ECSA Report
Quality Assurance 354

By Nabeela Parkar – 23622024

October 2022

Abstract

Data analysis of a company's sales data set is presented. The data was cleaned, prepared, and analysed to find correlations and relationships between certain features. Descriptive statistics for various features in the data set are presented and interpreted. Statistical process control metrics have also been computed and presented. A Multivariate ANOVA has also been carried out to check for dependencies between features within the data set. Lastly, some binomial distribution problems are presented, based on service and reliability.

Table of Contents

Table of Figures

Introduction

In this report, data analysis on the sales data set of a particular company is presented and discussed. The data set contains various continuous and categorical features relating to each sale. The data was first cleaned and prepared, analysed to find patterns and correlations between certain features, and to find potential problems that the data indicates. Descriptive statistics for various features in the data set have been presented and interpreted. Statistical process control metrics have also computed and presented for delivery times of the sales in the data set. Multivariate ANOVA has been carried out to check for dependencies between dependent and independent features within the data set. Lastly, some binomial distribution questions have been completed, relating to service and reliability.

Part 1: Data Wrangling

To carry out data analysis on the sales data set, the first important step is to ensure that clean data is used. This means ensuring that the data is free from data quality issues such as missing values, and values that do not make sense, which could be due to incorrect data inputting or human error.

To obtain a cleaned data set containing only valid data, all instances containing missing values were removed from the data set as well as instances containing negative values for features which should be strictly positive, for example price and delivery time. Once these invalid instances were removed, the valid data set was saved to a new file and used to carry out further data analysis.

Part 2: Descriptive Statistics and process capability metrics

The table below shows a summary of descriptive statistics for some of the continuous features within the sales data set.

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
AGE	179978	54.566	20.389	18	38	70	108
Price	179978	12294.098	20889.15	35.65	482.31	15270.97	116618.97
Year	179978	2024.855	2.783	2021	2022	2027	2029
Month	179978	6.521	3.454	1	4	10	12
Day	179978	15.539	8.649	1	8	23	30

