ECSA GRADUATE ATTRIBUTES PROJECT

Quality Assurance 344
Stellenbosch University

ML Rabie
23573821
21 October 2022

Table of Contents

Introduction **1**

1. Data Wrangling **2**

1.1. Invalid Data *2*

1.2. Valid Data *3*

2. Descriptive Statistics **4**

2.1. Continuous features *4*
Feature Analysis 5

2.2. Categorical Features *6*

2.3. Further Feature Analysis *8*

2.4. Process Capability Indices *9*

3. Statistical Process Control for S-chart and X-chart **10**

3.1. Initialise X&s-charts for Each Class of Sale *10*
3.1.1. X-Charts 10
3.1.2. S-Charts 12

3.2. X&s-charts for the Remaining Instances of Each Class of Sale *14*
3.2.1. Remaining Samples' Graphs and Analysis 14

4. Optimising the Delivery Processes **18**

4.1. Sample Numbers Demonstrating Out of Control *18*
4.1.1.A. Sample Means (X- Bar) 18
4.1.1.B. Most Consecutive Samples Between -0.3 and +0.4 Sigma Control Limits 19

4.2. Type I Error *19*
4.2.1. Likelihood of making a Type I Error for A 19
4.2.2. Likelihood of making a Type I Error for B 20

4.3. Optimising Delivery Times *20*

4.4. Type II Error *21*

5. DOE and MANOVA. **22**

6. Service and Product Reliability **25**

6.1. Problem 6 and 7 (Chapter 7, p.363) *25*
6.1.1. Problem 6 25
6.1.2. Problem 7 25

6.2. Problem 27 (Chapter 7) *26*

6.3. Binomial Probability *26*

7. Conclusion **28**

8. References **29**

Table of Figures

Figure 1: Continuous Features' Distributions 5
Figure 2: Sales per Month 6
Figure 4: Total Revenue per Class 6
Figure 3: Items Purchased per Class 6
Figure 5: Items Sold per Reason 7
Figure 6: Revenue per Reason 7
Figure 8: Age vs Class 8
Figure 10: X-bar Chart for Technology 11
Figure 11: X-bar Chart for Clothing 11
Figure 12: X-bar Chart for Household 11
Figure 13: X-bar Chart for Food 11
Figure 14: X-bar Chart for Sweets 11
Figure 15: X-bar Chart for Gifts 11
Figure 16: X-bar Chart for Luxury 12
Figure 17: S-bar Chart for Technology 13
Figure 18: S-bar Chart for Clothing 13
Figure 19: S-bar Chart for Household 13
Figure 20: S-bar Chart for Food 13
Figure 21: S-bar Chart for Sweets 13
Figure 22: S-bar Chart for Gifts 13
Figure 23: S-bar Chart for Luxury 14
Figure 24: Remaining Samples X-bar Chart for Technology 15
Figure 25: Remaining Samples S-bar Chart for Technology 15
Figure 26: Remaining Samples X-bar Chart for Clothing 15
Figure 27: Remaining Samples S-bar Chart for Clothing 15
Figure 28: Remaining Samples X-bar Chart for Household 15
Figure 29: Remaining Samples S-bar Chart for Household 15
Figure 30: Remaining Samples X-bar Chart for Food 16
Figure 31: Remaining Samples S-bar Chart for Food 16
Figure 32: Remaining Samples X-bar Chart for Sweets 17
Figure 33: Remaining Samples S-bar Chart for Sweets 17
Figure 34: Remaining Samples X-bar Chart for Gifts 17
Figure 35: Remaining Samples S-bar Chart for Gifts 17
Figure 36: Remaining Samples X-bar Chart for Luxury 18
Figure 37: Remaining Samples S-bar Chart for Luxury 18
Figure 38: Total Cost of Delivery Time Reduction per Hour 20
Figure 39: Summary of MANOVA output in R 22
Figure 40: Boxplot of Age per Class 23
Figure 41: Boxplot of Delivery Time per Class 23
Figure 42: Boxplot of Price per Class 24

List of Tables

Table 1: Complete Invalid Dataset 2
Table 2: Valid Dataset 3
Table 3: Valid Dataset Feature Description 4
Table 4: Capability Indices 9
Table 5: X-Chart Control Limits for Delivery Times 10
Table 6: X-bar Graphs 11
Table 7: S-Chart Control Limits for Delivery Times 12
Table 8: Out of Control Sample Means 18
Table 9: Most Consecutive Samples Between the Sigma Control Limits 19
Table 10: Iterative Calculation Conclusions 26

Introduction

Through this report, a data set is analysed to extract useful and meaningful information regarding the sales of a company. This is achieved by first cleaning the data, which involves removing invalid entries. The valid dataset of sales is chronologically ordered, and descriptive statistics methods are applied to the dataset to gain useful insight into the sales of the company. Following this, the Process Capability Indices are calculated to aid in determining the ability of the process to produce within the specification limits. The Delivery Time x-bar and s-bar Statistical Process Control charts are configured, and the graphs are interpreted to reveal trends in the quality control of the various classes, and whether the processes are in- or out-of-control. The delivery times undergo further analysis in order to optimize the delivery process. Aspects of the process, such as the sample means that fall outside of the control limits and the sample standard deviations that consecutively fall between specified sigma value limits, Type I and II Errors, and cost optimization, are discussed in detail. A MANOVA test is performed to determine the impact of “Class” on “Age”, “Price”, and “Delivery Time” of the items. The reliability of the delivery service and products is calculated and discussed.

1. Data Wrangling

The given dataset "salesTable2022" represents entries of the company's clients and includes several features relating to purchase information. The dataset was analysed to identify quality issues and then separated into two datasets, "valid" and "invalid". The "invalid" dataset contains the data instances with missing values or negative prices, whilst the "valid" dataset contains all other instances.

1.1. Invalid Data

As seen in Table 1 below, the invalid data, consisting of all data instances with missing values ("NA") or negative prices were compiled into a dataset. Additional quality issues, such as the large age range, could reflect a data entry issue or could be correct, thus, it was not extracted as invalid data. This dataset is created to remove these instances from the "cleaned" dataset and is not used for any further analysis.

Table 1: Complete Invalid Dataset

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

1.2. Valid Data

The valid data, which has been compiled into the dataset as seen in Table 2, contains all the instances except those deemed invalid as detailed above. The data instances have been chronologically ordered, from oldest to newest. This dataset will be used for further analysis.

Table 2: Valid Dataset

	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
463	463	47101	50	Clothing	1030.86	2021	1	1	9.0	Recommended
2627	2627	88087	21	Clothing	428.03	2021	1	1	10.0	Recommended
3374	3374	25418	68	Household	13184.41	2021	1	1	48.5	Website
5288	5288	13566	94	Household	7021.90	2021	1	1	42.0	Recommended
8182	8182	84692	35	Clothing	475.18	2021	1	1	9.0	Recommended
9272	9272	46305	72	Clothing	580.98	2021	1	1	8.5	Random
9712	9712	92105	45	Household	6877.00	2021	1	1	43.0	Recommended
12163	12163	21614	27	Clothing	513.13	2021	1	1	9.5	Recommended
12195	12195	12174	56	Household	14538.64	2021	1	1	41.5	EMail
20004	20004	84558	74	Food	255.41	2021	1	1	2.0	Recommended
20509	20509	15630	32	Clothing	164.56	2021	1	1	9.0	Recommended
21970	21970	81216	87	Clothing	173.76	2021	1	1	10.0	Recommended
27161	27161	56240	45	Household	17681.94	2021	1	1	45.5	Website
27638	27638	24396	30	Clothing	1018.21	2021	1	1	8.5	Recommended
30778	30778	12235	28	Technology	21096.86	2021	1	1	15.0	Website
34277	34277	30290	43	Household	10573.67	2021	1	1	51.0	Recommended
34950	34950	40035	77	Household	16548.61	2021	1	1	51.5	Recommended
35153	35153	36435	53	Technology	23304.75	2021	1	1	14.0	Browsing
37187	37187	49974	67	Sweets	332.46	2021	1	1	2.5	Recommended
42139	42139	36292	75	Food	205.96	2021	1	1	3.0	Recommended
43139	43139	39202	34	Clothing	353.28	2021	1	1	9.0	Recommended
44379	44379	92277	82	Food	219.76	2021	1	1	2.5	Recommended
45422	45422	12068	56	Gifts	1320.89	2021	1	1	6.5	Website

Showing 1 to 24 of 179,978 entries, 10 total columns

2. Descriptive Statistics

Using statistical analysis of the data, business trends, prediction accuracy and other meaningful insights into the business sales are extracted. The analysis is performed on the continuous and categorical features of only the valid dataset.

