	0.43725	0.38421	0.33118	0.27815	0.22511	0.17208	0.11905
Clothing	0.86656	0.76146	0.65635	0.55125	0.44614	0.34104	0.23593

Table 2: S Summary Table

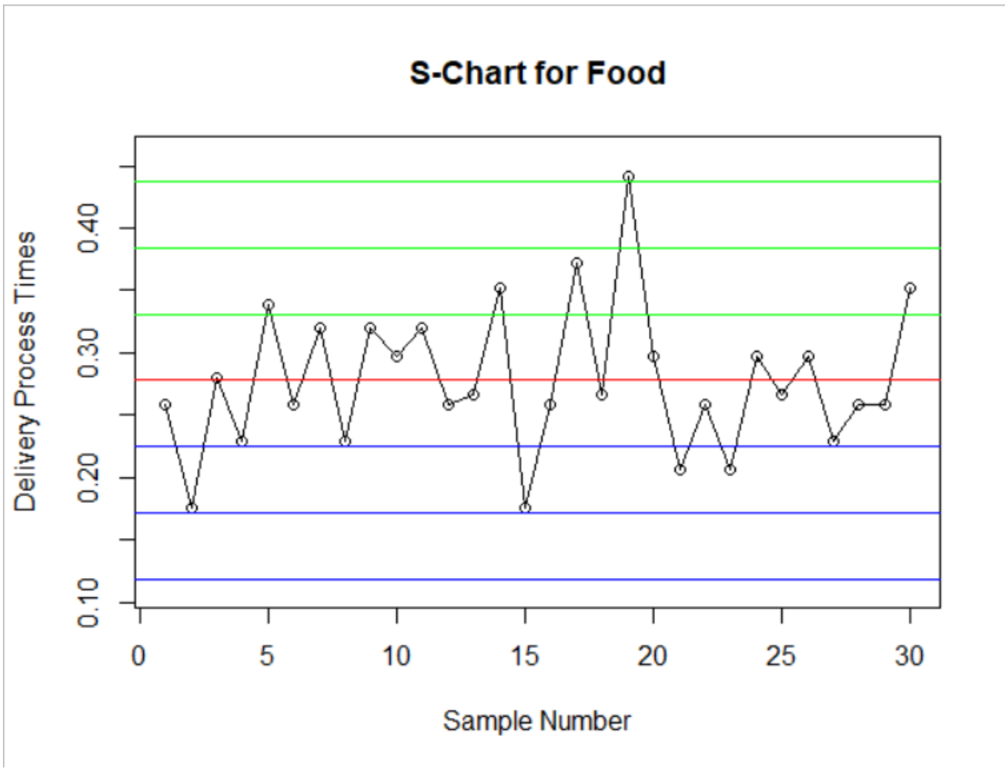

Figure 9: Standard deviation process control chart for all instances in the "Food" class

5 Optimising delivery process (Part 4)

5.1 Analysis of SPC graphics per class over all samples (4.1A+B)

5.1.1 Clothing

The table below summarises the first and last 3 instances, as well as the number of instances, in the sample for the "Clothing" class for which the instances are below or above the outer control limits.

QUANTITY OF INTEREST	SAMPLE OR VALUE
1ST OUT OF CONTROL SAMPLE	455
2ND OUT OF CONTROL SAMPLE	702
3RD OUT OF CONTROL SAMPLE	1152
3RD LAST OUT OF CONTROL SAMPLE	1677
2ND LAST OUT OF CONTROL SAMPLE	1723
LAST OUT OF CONTROL SAMPLE	1724
TOTAL NO. OUT OF CONTROL INSTANCES VALUE	17
LARGEST NO. OF CONSECUTIVE INSTANCES	10
ENDING SAMPLE	24

