ECSA Graduate Attribute Project (GA4)

Quality Assurance 344

Marcus Potgieter

23670207

Table of Contents

Table of Tables

Table of figures

Introduction:

Client data for an online business is given to be analysed. In some cases, the data is invalid, to fix this the data must be cleaned or removed. The clean data, or valid data, will then be analysed by using descriptive statistics. The data will also be used to determine process capability indices, as well as the delivery process times will be analysed using statistical process control.

Part 1: Data Wrangling

The data file, salesTable2022.csv, contains valid and invalid data and in order for to be used it has to be cleaned by getting rid of all the entries containing missing values as well as the duplicate entries, only then we can start analyzing the data. Firstly the function na.omit() will get rid of the original data, salesTable2022, to create the file cleanerData. Secondly the duplicated() function is used to get rid of all the duplicate entries, to finally obtain the valid that will be analysed, valData. The second data file called incomData will be comprised of all the entries containing missing values.

NA Values

The original data has a few entries containing missing values. The table below shows al these entries that will be taken out:

X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
12345	18973	93	Gifts	NA	2026	6	11	15.5	Website
16321	81959	43	Technology	NA	2029	9	6	22.0	Recommended
19541	71169	42	Technology	NA	2025	1	19	20.5	Recommended
19999	67228	89	Gifts	NA	2026	2	4	15.0	Recommended
23456	88622	71	Food	NA	2027	4	18	2.5	Random
34567	18748	48	Clothing	NA	2021	4	9	8.0	Recommended
45678	89095	65	Sweets	NA	2029	11	6	2.0	Recommended
54321	62209	34	Clothing	NA	2021	3	24	9.5	Recommended
56789	63849	51	Gifts	NA	2024	5	3	10.5	Website
65432	51904	31	Gifts	NA	2027	7	24	14.5	Recommended
76543	79732	71	Food	NA	2028	9	24	2.5	Recommended
87654	40983	33	Food	NA	2024	8	27	2.0	Recommended
98765	64288	25	Clothing	NA	2021	1	24	8.5	Browsing
144444	70761	70	Food	NA	2027	9	28	2.5	Recommended
155555	33583	56	Gifts	NA	2022	12	9	10.0	Recommended
166666	60188	37	Technology	NA	2024	10	9	21.5	Website
177777	68698	30	Food	NA	2023	8	14	2.5	Recommended

Figure 1: Data set containing missing values, to be removed

The Valid Data Set:

After getting rid of all the entries containing missing values as well as the duplicate entries we obtain a valid data set, this data we will use for analysing and figuring out trends. Below we can see the first 20 entries of the valid data set:

X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
238	60742	52	Food	620.54	2026	12	30	2.0	Recommended
252	73202	37	Clothing	973.36	2021	8	1	9.0	EMail
324	25014	50	Technology	32181.02	2023	9	23	17.5	Recommended
378	14808	48	Technology	56240.63	2029	1	5	23.0	Recommended
531	59070	60	Technology	23565.57	2027	1	3	19.5	Browsing
595	60539	44	Luxury	107703.71	2029	7	1	3.5	Recommended
612	70869	24	Luxury	43220.68	2022	8	24	3.5	Recommended
625	71083	38	Gifts	3949.33	2029	12	28	16.5	Recommended
631	28025	50	Luxury	98877.57	2023	5	1	3.0	Recommended
638	15964	81	Gifts	529.57	2022	3	5	9.5	Website
650	44760	85	Food	666.31	2025	1	6	2.0	Recommended
676	13289	68	Clothing	627.32	2022	8	22	7.5	Recommended
679	12335	67	Gifts	751.64	2028	4	17	18.0	Random
692	39674	67	Gifts	5188.68	2024	2	4	10.0	Random
719	42411	31	Clothing	1068.69	2025	1	9	10.5	Random
720	30379	57	Household	12668.28	2022	2	11	48.5	Browsing
724	16306	26	Technology	49054.75	2028	2	13	26.0	Recommended
732	48807	22	Clothing	754.74	2021	11	29	9.0	Browsing
737	20701	74	Technology	53933.85	2029	1	9	20.5	Website
744	96018	68	Sweets	293.36	2021	1	6	2.0	Recommended

Figure 2: Valid data set to be used for analysis

Part 2: Descriptive Statistics

Five-point summary of classes age, price, and delivery time.

AGE	
Minimum	18
1st Quartile	38
Median	53
Mean	55
3rd Quartile	70
Maximum	108

Table 2:5-point summary of class AGE.

PRICE [R]	
Minimum	35.65
1st Quartile	482.31
Median	2256.18
Mean	12294.94
3rd Quartile	15252.76
Maximum	116618.97

Table 1: 5-point summary of class PRICE.

DELIVERY TIME	
Minimum	0.50
1st Quartile	3
Median	10
Mean	14.49
3rd Quartile	18.50
Maximum	75

Table 3: 5-point summary of class DELIVERY TIME.

Visual Descriptive Statistics

Given is histograms of each class and their cash flow distributions.