Price

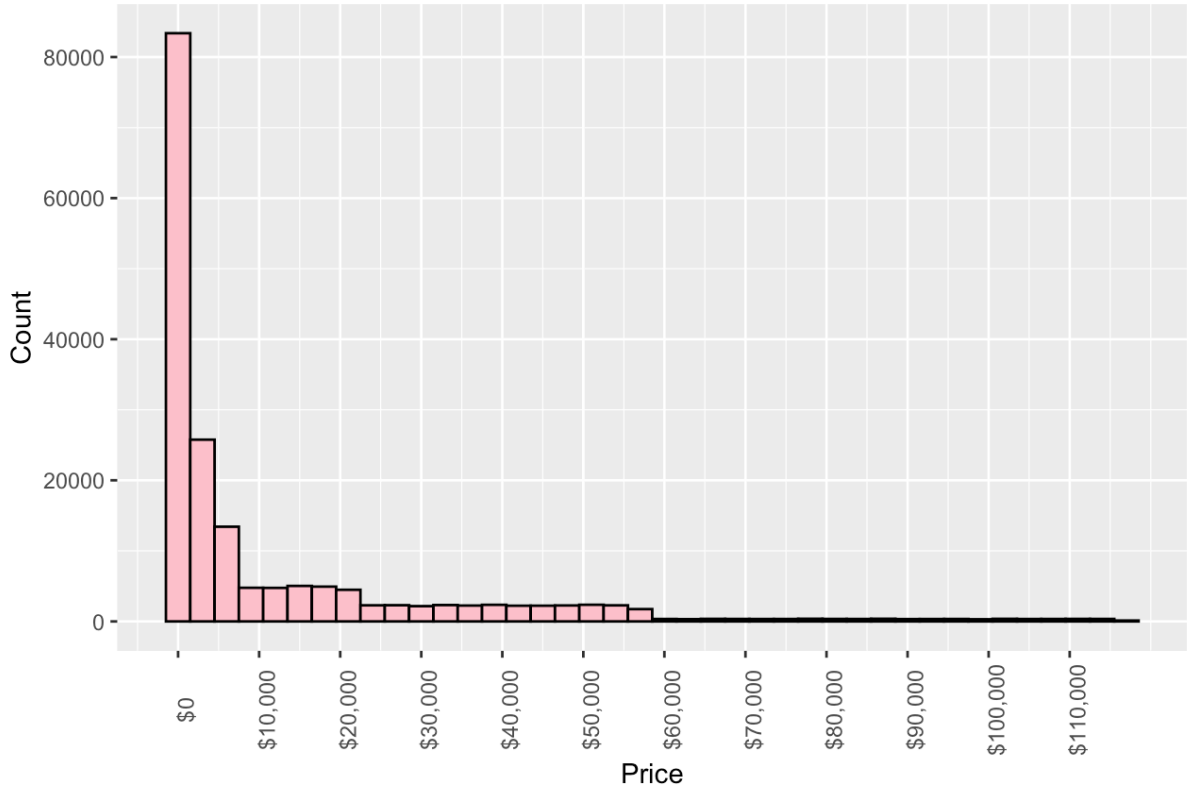

Figure 1: Number of sales by price

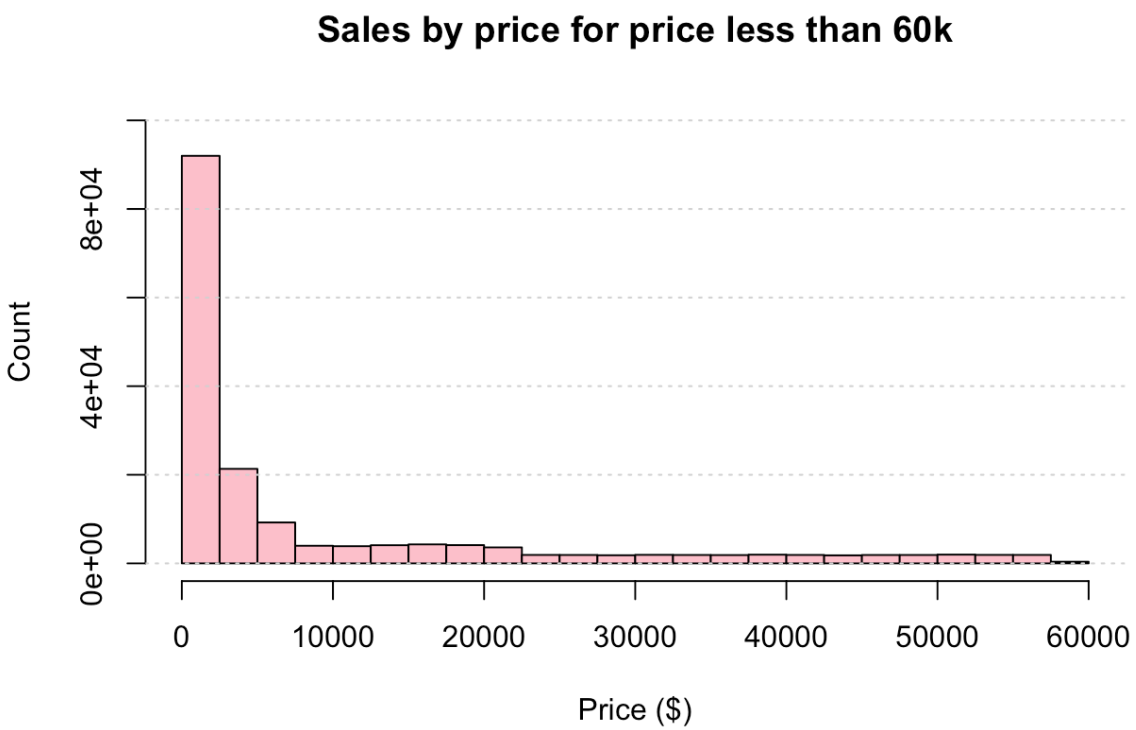

Figure 2: Sales by price for price less than 60k

The plots above show the number of sales categorised by price. The distribution of price data is positively skewed. From the first graph it can be seen that majority of the sales fall within the $0-25 000 price range. Around 50% of the sales in this data set are the price region below $10 000. This is further indicated by the second graph, which shows the distribution of data for sales below $60 000.