Table 3: summary of out-of-control samples, Clothing class

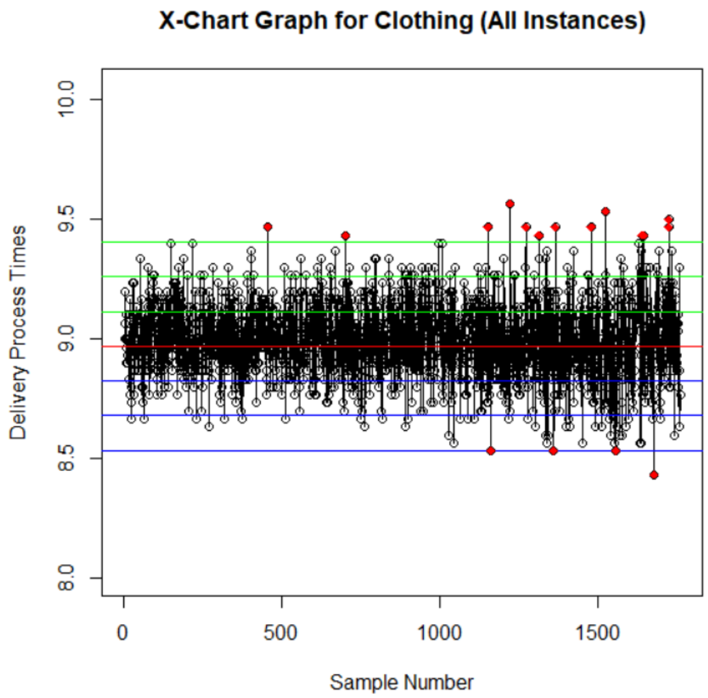

Figure 10: Process Control Chart ($\overline{X}$) for all instances in the "Clothing" class, with out-of-control instances highlighted in red.

5.1.2 Household

The table below summarises the first and last 3 instances, as well as the number of instances, in the sample for the "Household" class for which the instances are below or above the outer control limits.

QUANTITY OF INTEREST	**SAMPLE OR VALUE**
1ST OUT OF CONTROL SAMPLE	252
2ND OUT OF CONTROL SAMPLE	387
3RD OUT OF CONTROL SAMPLE	629
3RD LAST OUT OF CONTROL SAMPLE	1335
2ND LAST OUT OF CONTROL SAMPLE	1336
LAST OUT OF CONTROL SAMPLE	1337
TOTAL NO. OUT OF CONTROL INSTANCES VALUE	400
LARGEST NO. OF CONSECUTIVE INSTANCES	19
ENDING SAMPLE	473

Table 4: summary of out-of-control samples, Household class.

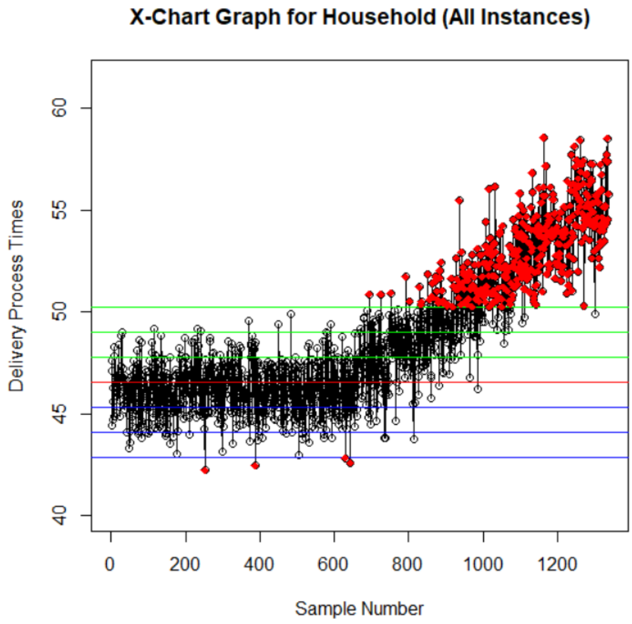

Figure 11: Process Control Chart ($\bar{X}$) for all instances in the "Food" class, with out-of-control instances highlighted in red

With modern homeowners having increasingly complicated needs, the delivery time would be impacting moving into the future. More complex items, larger items, and larger globalised distances to travel all play a role.

5.1.3 Food

The table below summarises the first and last 3 instances, as well as the number of instances, in the sample for the "Food" class for which the instances are below or above the outer control limits.

QUANTITY OF INTEREST	SAMPLE OR VALUE
1ST OUT OF CONTROL SAMPLE	75
2ND OUT OF CONTROL SAMPLE	633
3RD OUT OF CONTROL SAMPLE	1203
3RD LAST OUT OF CONTROL SAMPLE	N/A
2ND LAST OUT OF CONTROL SAMPLE	1467
LAST OUT OF CONTROL SAMPLE	1515
TOTAL NO. OUT OF CONTROL INSTANCES VALUE	5
LARGEST NO. OF CONSECUTIVE INSTANCES	11
ENDING SAMPLE	798

Table 5: summary of out-of-control samples, Food class.