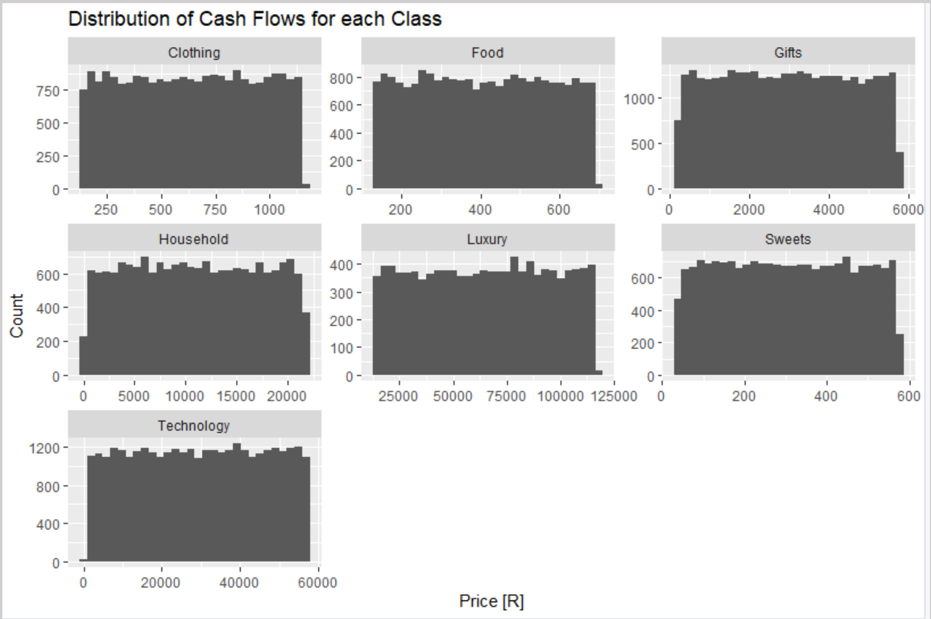


Figure 3: Distribution of Cash Flows for each Class

As we can see above, all the classes have Uniform Distributions of cash flows. This indicates that the cash flows per class are equally probable to occur, we also call this a constant trend.

Next, we look at the total cash flow per class:

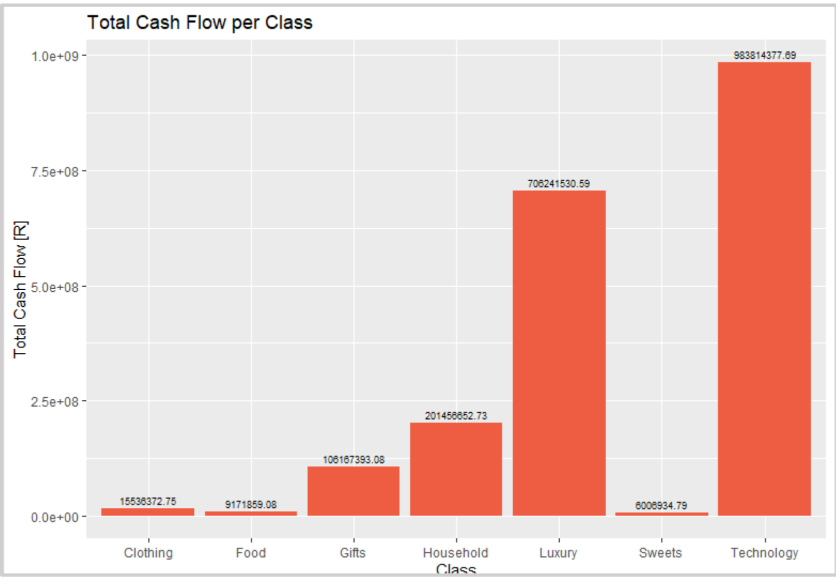


Figure 4: Total Cash Flow per Class

Here we can see that the most money moves through the Technology industry and least in the Sweets industry, but we can also see that the Clothing and Food industry is not far ahead of the Sweets industry.

The following histogram gives a visual description of Total cash flows per year:

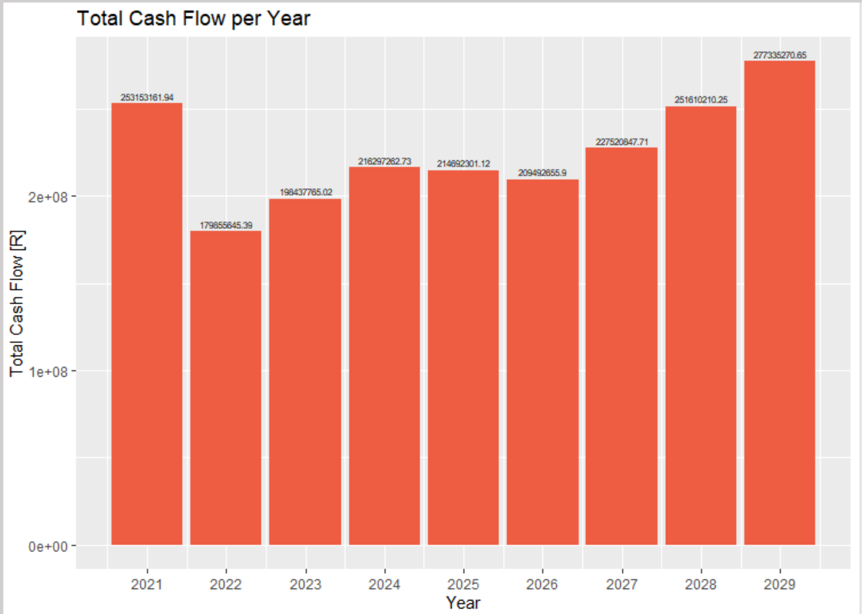


Figure 5: Total Cash Flow per Year

Analysing this histogram, we can that in the year 2021, when the company started, their sales was quite high. This can be due to the promotions the company ran for just opening up, like most companies will do. In the following year, year 2022, there was a drop in cash flow which then picked in the proceeding years, giving a skewed to the left distribution (Mathtec, n.d.).