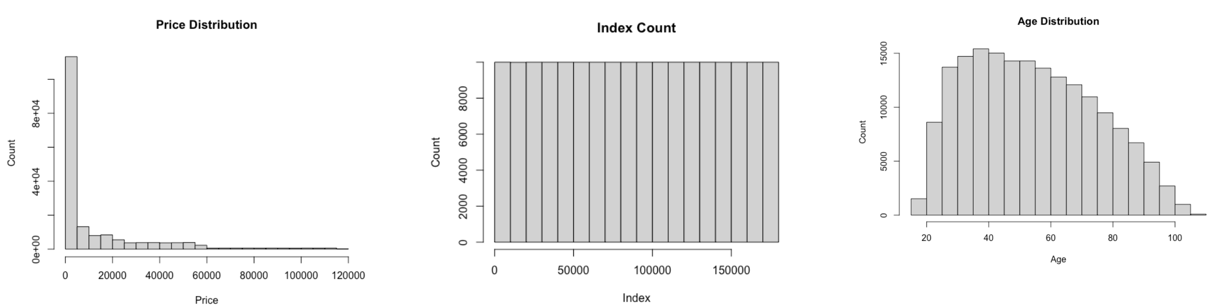
The valid dataset contains 179978 instances with 10 columns, known as features. In Table 3, the different features of the valid dataset are explored and described.

Table 3: Valid Dataset Feature Description

Feature	Description	Feature Type
X	Contains the original "salesTable2022" datafile indexes	Continuous
ID	Contains the unique value assigned to each customer as an identifier	Continuous
AGE	Age of customer	Continuous
Class	Department or category of the sold item	Categorical
Price	Amount the item is purchased for	Continuous
Year	Year in which the item is purchased	Continuous
Month	Month in which the item is purchased	Categorical
Day	Day on which the item is purchased	Continuous
Delivery.time	Number of days taken to deliver the item	Continuous
Why.bought	The client's reason for purchasing the item	Categorical

2.1. Continuous features

The continuous features that were identified include Price, X (Index), AGE, Month, Day, and Delivery Time. Visual representations of each of the continuous features are obtained through the histograms generated and shown below in Figure 1.



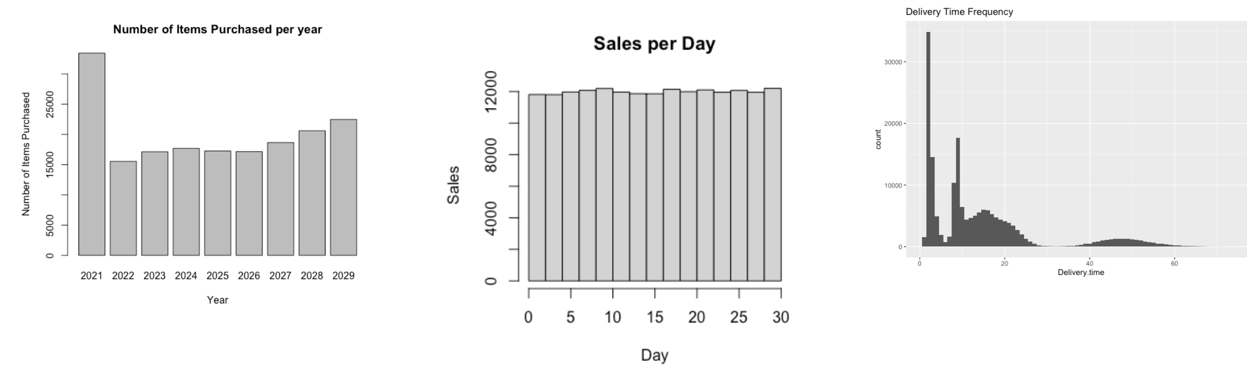


Figure 1: Continuous Features' Distributions

Feature Analysis

Price
The exponential distribution of the data is shown, meaning that more of the cheaper items were purchased.

Age
The age feature shows a skewed-to-the-right distribution, meaning most clients are below 50, with client number tapering off at a gradual decline.

Year
As illustrated on the graph, the most items were purchased in 2021. A sharp decline followed in 2022, after which a consistent increase in sales is seen year after year.

Day
A uniform distribution is shown in the Day feature, meaning that the frequency of sales remains the same throughout the days of the month.

Delivery time
The delivery time feature shows a multimodal distribution containing three peaks.

2.2. Categorical Features

The categorical features identified include Class (of item purchased), Month, and Why.bought (reason for purchase).

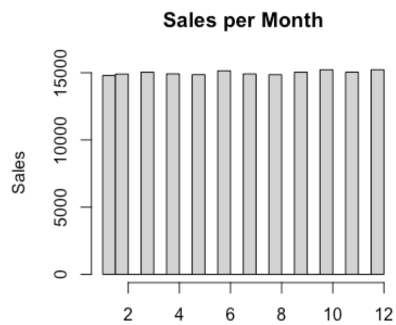


Figure 2: Sales per Month

Month

The month feature shows uniform distribution, meaning that the frequency of sales remains the same throughout the year, thus it is not seasonal.

Figure 4: Items Purchased per Class

Class

As shown in Table 5, "Gifts" and "Technology" have the largest number of items sold; however, as shown in Table 6, "Technology" and "Luxury" earn the largest revenue from sales.

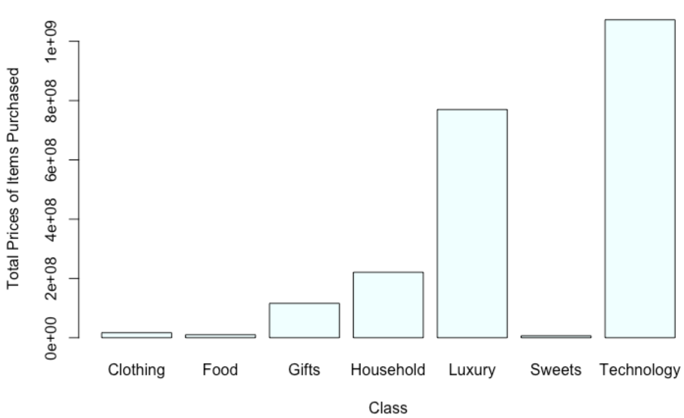


Figure 3: Total Revenue per Class

Figure 5: Items Sold per Reason

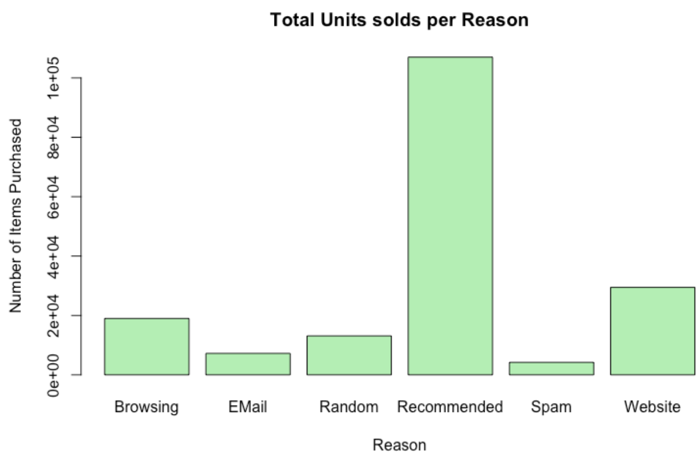
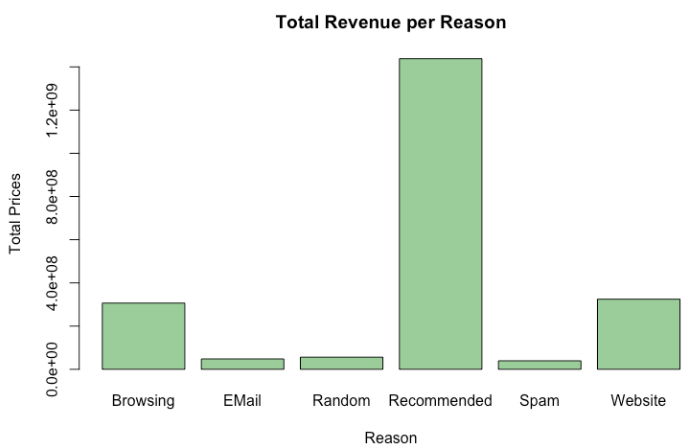


Figure 6: Revenue per Reason



Reason

As shown in Table 7 and 8, “Recommended” is by a large difference the main reason for sales.