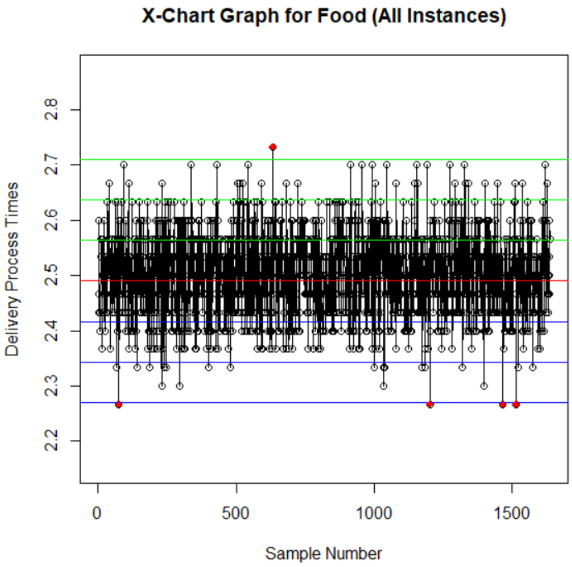

Figure 12: Process Control Chart ($\overline{X}$) for all instances in the "Food" class, with out-of-control instances highlighted in red.

Although there is random variation as time goes on, the graph depicts a normal distribution as there are more instances close to the mean and less moving away from the mean. Food is a stable item in the market in which there are not too many new introductions of different products, so there would be a lower likelihood of having out-of-control samples.

5.1.4 Technology

The table below summarises the first and last 3 instances, as well as the number of instances, in the sample for the "Technology" class for which the instances are below or above the outer control limits.

QUANTITY OF INTEREST	SAMPLE OR VALUE
1ST OUT OF CONTROL SAMPLE	37
2ND OUT OF CONTROL SAMPLE	398
3RD OUT OF CONTROL SAMPLE	483
3RD LAST OUT OF CONTROL SAMPLE	1872
2ND LAST OUT OF CONTROL SAMPLE	2009
LAST OUT OF CONTROL SAMPLE	2071
TOTAL NO. OUT OF CONTROL INSTANCES VALUE	17
LARGEST NO. OF CONSECUTIVE SAMPLES	15
ENDING SAMPLE	2308

Table 6: summary of out-of-control samples, Technology class.

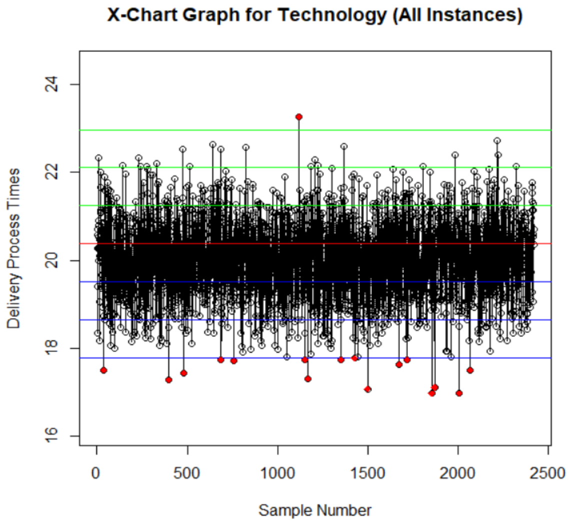

Figure 13: Process Control Chart ($\bar{X}$) for all instances in the "Technology" class, with out-of-control instances highlighted in red.

Although there is random variation as time goes on, the graph depicts a normal distribution as there are more instances close to the mean and less moving away from the mean. Technology nowadays is not necessarily expanding in macro form factor, so delivery times would not be affected so drastically as things like gifts.

5.1.5 Sweets

The table below summarises the first and last 3 instances, as well as the number of instances, in the sample for the "Sweets" class for which the instances are below or above the outer control limits.

QUANTITY OF INTEREST	SAMPLE OR VALUE
1ST OUT OF CONTROL SAMPLE	942
2ND OUT OF CONTROL SAMPLE	1104
3RD OUT OF CONTROL SAMPLE	483
3RD LAST OUT OF CONTROL SAMPLE	N/A
2ND LAST OUT OF CONTROL SAMPLE	1294
LAST OUT OF CONTROL SAMPLE	1403
TOTAL NO. OUT OF CONTROL SAMPLES VALUE	5
LARGEST NO. OF CONSECUTIVE SAMPLES	13
ENDING SAMPLE	317

Table 7: summary of out-of-control samples, Sweets class.