The following histogram gives the count of reasons why a product was bought:

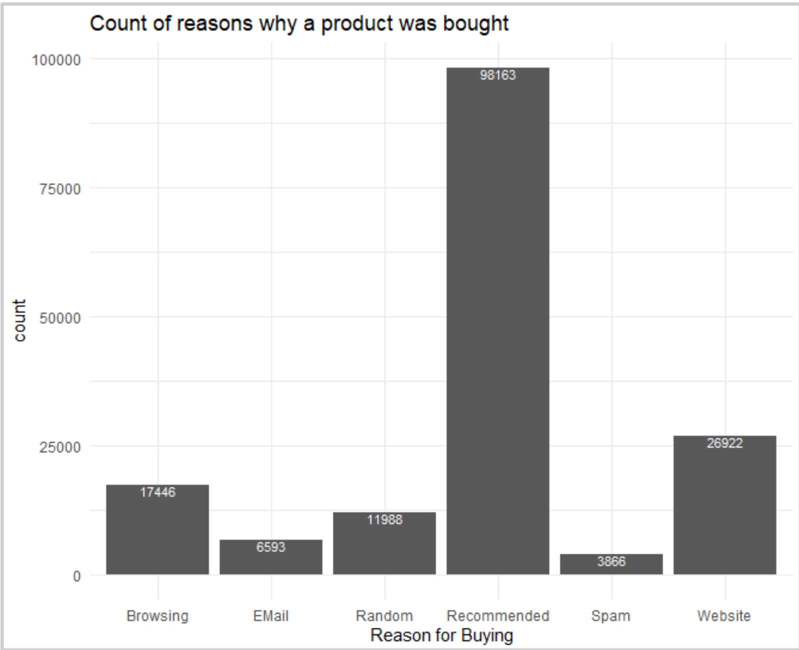


Figure 6: Count of reasons why a product was bought

Analysing this histogram, we can clearly see that the majority reason for buying from the company is through recommendations. This seems true, according to a Nielsen report, 92% of people trust word of mouth from people they know, above all other advertising (Mosley, 2022).

Here we see the distribution of ages buying products per class:

Figure 7: Distribution of ages of buying products per class.

<u>Analysis of distribution of ages buying products per class:</u>

Clothing: Here we see a unimodal skewed right distribution with the peak between 20 and 30 years old.

Food: For the food category we see sort of bimodal distribution that is non-symmetric, with peaks at 50 and 75 years old.

Gifts: For the gift category we see sort of bimodal distribution that is somewhat symmetric, with peaks at 45 and 75 years old.

Household: For the household category we see a unimodal skewed right distribution with the peak at about 45 years old.

Luxury: Here we see a unimodal skewed right distribution with the peak at about 37 years old.

Sweets: At the Sweets category we see a bimodal distribution that is non-symmetric, with peaks at 30 and 60 years old.

Technology: Here we see a unimodal skewed right distribution with the peak between 20 and 37 years old.

Process Capability Indices:

Calculating the Process Capability indices Cp, Cpu, Cpl and Cpk for the process delivery times of technology class items. Assuming the Upper Specification Limit (USL) = 24 hours and the Lower Specification Limit (LSL) = 0.

Cp	1.14206
CpU	0.37919
CpL	1.90493
CpK	0.37919

Table 4: Table of Process Capability Indices.

<u>Analysing the table above:</u>

The Cp is equal to 1.14206 which is greater than 1 and means the data is capable and centred.

The CpK is equal to 0.37919 which is less than 1 and means the data is not capable and is not centred.

A value of 0 was chosen for LSL, this is logical as an item cannot be delivered before an order is placed, thus you cannot have a time less than zero.

Part 3: Statistical Process Control (SPC)

SPC monitors a process to identify special causes of variation and thus whether to take corrective action and fix the process or decide to leave it alone. SPC is done with the aid of control charts.

Control Charts

S-Chart Table:

Class	UCL_s	U2Sigma_s	U1Sigma_s	CL_s	L1Sigma_s	L2Sigma_s	LCL_s
Clothing	0.8094184	0.6373038	0.5942752	0.5512465	0.5082179	0.4651893	0.2930747
Household	7.6447421	5.6628275	5.1673489	4.6718703	4.1763916	3.6809130	1.6989985
Food	0.4526136	0.3363023	0.3072245	0.2781467	0.2490689	0.2199910	0.1036797
Technology	5.4202688	4.0037748	3.6496513	3.2955278	2.9414043	2.5872808	1.1707868
Sweets	0.8588500	0.6405408	0.5859635	0.5313862	0.4768089	0.4222316	0.2039224
Gifts	2.2309957	1.6963087	1.5626370	1.4289652	1.2952935	1.1616217	0.6269347
Luxury	1.5056310	1.1426962	1.0519625	0.9612289	0.8704952	0.7797615	0.4168267

Figure 8: Statistical control S-chart for the delivery times of each of the classes.

X-Chart Table:

Class	UCL_x	U2Sigma_x	U1Sigma_x	CL_x	L1Sigma_x	L2Sigma_x	LCL_x
Clothing	9.406562	9.115521	9.042760	8.970000	8.897240	8.824479	8.533438
Household	50.108929	47.744458	47.153340	46.562222	45.971104	45.379987	43.015516
Food	2.685024	2.555008	2.522504	2.490000	2.457496	2.424992	2.294976
Technology	23.191594	21.313494	20.843969	20.374444	19.904919	19.435394	17.557295
Sweets	2.917475	2.624343	2.551061	2.477778	2.404495	2.331212	2.038081
Gifts	9.506742	8.742988	8.552050	8.361111	8.170173	7.979234	7.215480
Luxury	5.593755	5.021622	4.878589	4.735556	4.592522	4.449489	3.877356

Figure 9: Statistical control X-chart for the delivery times of each of the classes.

Where, UCL = Upper Control Limit, CL = Centre Line, U2Sigma = Upper 2-Sigma-Control limit, U1Sigma = Upper 1-Sigma-Control limit, L2Sigma = Lower 2-Sigma-Control limit, and L1Sigma = Lower 1-Sigma-Control limit.

X and s Charts of the Seven Processes

Here we will be constructing S-Charts to analyse the delivery times per class. If these samples is within the acceptable range, which is +-3 sigma from the centre line the X-Bar Chart can be constructed, with accurate control limits, of the first 30 samples.