2.3. Further Feature Analysis

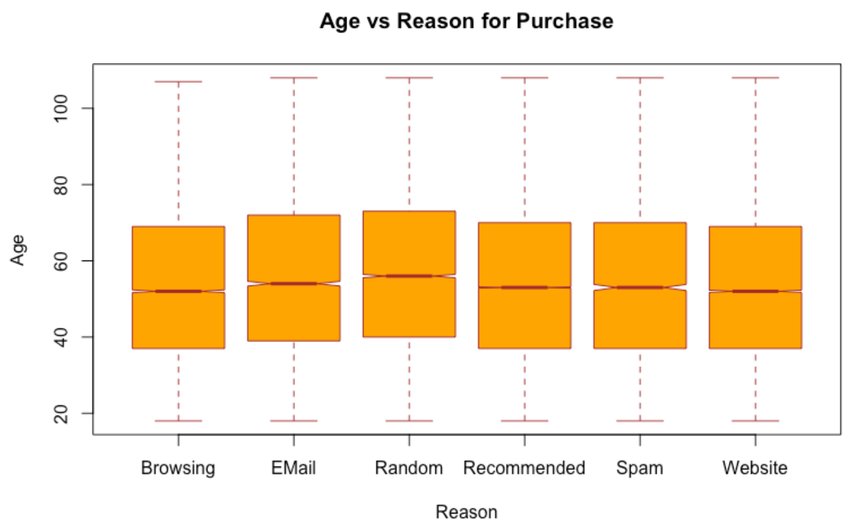


Figure 7: Age vs Reason for Purchase

Table 9 above shows the distribution of age across the different reasons for purchase. This boxplot shows that age is uniformly distributed across the reasons and likely does not have an effect.

Figure 7: Age vs Class

Figure 6 above shows that “Food” is significantly more popular with older demographics and “Clothing” and “Technology” are more popular amongst younger clients.

2.4. Process Capability Indices

Process capability indices aid in determining the ability of a process to produce within certain specification limits.

The following assumptions were made in Process Capability Calculations:
USL (Upper Specification Limit) = 24 days
LSL (Lower Specification Limit) = 0 days

An LSL of 0 makes logical sense as the delivery time cannot be negative (less than 0), as that would imply that the item is delivered before it is ordered, which is not possible. A value of 0 is possible as that would mean that the item is ordered and delivered on the same day.

The mean and standard deviation were obtained by calculations:

$\sigma = 3.501993$

$X = 20.01095$

By using the values mentioned above, the following process capability indices are calculated:

Table 4: Capability Indices

Indices	**C_p**	**C_{pu}**	**C_{pl}**	**C_{pk}**
Formula	$\frac{USL - LSL}{6\sigma}$	$\frac{USL - \bar{X}}{3\sigma}$	$\frac{\overline{X - LSL}}{3\sigma}$	$Min(C_{pu}, C_{pl}$
Value	1.142207	0.3796933	1.90472	0.3796933

The C_{pk} of 0.38 as calculated above signifies that the process is not capable of producing within the specification limits and is not consistent around the average performance. To improve the process capability index, variability within the processes should be decreased.

3. Statistical Process Control for S-chart and X-chart

3.1. Initialise X&s-charts for Each Class of Sale

The valid dataset, “Sales”, is chronologically ordered from the oldest transaction to the newest according to the Year, Month, Day and then X (the original index). This ordered dataset was then saved to dataset, “ordData”. The data is subset for each class, which is further split into 30 samples of 15 instances each, which totals to 450 values. The target feature “Delivery Time” of each of the instances in each class is analysed to produce the SPC charts.

3.1.1. X-Charts

The X-charts are generated to determine outer control limits, the centre line, the 2-sigma-control limits, and the 1-sigma-control limits for the delivery times of each of the seven classes.

X-Chart Table

The respective control limits are summarized in the table below.

Table 5: X-Chart Control Limits for Delivery Times

CLASS	LCL	L2SIGMA	L1SIGMA	CL	U1SIGMA	U2SIGMA	UCL
CLOTHING	8.535066	8.680044	8.825022	8.970000	9.114978	9.259956	9.404934
HOUSEHOLD	42.876117	44.104818	45.333520	46.562222	47.790924	49.019626	50.248328
FOOD	2.270542	2.343695	2.416847	2.490000	2.563153	2.636305	2.709458
TECHNOLOGY	17.774273	18.640997	19.507721	20.374444	21.241168	22.107892	22.974616
SWEETS	2.058514	2.198269	2.338023	2.477778	2.617532	2.757287	2.897042
GIFTS	7.233658	7.609475	7.985293	8.361111	8.736929	9.112747	9.488565
LUXURY	3.977146	3.977146	4.482752	4.735556	4.988359	5.241162	5.493965

X-bar Control Charts

The x-bar graphs in Table 8 visualize the distribution of the first 30 samples with the Centre Line shown in red, the 1-sigma-control limits in yellow, the 2-sigma-control limits in green, and the outer control limits in blue. The area between the blue outer control limits represents the control area.