Delivery time

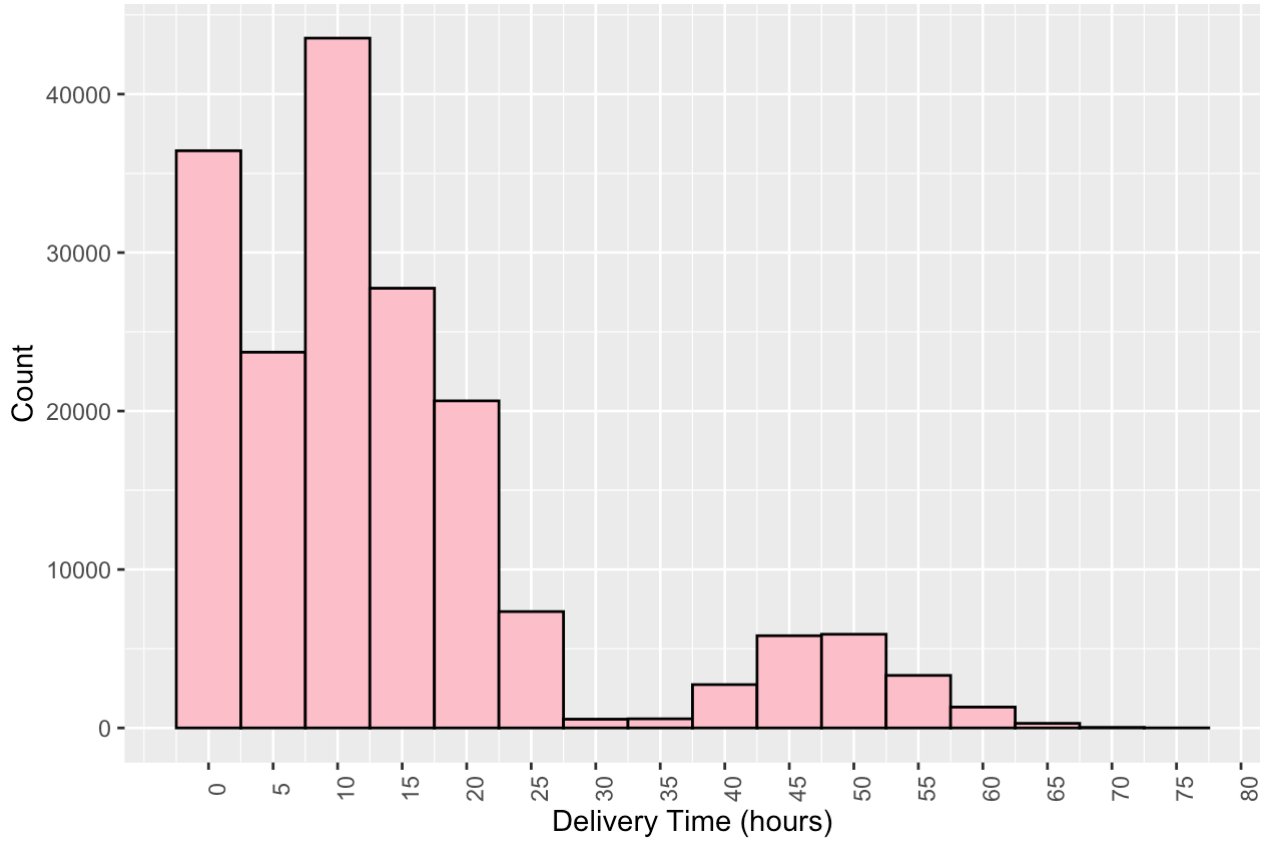

Figure 3: Distribution of sales by delivery time

The plot above shows the distribution of sales in terms of delivery time. Majority of the sales are delivered within 20 hours. There is another small peak in the interval from 40 hours to 60 hours but this is still a small percentage of the sales overall. The modal delivery time period is 10 hours.