Clothing:

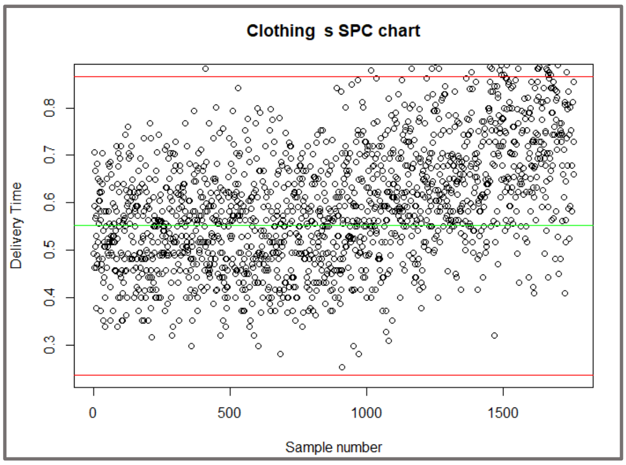


Figure 10: S-Chart for the Clothing class.

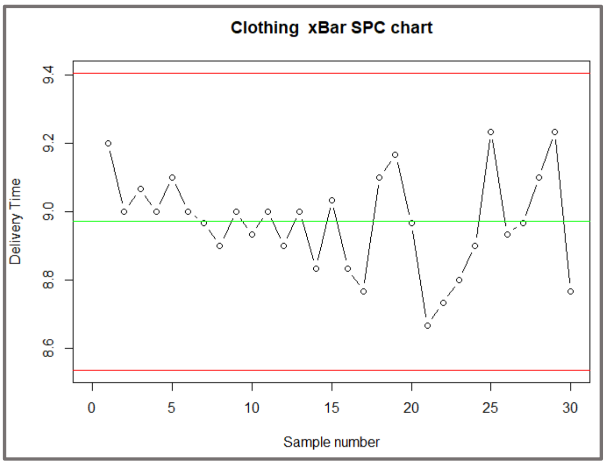


Figure 11: X-Chart for the Clothing class, first 30 samples.

The S-chart shows that most of the points are in the control limits but there are still some outliers, thus we can still say the process variation is in control. Then, we see the X-Bar chart has all the points in control which means the process mean is in control.

Household:

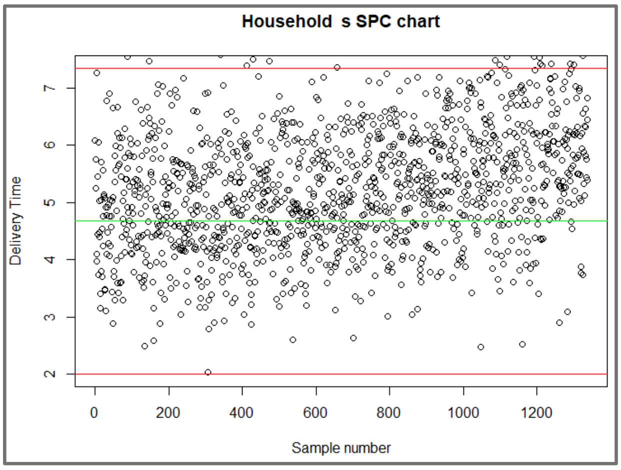


Figure 12: S-Chart for the Household class.

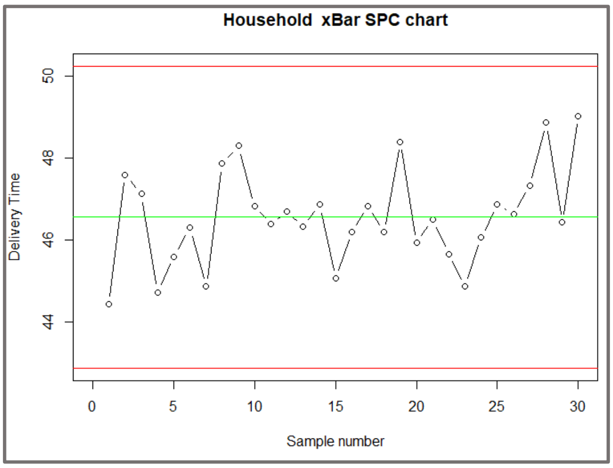


Figure 13: X-Chart for the Household class, first 30 samples.

The S-chart shows that most of the points are in the control limits but there are still some outliers, thus we can still say the process variation is in control. Then, we see the X-Bar chart has all the points in control which means the process mean is in control.

Food:

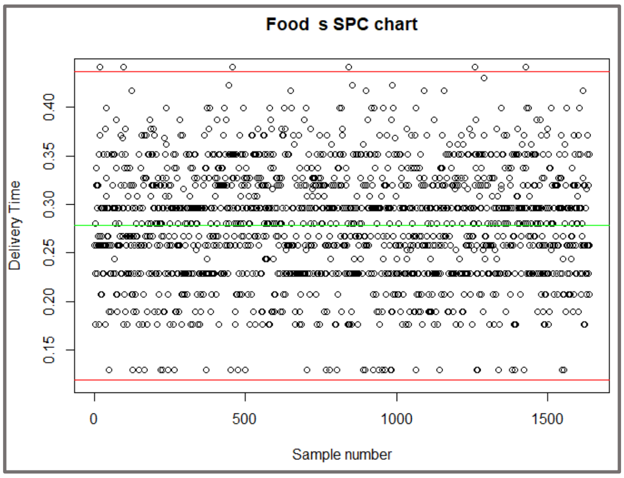


Figure 15: S-Chart for the Food class.