Table 6: X-bar Graphs

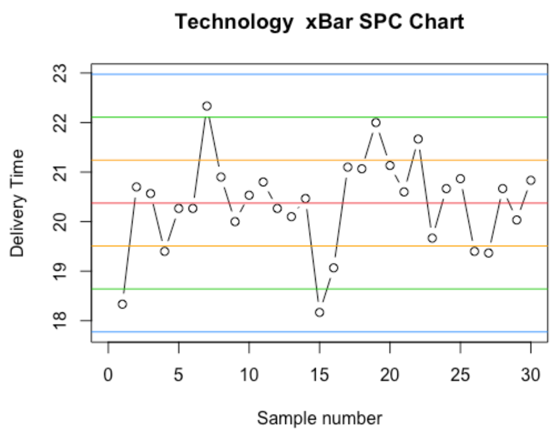


Figure 8: X-bar Chart for Technology

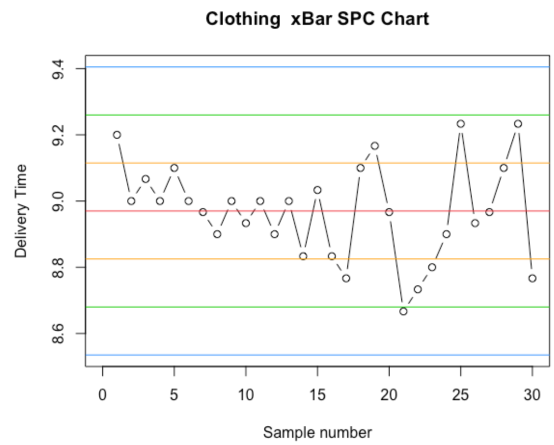


Figure 9: X-bar Chart for Clothing

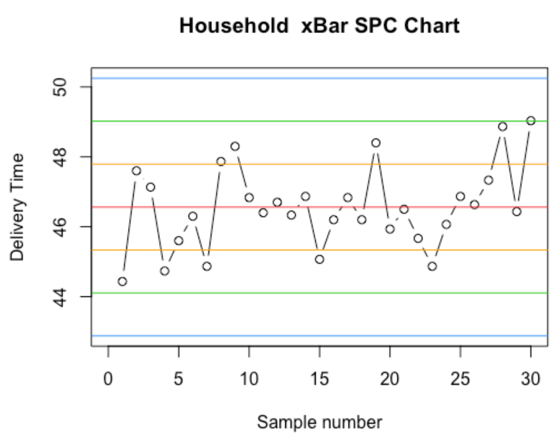


Figure 10: X-bar Chart for Household

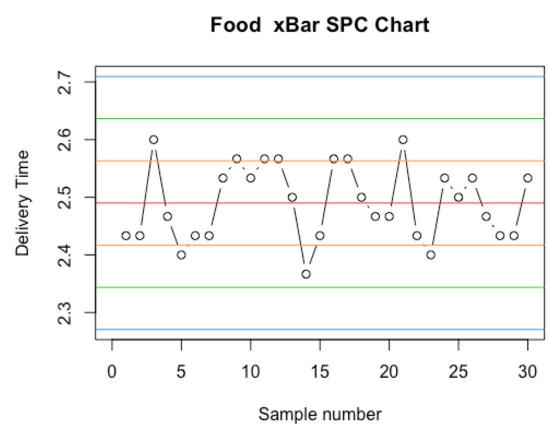


Figure 11: X-bar Chart for Food

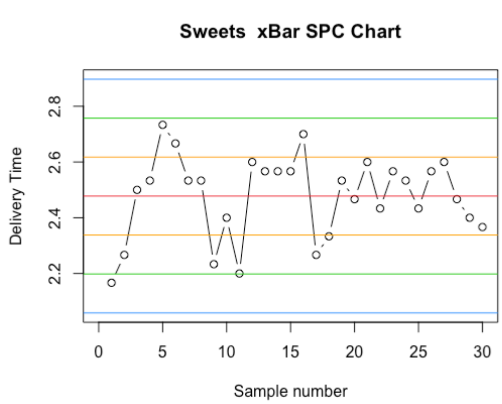


Figure 12: X-bar Chart for Sweets

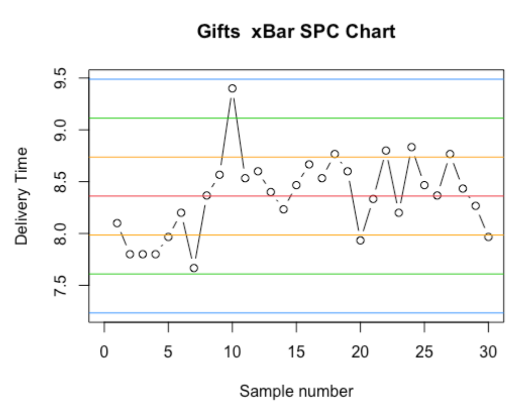


Figure 13: X-bar Chart for Gifts

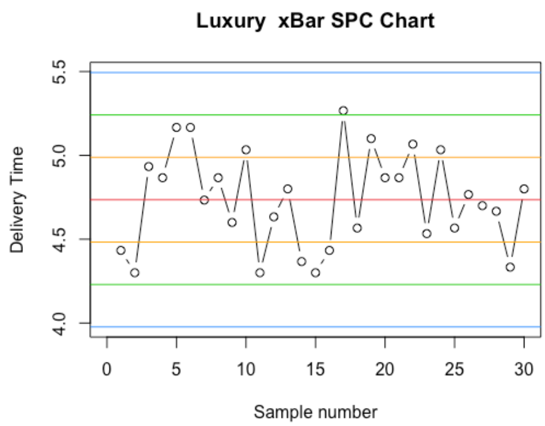


Figure 14: X-bar Chart for Luxury

3.1.2. S-Charts

S-Chart Table

The respective control limits are summarized in the table below.

Table 7: S-Chart Control Limits for Delivery Times

CLASS	LCL	L2SIGMA	L1SIGMA	CL	U1SIGMA	U2SIGMA	UCL
CLOTHING	0.2359335	0.3410379	0.4461422	0.5512465	0.6563509	0.7614552	0.8665596
HOUSEHOLD	1.9995605	2.8903304	3.7811003	4.6718703	5.5626402	6.4534101	7.3441801
FOOD	0.1190468	0.1720801	0.2251134	0.2781467	0.3311800	0.3842133	0.4372466
TECHNOLOGY	1.4104859	2.0388332	2.6671805	3.2955278	3.9238751	4.5522224	5.1805697
SWEETS	0.2274333	0.3287509	0.4300686	0.5313862	0.6327039	0.7340215	0.8353391
GIFTS	0.6115971	0.8840532	1.1565092	1.4289652	1.7014213	1.9738773	2.2463333
LUXURY	0.4114060	0.5946803	0.7779546	0.9612289	1.1445032	1.3277775	1.5110518

S-bar Control Charts

Similarly, to the x-bar graphs in Table 8, the s-bar graphs below visualize the distribution of the first 30 samples with the Centre Line shown in red, the 1-sigma-control limits in yellow, the 2-sigma-control limits in green, and the outer control limits in blue. The area between the blue outer control limits represents the control area.