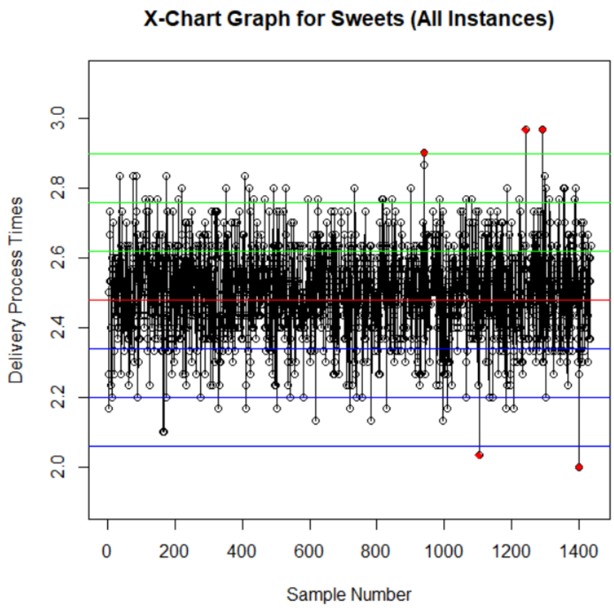

Figure 14: Process Control Chart ($\bar{X}$) for all instances in the "Sweets" class, with out-of-control instances highlighted in red.

Although there is random variation as time goes on, the graph depicts a normal distribution as there are more instances close to the mean and less moving away from the mean. Sweets are a relatively stable item in the market and thus less variation can be expected.

5.1.6 Gifts

The table below summarises the first and last 3 instances, as well as the number of instances, in the sample for the "Gifts" class for which the instances are below or above the outer control limits.

QUANTITY OF INTEREST	SAMPLE OR VALUE
1ST OUT OF CONTROL SAMPLE	213
2ND OUT OF CONTROL SAMPLE	216
3RD OUT OF CONTROL SAMPLE	1243
3RD LAST OUT OF CONTROL SAMPLE	2607
2ND LAST OUT OF CONTROL SAMPLE	2608
LAST OUT OF CONTROL SAMPLE	2609
TOTAL NO. OUT OF CONTROL INSTANCES VALUE	5
LARGEST NO. OF CONSECUTIVE INSTANCES	17
ENDING SAMPLE	2191

Table 8: summary of out-of-control samples, Gifts class.

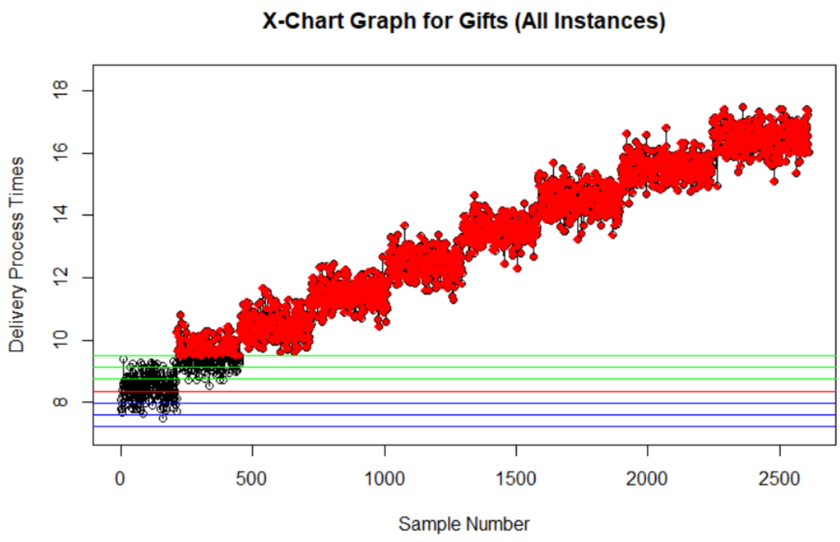

Figure 15: Process Control Chart ($\overline{X}$) for all instances in the "Sweets" class, with out-of-control instances highlighted in red.