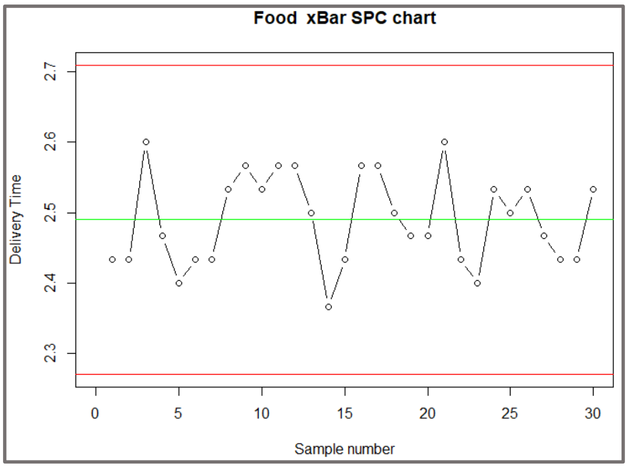


Figure 14: X-Chart for the Food class, first 30 samples.

The S-chart shows that all most all of the points are in the control limits but there, thus we say the process variation is in control. Then, we see the X-Bar chart has all the points in control which means the process mean is in control.

Technology:

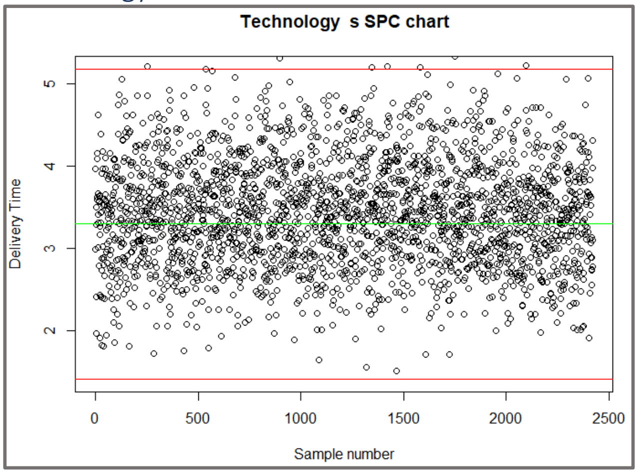


Figure 17: S-Chart for the Technology class.

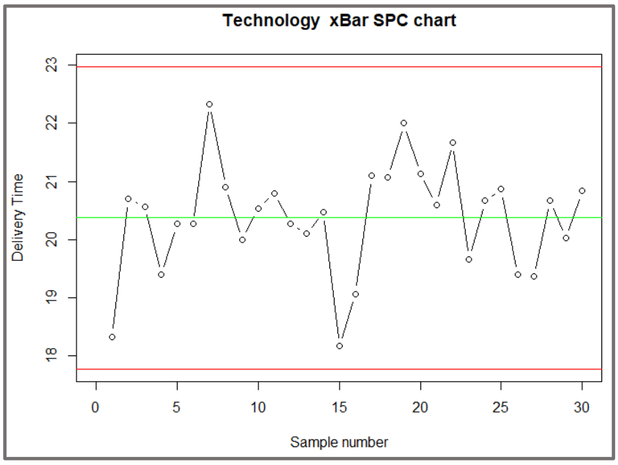


Figure 16: X-Chart for the Technology class, first 30 samples.

The S-chart shows that all most all of the points are in the control limits but there, thus we say the process variation is in control. Then, we see the X-Bar chart has all the points in control which means the process mean is in control.

Sweets

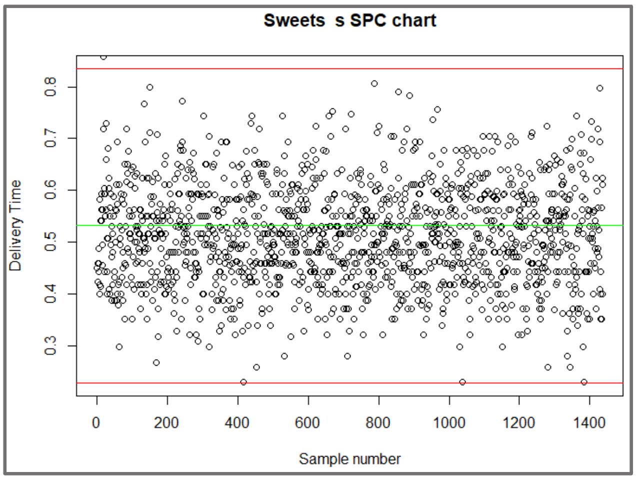


Figure 18: S-Chart for the Sweets class.

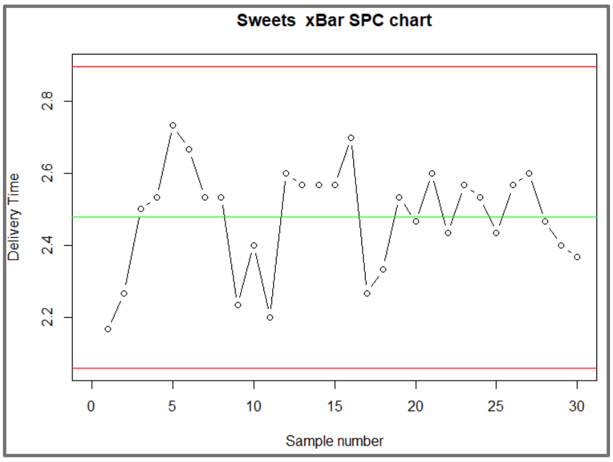


Figure 19: X-Chart for the Sweets class, first 30 samples.

The S-chart shows that all most all of the points are in the control limits but there, thus we say the process variation is in control. Then, we see the X-Bar chart has all the points in control which means the process mean is in control.

Gifts:

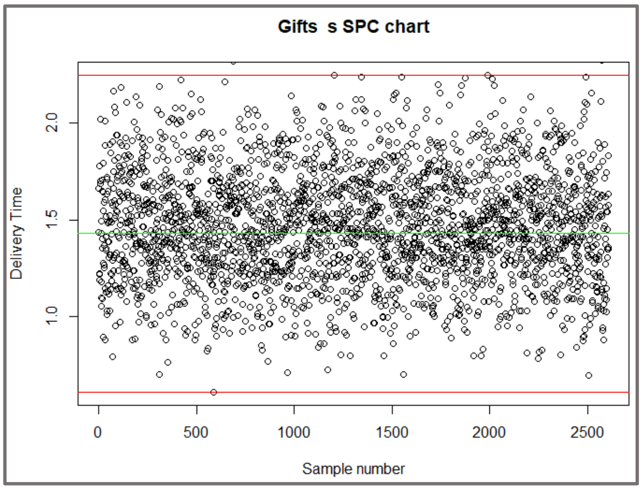


Figure 21: S-Chart for the Gifts class.