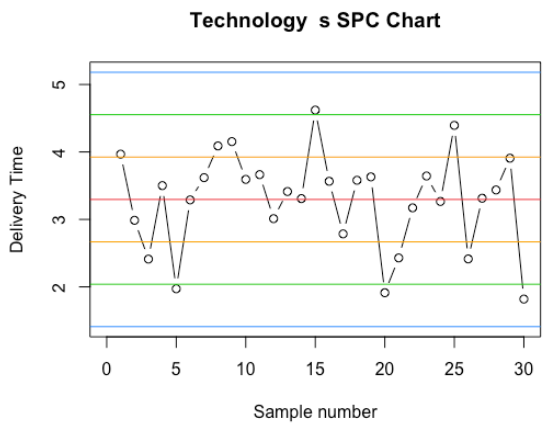


Figure 15: S-bar Chart for Technology

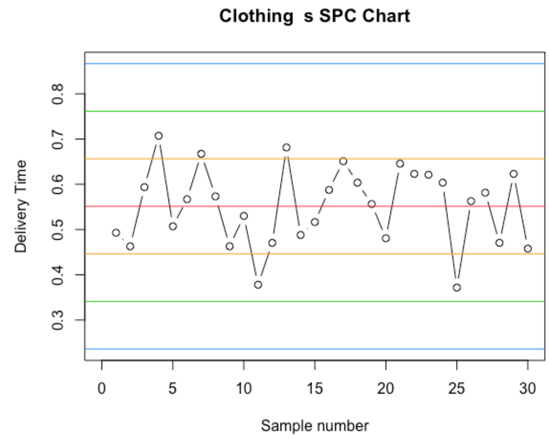


Figure 16: S-bar Chart for Clothing

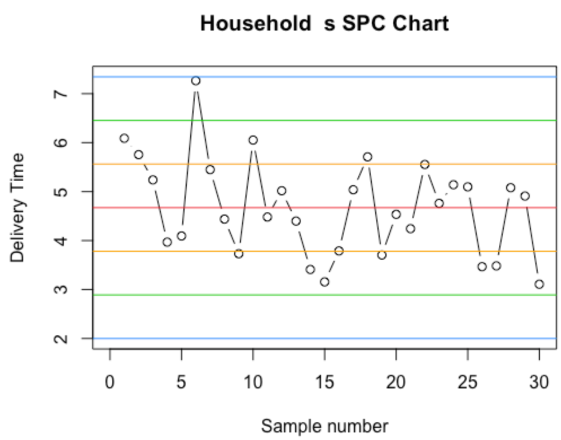


Figure 17: S-bar Chart for Household

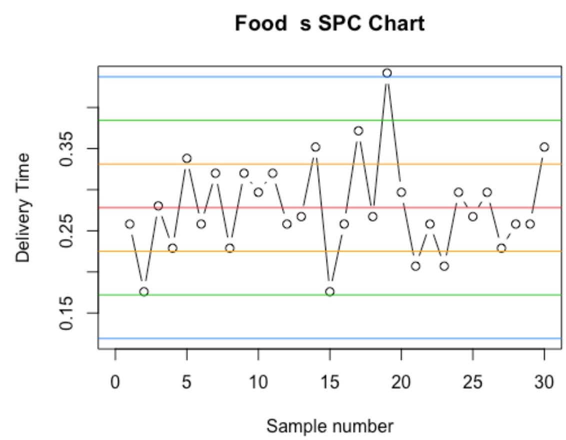


Figure 18: S-bar Chart for Food

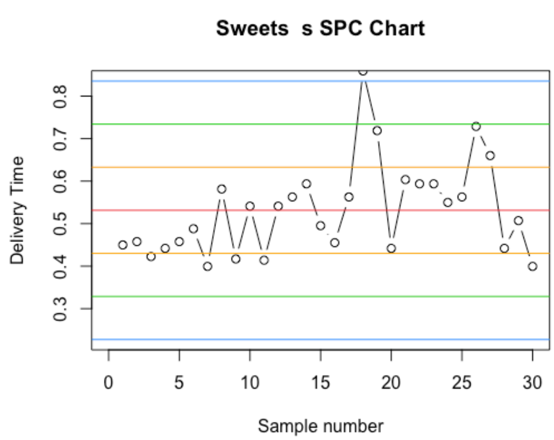


Figure 19: S-bar Chart for Sweets

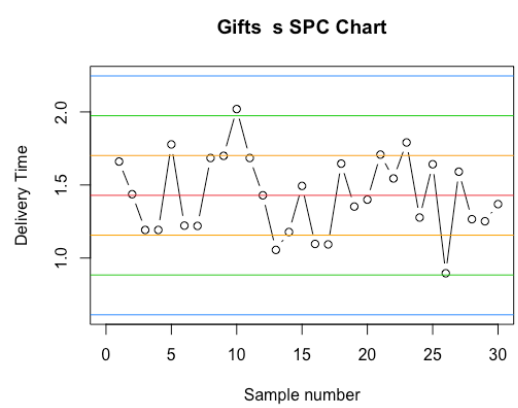


Figure 20: S-bar Chart for Gifts

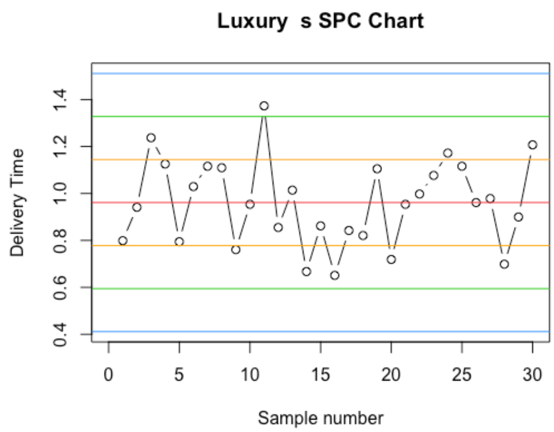


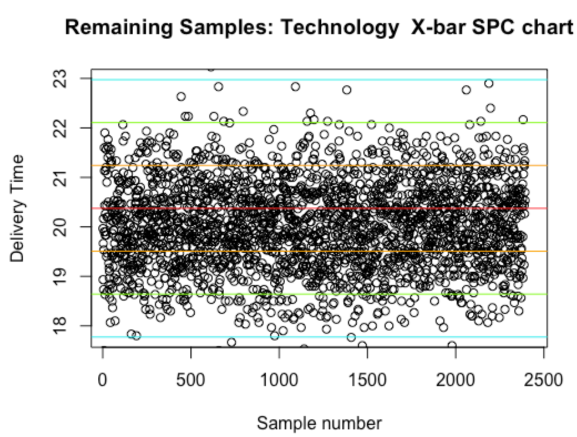
Figure 21: S-bar Chart for Luxury

3.2. X&s-charts for the Remaining Instances of Each Class of Sale

The SPC charts are configured for the delivery times of the remaining samples in each class. Each of the control lines are shown in their respective colours, thus revealing instances that fall outside of the blue outer control limits. The visualization of the data reveals trends in the quality control of the delivery times, with the area between the blue lines representing the in-control area.

3.2.1. Remaining Samples' Graphs and Analysis

Technology



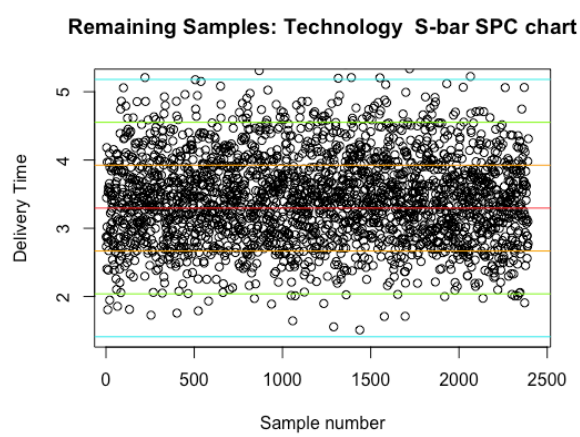


Figure 22: Remaining Samples X-bar Chart for Technology

Figure 23: Remaining Samples S-bar Chart for Technology

As shown in Figure 24 and Figure 25, the respective x-bar and s-bar instances are condensed around the centre line, with few instances outside of the outer control limits and no substantial upwards or downwards trend. From this information, it can be inferred that the process is in control.

Clothing

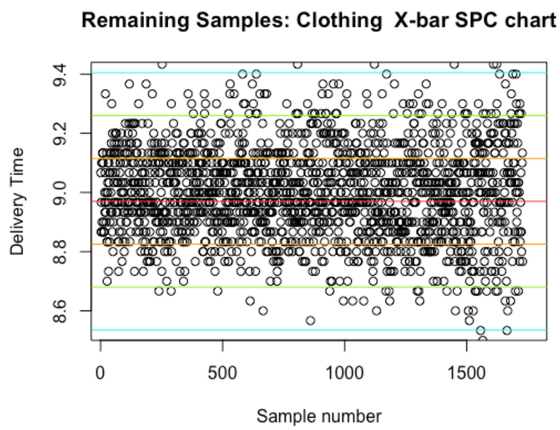


Figure 24: Remaining Samples X-bar Chart for Clothing

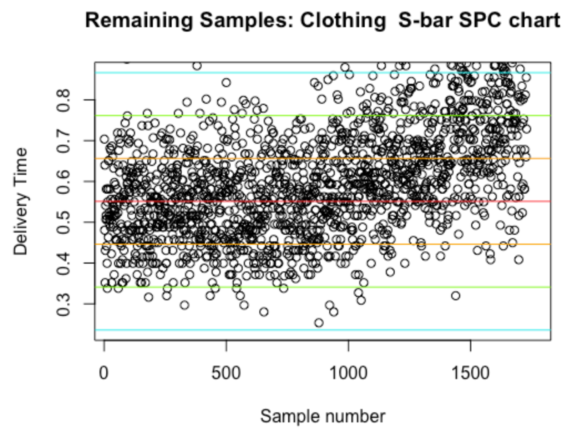


Figure 25: Remaining Samples S-bar Chart for Clothing

As shown by the higher density of instances around the centre line and relatively few instances outside of the control limits in Figure 26, the x-bar chart for Clothing is in control; however, the graph reveals a developing out-of-control trend, with instances spreading out from the centre line and more instances falling outside of the control limits. This is verified by the upwards trend present in the s-bar chart shown in Figure 27, which represents a greater degree of variation in the delivery times.

Household

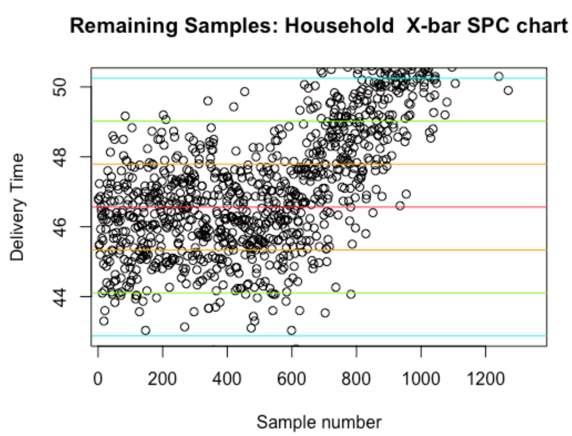


Figure 26: Remaining Samples X-bar Chart for Household

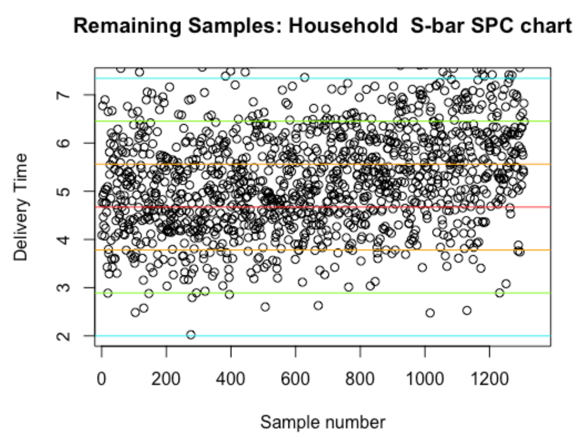


Figure 27: Remaining Samples S-bar Chart for Household

Figure 28 illustrates the x-bar of the process rapidly moving out of control, with the whole process remaining in a similar pattern that moves upwards, beyond the upper control limit. The s-bar chart in Figure 29 remains relatively in control, with an upwards trend that moves further out of upper control limit, which corresponds with the consistent density pattern in Figure 28. These developments in the graphs indicate that there could be specific issue or change to the process that has resulted in the upwards trend (longer delivery times), which could be fixed to return the process to within the control limits.