The trend is a strong upwards movement each year. This could very well be attributed to year-on-year increase in interest in the various types of products offered, as well as an increase in the variety of products sold. Other factors like global trends, advertising and social media could have an impact on such an increase, as there would be a greater strain on supply chain to fulfil such needs in a timeous manner.

5.1.7 Luxury

The table below summarises the first and last 3 instances, as well as the number of instances, in the sample for the "Luxury" class for which the instances are below or above the outer control limits.

QUANTITY OF INTEREST	SAMPLE OR VALUE
1ST OUT OF CONTROL SAMPLE	142
2ND OUT OF CONTROL SAMPLE	171
3RD OUT OF CONTROL SAMPLE	184
3RD LAST OUT OF CONTROL SAMPLE	789
2ND LAST OUT OF CONTROL SAMPLE	790
LAST OUT OF CONTROL SAMPLE	791
TOTAL NO. OUT OF CONTROL INSTANCES VALUE	434
LARGEST NO. OF CONSECUTIVE INSTANCES	8
ENDING SAMPLE	171

Table 9: summary of out-of-control samples, Luxury class.

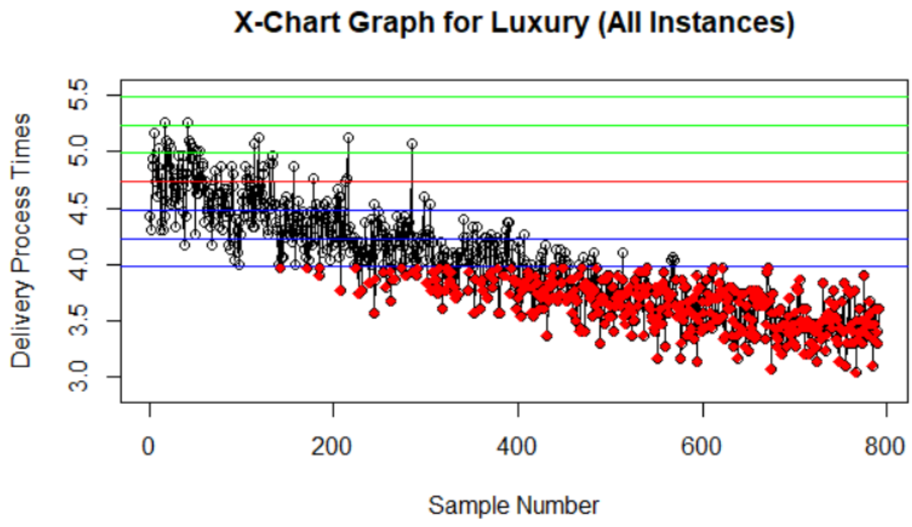

Figure 16: Process Control Chart ($\bar{X}$) for all instances in the "Luxury" class, with out-of-control instances highlighted in red.

The above graphic clearly depicts that with time, the delivery time as decreased. This is to be expected, as new technologies and delivery methods are introduced each year, while production processes and scheduling are improved. All these factors help to reduce the time the products take to get from the business to the customer.

Over time, the process control charts would need to be adjusted, with differing averages as the production and delivery scenario for the business would not remain the same year on year. There needs to be an investigation as to why the out-of-control instances deviate to the extent shown above, as the company need to find reasoning to change averages and specification limits. The samples look at a time horizon of almost a decade, so having all the instances on a single chart may also give the impression of large error.

5.2 Type I error

4.2) The null hypothesis assumes: "The process is in control and centred on the centreline calculated using the first 30 samples". A type I error means that the null is rejected even though it should have been accepted. With regards to this situation, this would mean rejecting the null hypothesis and concluding that the process is out of control even though the process was indeed in control.

After calculations, the probability of making a type I error in the A scenario would be 0.26997960632602%.

The probability of making a type I error in the B scenario would be 0.231937762888541% (72.6666836200723^{19}%, as the longest length of consecutive samples calculated in 4.1B was 19)

5.3 Centring the process around a new delivery time for technology to maximise profits

4.3) A graph was constructing to demonstrate the minimisation of cost versus hours. An optimal point was reached between the two variables, and it was found that a delivery time of 23 hours was the optimal length to optimise costs. The loss is similar to a Taguchi loss as there is no vertical lines as seen in a good/no good loss at the LCL and UCL, allowing more lenience on either side of the tolerances