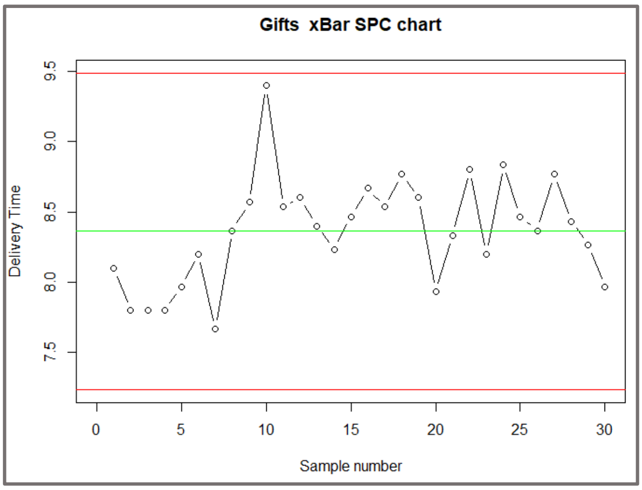


Figure 20: X-Chart for the Gifts class, first 30 samples.

The S-chart shows that all most all of the points are in the control limits but there, thus we say the process variation is in control. Then, we see the X-Bar chart has all the points in control which means the process mean is in control.

Luxury:

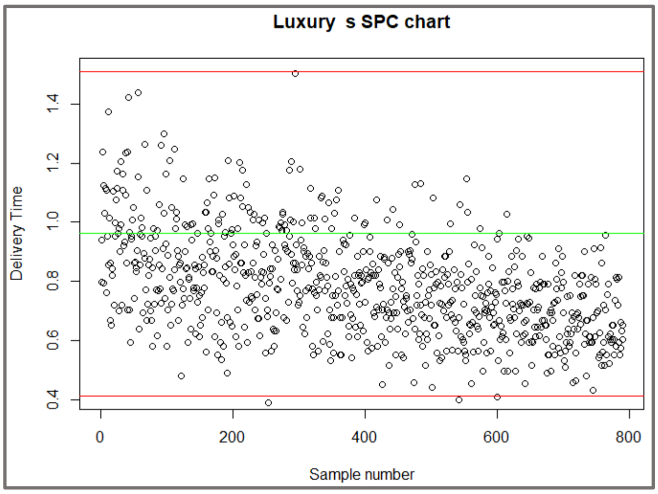


Figure 23: S-Chart for the Luxury class.

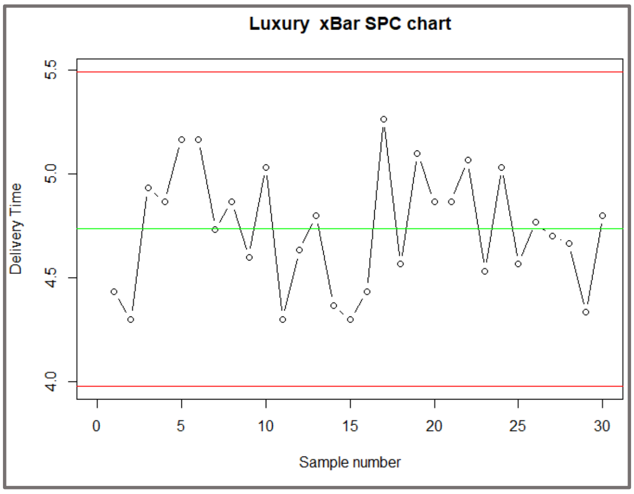


Figure 22: X-Chart for the Luxury class, first 30 samples.

The S-chart shows that all most all of the points are in the control limits but there, thus we say the process variation is in control. Then, we see the X-Bar chart has all the points in control which means the process mean is in control.

Part 4: Optimising the Delivery Process

Sample Means outside control limits:

Here we can see the total number of outliers per class as well as the values of the first three and last three outliers also per class:

Class	Total Found	First	Second	Third	3rd Last	2nd Last	Last
Clothing	15	419	1021	1061	1541	1561	1584
Household	359	138	154	231	1220	1221	1222
Food	3	310	-	-	-	1051	1386
Technology	14	396	562	632	1900	2034	2053
Sweets	1	1185	-	-	-	-	-
Gifts	2091	197	201	202	2391	2392	2393
Luxury	404	131	159	169	724	725	726

Table 5: Table showing summary of outliers.

Looking at the table we can see that the Household, Gifts, and Luxury classes has many subgroups that are out of control and thus there still exists a problem in the process. Clothing, Food Technology, and Sweets show very few out of control values and thus these processes are within expectations.

Consecutive samples of S bar between -0.3 and +0.4 sigma-control limits:

The table below shows the most consecutive samples of “s-bar” between -0.3 and+0.4 sigma-control limits and the ending sample number, where the last sample is in the given range:

Class	Maximum Pattern Length	Last Sample Position
Clothing	14	372
Household	14	830
Food	11	1069
Technology	16	1988
Sweets	15	525
Gifts	23	1485
Luxury	10	268

Table 6: Table showing consecutive samples of the S-Bar chart between -0.3 and +0.4 sigma-control limits.

Part 5: Hypothesis Testing

For Hypothesis testing, tests will be conducted to see if the mean of certain data is the same or different for the two cases. This will be done by using the Manova function on RSTudio. A MANOVA test is used In the situation where there multiple response variables you can test them simultaneously using a multivariate analysis of variance (MANOVA) (STHDA, 2022).