Food

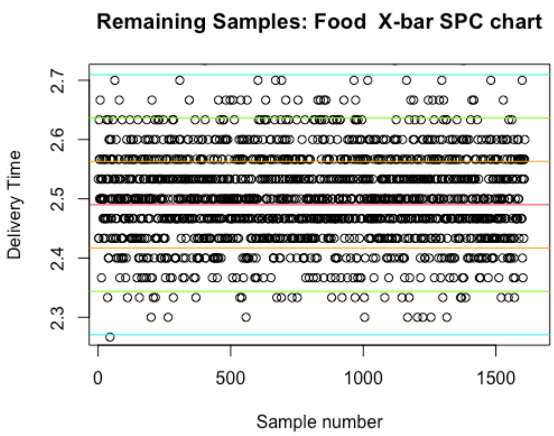


Figure 28: Remaining Samples X-bar Chart for Food

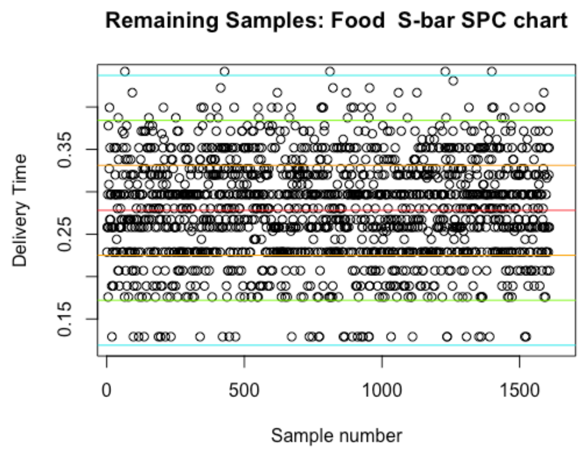


Figure 29: Remaining Samples S-bar Chart for Food

Figure 30 and Figure 31 shows that the process for Food is completely in control, with a negligible number of instances outside of the control limits for both the x-bar and s-bar charts. The distribution of the x-bar instances between the UCL and LCL is explained by the corresponding distribution of the s-bar instances between its outer control limits. Measures could be implemented to reduce the variation in standard deviation to condense the s-bar and x-bar instances around the centre lines; however, the process is currently in control, so a cost-benefit analysis would be required to determine the feasibility of such a project.

Sweets

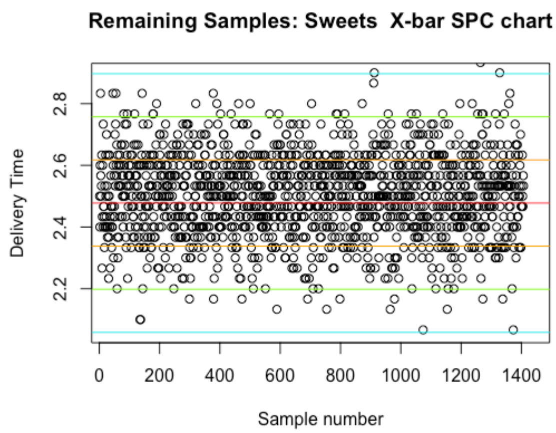


Figure 30: Remaining Samples X-bar Chart for Sweets

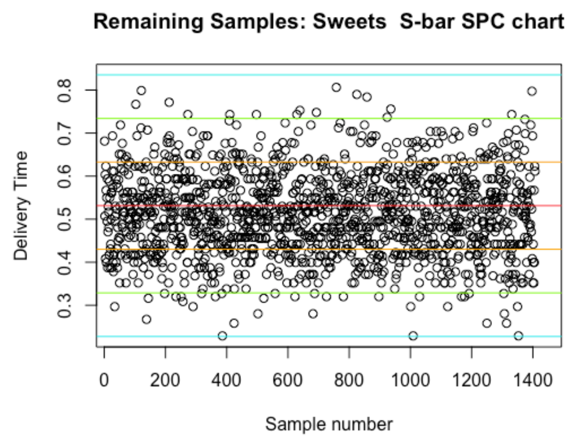


Figure 31: Remaining Samples S-bar Chart for Sweets

As shown in Figure 32 and Figure 33, the respective x-bar and s-bar instances are relatively condensed around the centre line, with negligible instances outside of the outer control limits and no substantial upwards or downwards trend, which reveals that the process is in control.

Gifts

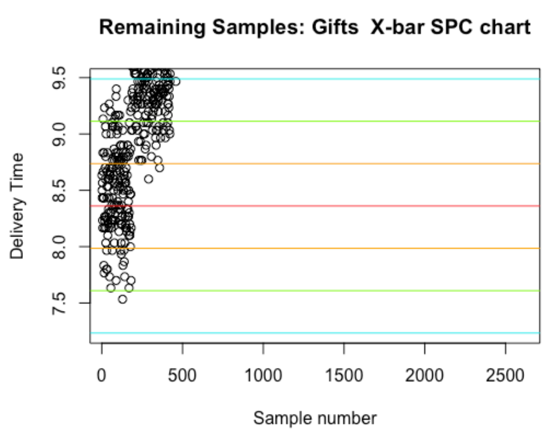


Figure 32: Remaining Samples X-bar Chart for Gifts

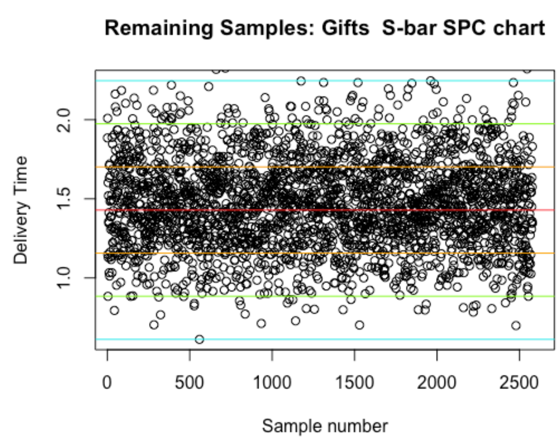


Figure 33: Remaining Samples S-bar Chart for Gifts

As shown in Figure 35, the s-bar instances are mostly within the control limits and are well-distributed, with majority of the instances condensed around the centre line, showing that the standard deviations of the Gifts' Delivery Times are in control. Figure 34, however, shows that the process is completely out of control as the x-bar trends upwards with a dramatic gradient. This shift in the delivery times needs to be further analysed to determine the cause, which will determine whether a solution can be found to return the process to within its control limits or, if no solution can be found (the cause is external and no alternative is available), new control limits need to be set.

Luxury

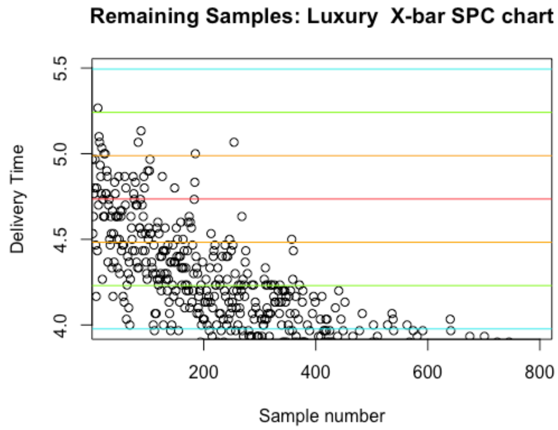


Figure 34: Remaining Samples X-bar Chart for Luxury

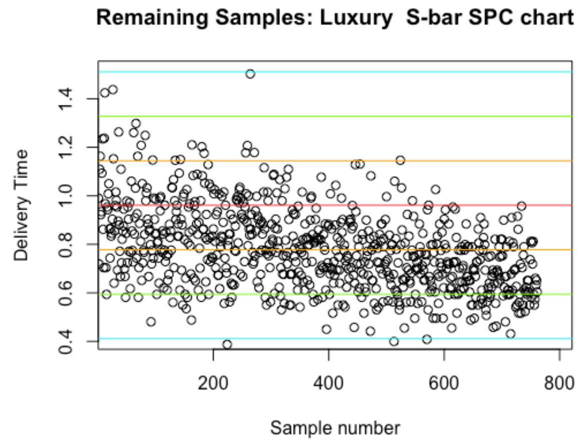


Figure 35: Remaining Samples S-bar Chart for Luxury

Figure 36 and Figure 37 reveal that the process is initially in control, with a downwards trend that shifts the x-bar chart below and s-bar chart towards the LCL. This process appears out of control, but this trend is desirable as it shows a decrease in mean delivery time and a decrease in variation. The cause of this reduction should be identified to potentially replicate in other classes.