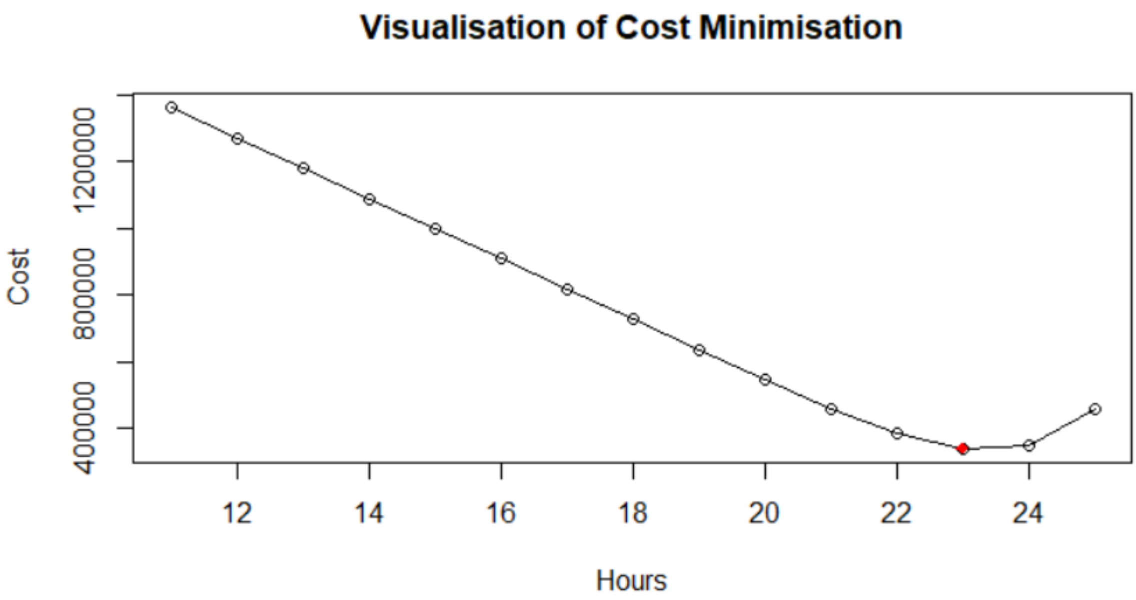

Figure 17: Graph with cost minimisation visualisation

5.4 Type II error

4.4) The null hypothesis assumes: “The process is in control and centred on the centreline calculated using the first 30 samples”. A type II error means that the null hypothesis is accepted even though it should have been rejected. With regards to this situation, this would lead to the conclusion that the process is in control even though the process was indeed not in control.

After calculations, the probability of making a type II error in this scenario would be 48.6193216851394%.

6 DOE and MANOVA(Part 5)

6.1 Testing relationship of Price and Age vs Reason for Buying

The first relationship that will be observed is that of price and age vs reason for buying. These are often common factors that play a role in people's decision to purchase, so this will be checked based on the current dataset.

The null hypothesis tested is: "There is no significant relationship between the two features, price and age, and the target, reason for buying". A standard alpha value used is 0.05. The results from the MANOVA summary from are tabulated as follows:

Reponse	DF	Sum sq error	Mean sq error	F	Pr(>F)
Price	5	1.57421E+12	3.14842E+11	736.26	<2.2e-16
Age	5	106542	21308.4	51.33	<2.2e-16
Price*Age	5	3.75281E+15	7.50562E+14	599.54	<2.2e-16

Table 10: MANOVA testing relationship between price and age, against reason for buying

The P values are significantly smaller than the alpha value, meaning we need to reject the null hypothesis. This leads us to conclude that price and age both have a significant impact on reason for buying.

6.2 Testing relationship of Delivery Time and Price vs Class

The second relationship that will be tested is that of price and delivery time versus class. Different types of products require different kinds of components, have different form factors, and can differ vastly in price, and this test is used to confirm whether this is true based on the dataset.

The null hypothesis tested is: "There is no significant relationship between the two features, price and delivery, and the target, class". A standard alpha value used is 0.05. The results from the MANOVA summary from are tabulated as follows:

Reponse	DF	Sum sq error	Mean sq error	F	Pr(>F)
Delivery Time	6	33458565	5576427	629429	<2.2e-16
Price	6	5.71684E+13	9.52807E+12	80258	<2.2e-16
Delivery Time*Price	6	1.17012E+16	1.95021E+15	52792	<2.2e-16

Table 11: MANOVA testing relationship between price and age, against class

The P values are significantly smaller than the alpha value, meaning we need to reject the null hypothesis. This leads us to conclude that delivery time and price are both significantly impact class.