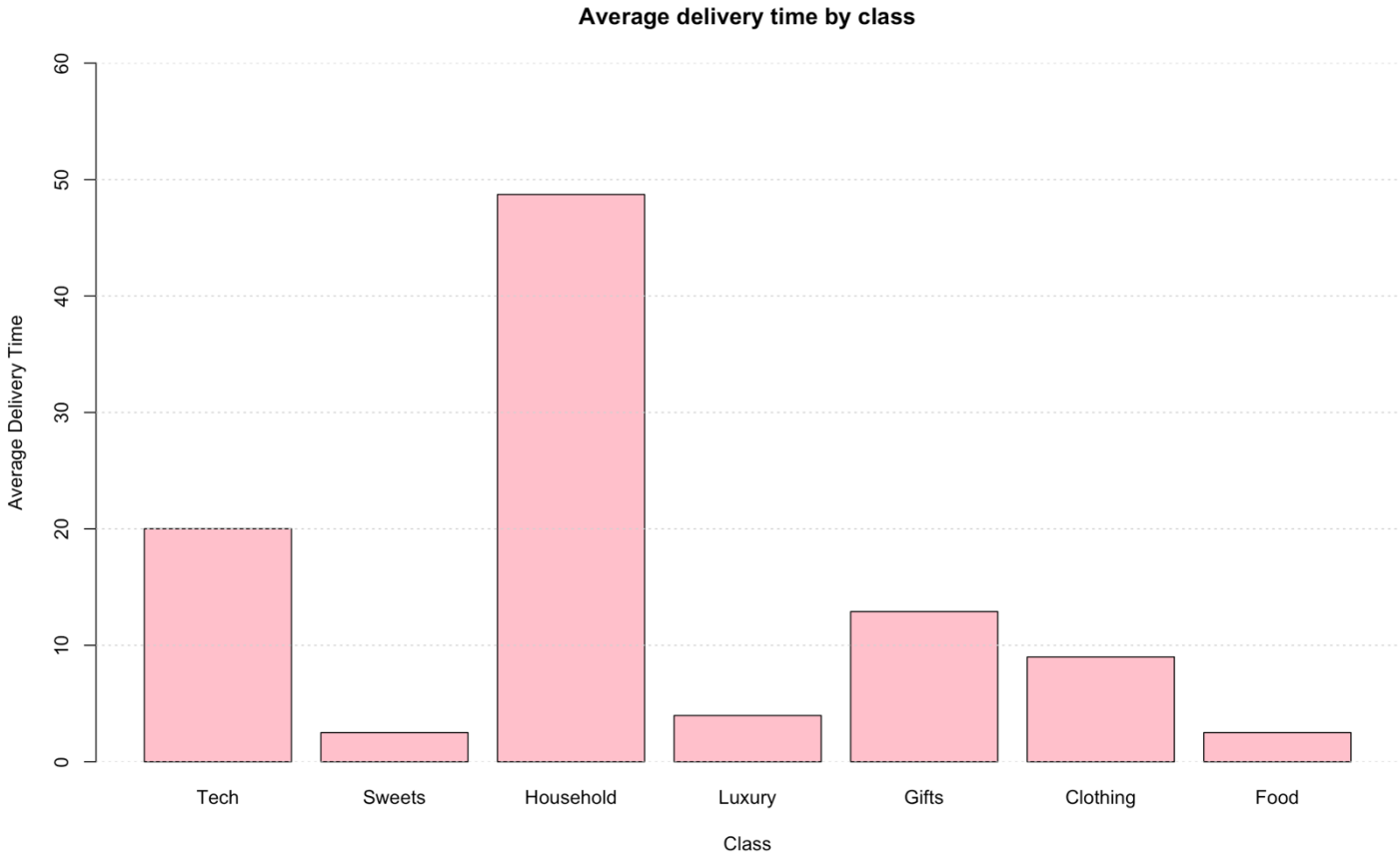

Figure 4: Average delivery time by class

The average delivery time for household items is much higher than all the other classes, sitting at around 48 hours. Gifts and technology have an average sitting between 10 and 20 hours and sweets and luxury items are under 5 hours. Household items are the only class of items that have such a high average delivery time. Management should aim to decrease the delivery time of household items to make it more in line with the expected delivery time of the other classes, as this would provide a better level of service to customers.