4. Optimising the Delivery Processes

All available valid data is used in this section. Through careful analysis, insight into the current state of the delivery process can be obtained in order to locate areas where optimization is necessary and will be most impactful.

4.1. Sample Numbers Demonstrating Out of Control

4.1.1.A. Sample Means (X- Bar)

As tabulated in Table 8 below, the indices of the first three and last three delivery time x-bar samples, as well as the total number of samples, that fall outside of the control limits (above and below the UCL and LCL, respectively) for each class.

Table 8: Out of Control Sample Means

CLASS	1ST	2ND	3RD	3RD LAST	2ND LAST	LAST	TOTAL NO. OF INSTANCES
CLOTHING	282	837	1048	1695	1723	1756	17
HOUSEHOLD	252	387	643	1335	1336	1337	395
FOOD	75	633	NA	NA	1149	1408	5
TECHNOLOGY	37	345	353	1933	2009	2071	19
SWEETS	942	1243	NA	NA	1294	1358	4

GIFTS	213	216	218	2607	2608	2609	2287
LUXURY	142	171	184	789	789	789	440

As seen above, “Household”, “Gifts”, and “Luxury” have a significantly larger number of instances outside of the control limits, which correspond with the x-bar and s-bar graphs seen in Section 3.2. that determines that these processes are out of control. It is noted, however, that the out-of-control trend in “Luxury” is desirable in this context as it moves below the LCL.

4.1.1.B. Most Consecutive Samples Between -0.3 and +0.4 Sigma Control Limits

As shown in Table 9 below, the maximum number of consecutive numbers in each class between the -0.3 and +0.4 sigma control limits, as well as the value of the last number in the respective sequence.

Table 9: Most Consecutive Samples Between the Sigma Control Limits

CLASS	NO. OF CONSECUTIVE SAMPLES	LAST NO. IN SEQUENCE
CLOTHING	8	578
HOUSEHOLD	10	214
FOOD	8	746
TECHNOLOGY	9	199
SWEETS	6	284
GIFTS	9	313
LUXURY	4	26

4.2. Type I Error

In this section, the likelihood of making a Type I (Manufacturer's) Error for part A and B in section 4.1. above is estimated. A Type I Error occurs if an investigator rejects a null hypothesis that is actually true in the population — the error is a false-positive (Amitav Banerjee, 2009). A Type I Error is a theoretical value that holds true for each process and can thus be calculated independently of the application.

4.2.1. Likelihood of making a Type I Error for A

Following the normal distribution curve, the upper and lower control limits are +3 σ and -3 σ, respectively, away from the centre line. The likelihood of making a Type I Error is determined by calculating the probability of falling between the two Z values of +3 and -3.

$$\begin{aligned} P(Type\ I\ Error) &= pnorm(-3) \times 2 \\ &= 0.002699796 \\ &= 0.2699796\% \end{aligned}$$

The probability is relatively low, suggesting that it is not likely that a Type I Error will occur, and the associated risk is acceptable without much concern on the company’s behalf.

4.2.2. Likelihood of making a Type I Error for B

The upper and lower control limits for B are +0.4 σ and -0.3 σ, respectively, away from the centre line. The likelihood of making a Type I Error for the Household class is determined by calculating the probability of falling between the two Z values of +0.4 and -0.3.

$$\begin{aligned} P(Type\ I\ Error) &= (pnorm(0.4) - pnorm(-0.3))^{10} \\ &= 2.32768e - 06 \\ &= 0.000232768\% \end{aligned}$$

The likelihood of a Type I error occurring for part B is negligible, which suggests that the associated risk is insignificant.

4.3. Optimising Delivery Times

Additional information is given regarding the delivery times of the Technology class, including a penalty cost of R329 per item per hour delivered slower than 26 hours, and a reduction cost of R2.50 per hour that the mean delivery time is reduced by. A reduced mean delivery time is calculated around which the process should be cantered in order to optimize profits.

The initial mean delivery time is 20.01095 hours. This incurs a penalty cost of R758 674 for late deliveries. By calculating and plotting, as seen in Figure 38 below, the total of the penalty and reduction costs for each hour the delivery time of each item is reduced by, it can be gathered that a reduction of 3 hours is optimal, beyond which, the total reduction cost offsets any savings in late delivery penalty costs.

Figure 36: Total Cost of Delivery Time Reduction per Hour

After a reduction of 3 hours, the new mean of 17.01095 hours is calculated.

4.4. Type II Error

In this section, the likelihood of making a Type II Error for the Technology class in part A of section 4.1. is estimated. A Type II (Consumer's) Error occurs if the investigator fails to reject a null hypothesis that is actually false in the population — the error is a false-negative (Amitav Banerjee, 2009). The acceptance of subpar quality can have severe repercussions; thus, companies must ensure that preventing Type II Errors is a high priority. Given that the delivery process average for class "Technology" moves to 23 hours, the likelihood of making a Type II Error between the Upper Control Limit and Lower Control Limit is as calculated below.

$$
\begin{aligned}
P(Type\ II\ Error) &= pnorm(UCL, \mu, \sigma\ ) - pnorm(LCL, \mu, \sigma) \\
&= 0.4955224 - 0.01042696 \\
&= 0.4850954 \\
&= 48.50954\%
\end{aligned}
$$

The probability is quite high, consequently the associated risk of committing a Type II Error is elevated. A recommendation of prioritizing lowering this risk would be appropriate.

5. DOE and MANOVA.

MANOVA is a technique which determines the effects of several independent categorical variables on multiple continuous dependent variables (Anon., 2022).

The MANOVA was performed on the “Class” categorical feature, which specifies the reason for purchase by the customer, and the continuous features “Age”, “Price” and “Delivery Time” to determine their influence on customers’ purchasing specific classes of items.

The following hypotheses were initialized:

Null Hypothesis	The features “Age”, “Price” and “Delivery Time” do not significantly impact the class of items purchased by customers.
Alternative Hypothesis:	At least one feature has a significant impact on the class of item purchased.