7 Reliability of Service and Products (Part 6)

7.1 Taguchi Loss

7.1.1 Problem 6

The Taguchi loss function for problem 6 was found to be:

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

$$\therefore L(y) = 28125(y - 0.06)^2$$

7.1.2 Problem 7

a) The Taguchi loss function for problem 7a) was found to be:

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

$$\therefore L(y) = 21875(y - 0.06)^2$$

b) The Taguchi loss \for problem 7b) was found to be:

$$L(y) = 15.946875$$

7.2 System Reliability

7.2.1 Problem 7

a) The probability of the system running (the system reliability) with only one machine per stage was found to be:

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

$$\therefore System\ Reliability(Series) = 0.85 \times 0.92 \times 0.90 = 70.38\%$$

b) The probability of the system running (the system reliability) with both machines running per stage was found to be:

$$Reliability(stage\ A) = 2 \times Reliability(A) - Reliability(A)^2$$

$$Reliability(stage\ B) = 2 \times Reliability(B) - Reliability(B)^2$$

$$Reliability(stage\ C) = 2 \times Reliability(C) - Reliability(C)^2$$

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

$$\therefore System\ Reliability(Parallel) = 0.9775 \times 0.9936 \times 0.99 = 96.15316\%$$

This gives an overall system reliability improvement of:

$$System\ Reliability\ Improvement = System\ Reliability(Parallel) - System\ Reliability(Series)$$

$$System\ Reliability\ Improvement = 96.15316 - 70.38 = 25.77316\%$$

7.3 System Reliability

7.3.1 21 vehicles

The question states that 19 of the 21 vehicles are required for reliable service. Therefore, we need to calculate the binomial probability of 0 and 1 vehicles being unavailable and sum those to multiply with the number of days in a year to find out how many days a year the vehicles were available.

Days with all vehicles available = 314.4625 days (based on 0.8615411 probability)
Days with one vehicle unavailable = 47.03181 days (based on 0.1288543 probability)
∴ Reliable days (vehicles) = 361.4943 days

We are also given the number of days drivers are available, and the same approach as above is followed.

Days with all drivers available = 341.0658 days (based on 0.8615411 probability)
Days with one driver unavailable = 23.16911 days (based on 0.06347701 probability)
∴ Reliable days (drivers) = 364.2349 days

∴ Total combined reliable days = 360.7366 days

7.3.2 22 vehicles

The question states that 19 of the 22 vehicles are required for reliable service. Therefore, we need to calculate the binomial probability of 0, 1 and 2 vehicles being unavailable and sum those to multiply with the number of days in a year to find out how many days a year the vehicles were available.

Days with all vehicles available = 314.4627 days (based on 0.8615415 probability)
Days with one vehicle unavailable = 47.0241 days (based on 0.1288332 probability)
Days with two vehicles unavailable = 3.356128 days (based on 0.009194871 probability)
∴ Reliable days (vehicles) = 364.8203 days

∴ Total combined reliable days = 364.0782 days

8 Conclusion

The dataset was thoroughly analysed using statistical, graphical and insight analysis. Trends in the data were discovered and discussed, and relationships between the various features were observed and conclusions were drawn. Methods of statistical analysis were practiced and improved upon which can be used in future data-driven projects.

9 References

Bhandari, P. 2021. [Online]. Available: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/ [19 October 2022]

ISixSigma. 2022. [Online]. Available: https://www.isixsigma.com/dictionary/cp/ [15 October 2022].

Learn Six Sigma. 2022. [Online]. Available: https://www.leansixsigmadefinition.com/glossary/taguchi-loss-function/ [18 October 2021]

Mountz. 2018. [Online]. Available: https://www.mountztorque.com/Understanding-and-Using-Cpk#:~:text=The%20higher%20the%20Cpk%2C%20the,capable%20of%20meeting%20its%20requirements [18 October 2022]

Stellenbosch Engineering Faculty. 2022. Engineering Academic programmes and faculty information. [Online] Available at: https://www.sun.ac.za/english/Documents/Yearbooks/Current/2021-Engineering.pdf [10 October 2022].

Wikipedia. 2022. [Online]. Available: https://en.wikipedia.org/wiki/Process_capability_index [17 October 2021]