Hypothesis 1:

Testing if the delivery time and sales price is different for all sales classes. For this hypothesis we will say that the mean for each class's delivery time and price is the same.

Ho: Class has no impact on the price or the delivery time of the products.

Ha: Class has an impact on Price as well as the delivery time for products.

ANOVA Table for Hypothesis 1:

Class	n	Delivery Time	Price
Clothing	24266	8.999	640.25
Food	22491	2.502	407.80
Gifts	35900	12.887	2957.31
Household	18332	48.702	10989.34
Luxury	10899	3.972	64798.75
Sweets	19739	2.500	304.32
Technology	33351	20.016	29498.80
P-Value	<0.001		

Table 7: ANOVA Table of the first hypothesis that includes the Delivery Time and Price of each class.

Inspecting from Anova table the P-value is extremely small. This means that the mean for all classes is not the same for delivery time & Price. The null hypothesis was rejected with high certainty seeing as the class of the product has a significant impact on the price as well as the delivery time of the product. The company cannot promise the same delivery time per product & in order to be reliable they should inform customers on the average delivery time per class. The company cannot promise the same service level for every product.

Boxplot of Delivery Time and Price for each class:

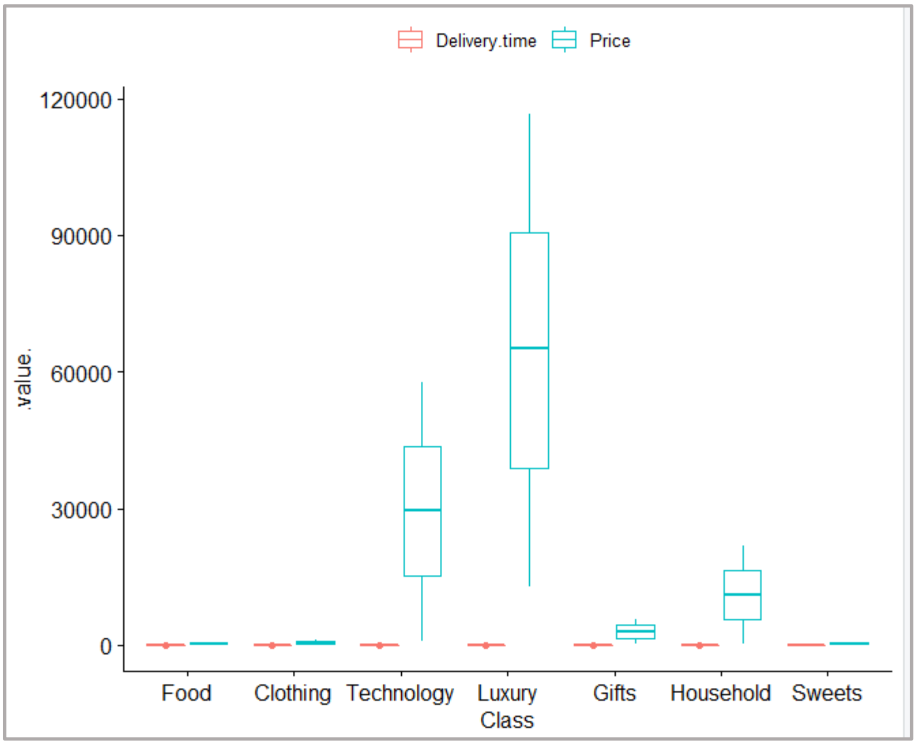


Figure 24: Boxplot showing count of Delivery Time and Prices data for each class.

From the plot it is evident that the mean value for both delivery time & price is different for each class, helping us to confirm our hypothesis test that class has no impact on the price or the delivery time of the products.

Hypothesis 2:

In this hypothesis we will be looking at the Age category to determine if certain ages prefer certain class items or not. This hypothesis will be inspected since the company should determine wheter or not they should solely focus on a certain age group per class.

Ho: Customer age has no impact on the class they are buying products from.

Ha: Customer age has a impact on the class they are buying products from.

ANOVA Table for Hypothesis 2:

Class	**n**	**AGE**
Clothing	24266	48
Food	22491	65
Gifts	35900	60
Household	18332	52
Luxury	10899	51
Sweets	19739	57
Technology	33351	47
P-Value	<0.001	

Table 8: ANOVA Table of the second hypothesis that includes Age of each class.

Inspecting from Anova table the P-value is extremely small. This means that the mean for all classes is not the same for Age. Thus, we reject the null hypothesis that age does not have an impact on the number of products sold

Boxplot of count of ages for each class:

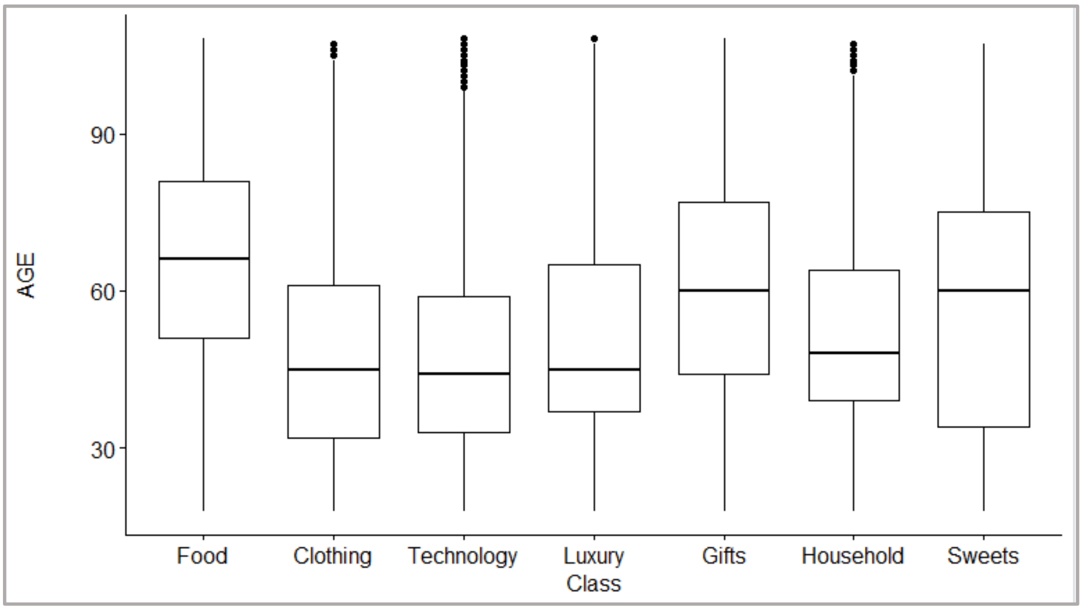


Figure 25: Boxplot showing count of AGE data for each class.

From the boxplot we can see that the mean values of ages are different for each class, helping us to confirm our hypothesis test that customer age does not impact on the class they are buying products from.

Part 6: Reliability of the service and products.

For computing the reliability of the service and products the Taguchi loss function will be used.

"The Taguchi loss function is graphical depiction of loss developed by the Japanese business statistician Genichi Taguchi to describe a phenomenon affecting the value of products produced by a company. Praised by Dr. W. Edwards Deming, it made clear the concept that quality does not suddenly plummet when, for instance, a machinist exceeds a rigid blueprint tolerance. Instead 'loss' in value progressively increases as variation increases from the intended condition. This was considered a breakthrough in describing quality, and helped fuel the continuous improvement movement." (SunLearn)

Taguchi Loss Function

6.1

Food deliveries must be kept cool during transit. The company has a subsidiary, Lafrideradora, who makes components for their units. Problem 6 and 7 of chapter 7 (page 363):

<u>Problem 6:</u>

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:

The Taguchi Loss function: $L = k(y - m)^2 = 45$, when $(y - m) = 0.04$

$k = \frac{45}{0.04^2} = 28125$

Thus, we have: $L = 28125(y - 0.06)^2$

<u>Problem 7:</u>

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

a.

Taguchi loss function: $L = k(y - m)^2 = 45$, when $(y - m) = 0.04$

$k = \frac{35}{0.04^2} = 21875$

Thus, we have: $L = 21875(y - 0.06)^2$

b.

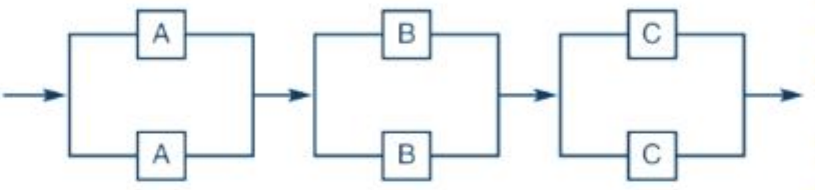
The Taguchi loss: $L = 21875 * 02027^2 = 15.95$

6.2

Another subsidiary, Magnaplex, manufactures some your technology items. Management thinks their current process is wasteful due to the identical machines used as backup in case of failures. Problem 27 of chapter 7:

<u>Problem 27:</u>

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 figure).



Production System

The reliabilities of the machines are as follows:

Machine	Reliability
A	0.85
B	0.92
C	0.90

a. Analyse the system reliability, assuming only one machine at each stage (all the backup machines are out of operation).
b. How much is the reliability improved by having two machines at each stage.

a. A reliability block diagram (RBD) may be used to demonstrate the interconnection between individual components. See below left a RBD for Magnaplex. If only one machine each in A, B and C is working, the reliability is simply: **REPHRASE**

$$R_A * R_A * R_A = 085 * 0.92 * 0.9 = 0.7038$$

b. If both of each machine's A, B and C are working, then the first process reliability will be higher. For parallel connected components A, their combined reliability is 1 – probability(both fail).

   Both A's fail with a probability $(1 - 0.85^2) = 0.0225$ and then their combined reliability is $R_{AA} = 1 - 0.0225 = 0.9775$.

Both B’s fail with a probability $(1 - 0.92^2) = 0.0064$ and then their combined reliability is $R_{BB} = 1 - 0.0064 = 0.9936$.

Both C’s fail with a probability $(1 - 0.90^2) = 0.01$ and then their combined reliability is $R_{CC} = 1 - 0.01 = 0.99$.

The total reliability is: $R_{AA} * R_{BB} * R_{CC} = 0.9775 * 0.9936 * 0.99 = 0.9615$ or we say that the reliability improved by: [(0.9615-0.7038)/0.7038]*100 = 26%.

Conclusion

The analysis of the valid dataset made us aware of the different relationships between different features. These relationships can be vital to the company's owners as they can be aware of the impacts changes have on different criteria of the customers. It has been found that definite changes should be made in the following classes: Household, Gifts, and Luxury. These classes still have variation in the data which makes it more difficult to use, thus, a solution should be thought of.

References

Mathtec. (n.d.). *GRAPH SHAPES*. Retrieved from mathtec.weebly.com: https://mathtec.weebly.com/graph-shapes.html

Mosley, M. (2022, October 20). *Word of Mouth Marketing: 6+ Proven Strategies [2022 Guide]*. Retrieved from referralrock.com: https://referralrock.com/blog/word-of-mouth-marketing/

STHDA. (2022). *MANOVA Test in R: Multivariate Analysis of Variance - Easy Guides - Wiki*. Retrieved from STHDA - Statistical tools for high-throughput data analysis: http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance

SunLearn. (n.d.). *Statistical Methods in Quality Assurance Part 1 summary.* Retrieved from SunLearn: https://learn.sun.ac.za/pluginfile.php/3514418/mod_resource/content/1/QA344%20Statistics.pdf