The MANOVA results are obtained when comparing “Price” and “Delivery Time” to the “Class” feature and are summarized as follows:

```
 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 37: Summary of MANOVA output in R

As seen in Figure 39, a p-value of 2.2×10^{-16} is obtained, which reveals that there is strong evidence in favour of the alternative hypothesis and that a significant difference is observed (Beers, 2022).

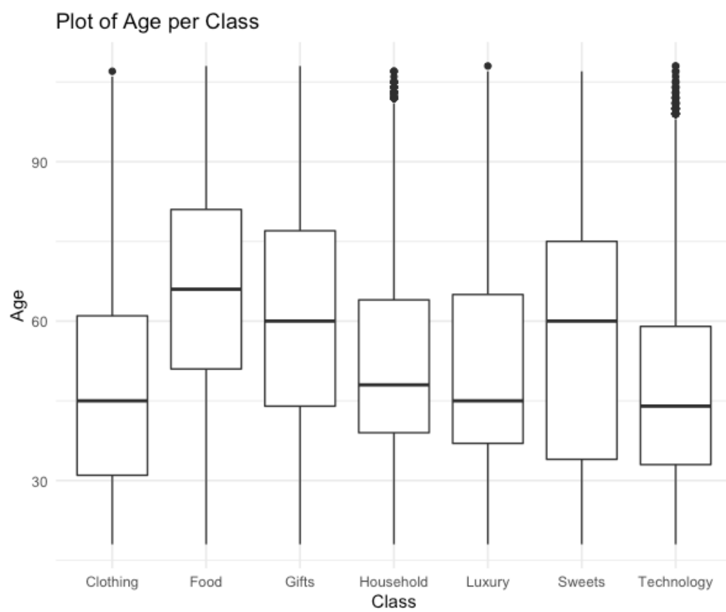


Figure 38: Boxplot of Age per Class

Figure 40 illustrates that Age does influence the class of item purchased; however, the influence is not severe.

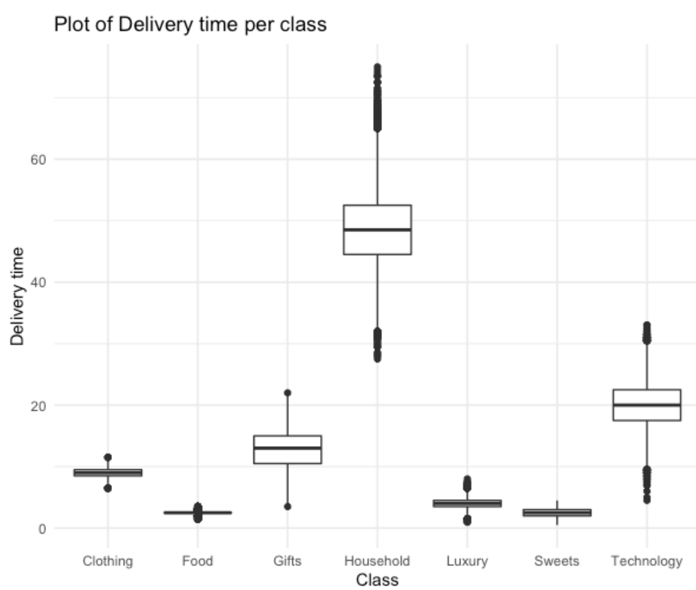


Figure 39: Boxplot of Delivery Time per Class

As shown in Figure 41, Delivery Time is significantly impacted by Class, with the greatest disparity being between the "Household" and "Food"/"Sweets" classes. This may be due to the perishable nature of the "Food" and "Sweets" classes, and the potential difficulty in transporting large household items, such as furniture.

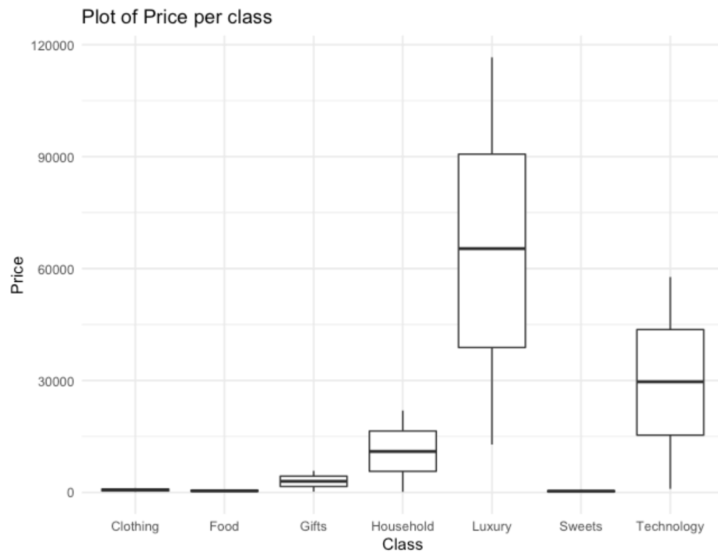


Figure 40: Boxplot of Price per Class

Figure 42 reveals that Price is significantly influenced by the class of the item, with the greatest difference found in the “Luxury” and “Technology” classes.

6. Service and Product Reliability

6.1. Problem 6 and 7 (Chapter 7, p.363)

6.1.1. Problem 6

A blueprint specification for the thickness of a refrigerator part at Lafrigeradora, 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.

Thickness = 0.06 ± 0.04 cm
Cost= $45

Taguchi loss function: $L(x) = k(x - T)^2$

Where:

$$L(x) = 45$$
$$\mathrm{T} = 0.06$$
$$(x - T) = 0.04$$

Solve for k:

$$45 = k(0.04)^2$$
$$k = 28125$$

Therefore the complete Taguchi loss function is as follows:

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

6.1.2. Problem 7

A team was formed to study the refrigerator part at Lafrigeradora 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) $k = \frac{35}{(0.04)^2} = 21875$

With a resulting Taguchi loss function of:
$L(x) = 21875(x - 0.06)^2$

b) $L(x) = 21875(0.027)^2 = 15.95$

6.2. Problem 27 (Chapter 7)

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.
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) $Reliability = R_A \times R_B \times R_C$
$= 0.85 \times 0.92 \times 0.90$
$= 0.7038$
$= 70.38\%$

b) $Reliability = [1-(1-R_A)\times(1-R_A)]\times[1-(1-R_B)\times(1-R_B)]\times[1-(1-R_C)\times(1-R_C)]$
$=[1-(1-0.85)\times(1-0.85)]\times[1-(1-0.92)\times(1-0.92)]\times[1-(1-0.90)$
$\times(1-0.90)]$
$=(0.9775)(0.9936)(0.99)$
$=0.9615315$

6.3. Binomial Probability

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 were only 20 vehicles available was 190 days, only 19 vehicles available was 22 days, only 18 vehicles available was 3 days and 17 vehicles available only once.
There are also 21 drivers, who each work an 8-hour shift per day. During the past 1560 days, the number of days that there were only 20 drivers available was 95 days, only 19 drivers available was 6 days and only 18 drivers available, once only.

Estimate on how many days per year we should expect reliable delivery times, given the information above. If we increased our number of vehicles by one to 21, how many days per year we should expect reliable delivery times?

Through iterative calculations using the **dbinom()** function in R, the following may be determined:

Table 10: Iterative Calculation Conclusions

		Output
Probability of Reliable Number of Vehicles:		0.9278192
Probability of Reliable Number of Drivers:		0.9471258
P(Reliability) — Probability of Reliability:	P(Vehicles) * P(Drivers) =	0.8787615
		87.88%
Expected Reliable Days:	87.87615% * 365 =	319.869

		319 Days

As detailed above, 319 out of 365 days per year are expected to have reliable delivery.

With the addition of a vehicle, increasing the total from 21 to 22, the following information is determined using the iterative calculations shown in Table 10:

		Output
Probability of Reliable Number of Vehicles:		0.96018
Probability of Reliable Number of Drivers:		0.94713
P(Reliability) — Probability of Reliability:	P(Vehicles) * P(Drivers) =	0.90941
		90.94%
Expected Reliable Days:	90.941% * 365 =	331.0247
		331 Days

The addition of a vehicle increased the delivery reliability considerably, with a *Probability of Reliability* increase of 3.06% and an increase of 12 days in *Expected Reliable Days.* Based on these findings, adding a vehicle to the available fleet proves impactful to the quality of delivery; however, an analysis of the cost implications would be required to determine the feasibility of the project.

7. Conclusion

Through this report, meaningful information regarding the sales of a company is extracted from a dataset. The target feature of this report is the Delivery Time of the various classes. Quality issues in the dataset, such as negative and missing values, are addressed and the data is chronologically ordered from oldest to newest sales. This allows descriptive statistics methods to be applied to the dataset to gain useful insight into the sales of the company. It is determined that “Gifts” has the largest volume of sales, “Technology” returns the largest revenue, and recommendation is the leading reason for customers’ purchases. Following this, the Process Capability Indices are calculated to aid in determining the ability of the process to produce within the specification limits. The Delivery Time x-bar and s-bar Statistical Process Control charts are configured to reveal trends in the quality control of the various classes. It is concluded that the delivery processes for the classes “Household”, “Gifts”, and “Luxury” are out of control; however, the delivery times of items in the “Luxury” are reduced by this shift in the process. It is recommended to identify the reason behind this reduction and to explore a potential implementation of the findings on the other classes. The delivery times undergo further analysis to optimize the delivery process. The sample means that fall outside of the control limits and the sample standard deviations that consecutively fall between specified sigma value limits, Type I and II Errors, and cost optimization are determined. A MANOVA test is performed to determine the impact of “Class” on “Age”, “Price”, and “Delivery Time” of the items, revealing that “Price” and “Delivery Time” are significantly influenced by “Class”. The reliability of the delivery service and products is calculated, and it is concluded that the addition of a vehicle increases the delivery reliability considerably and is thus recommended. The valuable information obtained in this report using the methods described above emphasizes the significance of comprehensive quality control measures.

8. References

Amitav Banerjee, U. B. C. S. L. J. J. S. B. a. S. C., 2009. Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal.*

Anon., 2022. *Multiple analysis of variance (MANOVA).* [Online]
Available at: https://www.statistics.com/glossary/multiple-analysis-of-variance-manova/

Beers, B., 2022. *P-Value: What It Is, How to Calculate It, and Why It Matters.* [Online]
Available at: https://www.investopedia.com/terms/p/p-value.asp