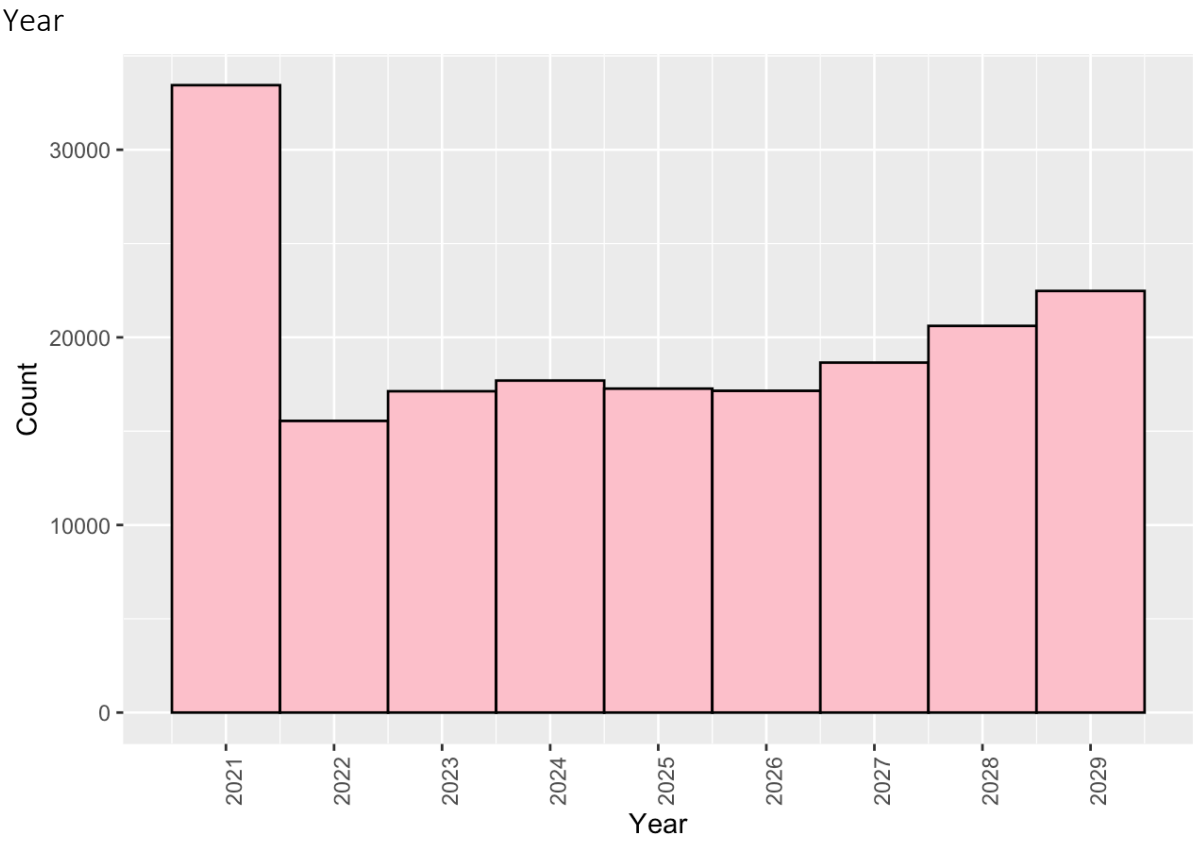

Figure 5: Distribution of sales by year

The distribution of sales over the past 9 years shows that sales dropped dramatically from 2021 to 2022. However, they have been fairly uniform from 2022 until 2026, with a slight positive increase from 2027 to 2029. There is value in exploring what changed from 2021 to 2022 that led to the sales decreasing so drastically.

Class

The table below shows some descriptive statistics for each class of product sold by the company. The technology, luxury and household classes have the highest prices.

Class	max_price	min_price	median_price
Clothing	1,154.02	127.76	642.04
Food	691.96	127.76	408.37
Gifts	5,774.49	172.61	2,961.59
Household	21,935.33	127.76	10,960.88
Luxury	116,618.97	12,825.37	65,342.14
Sweets	576.38	35.65	303.25
Technology	57,735.40	935.18	29,653.90

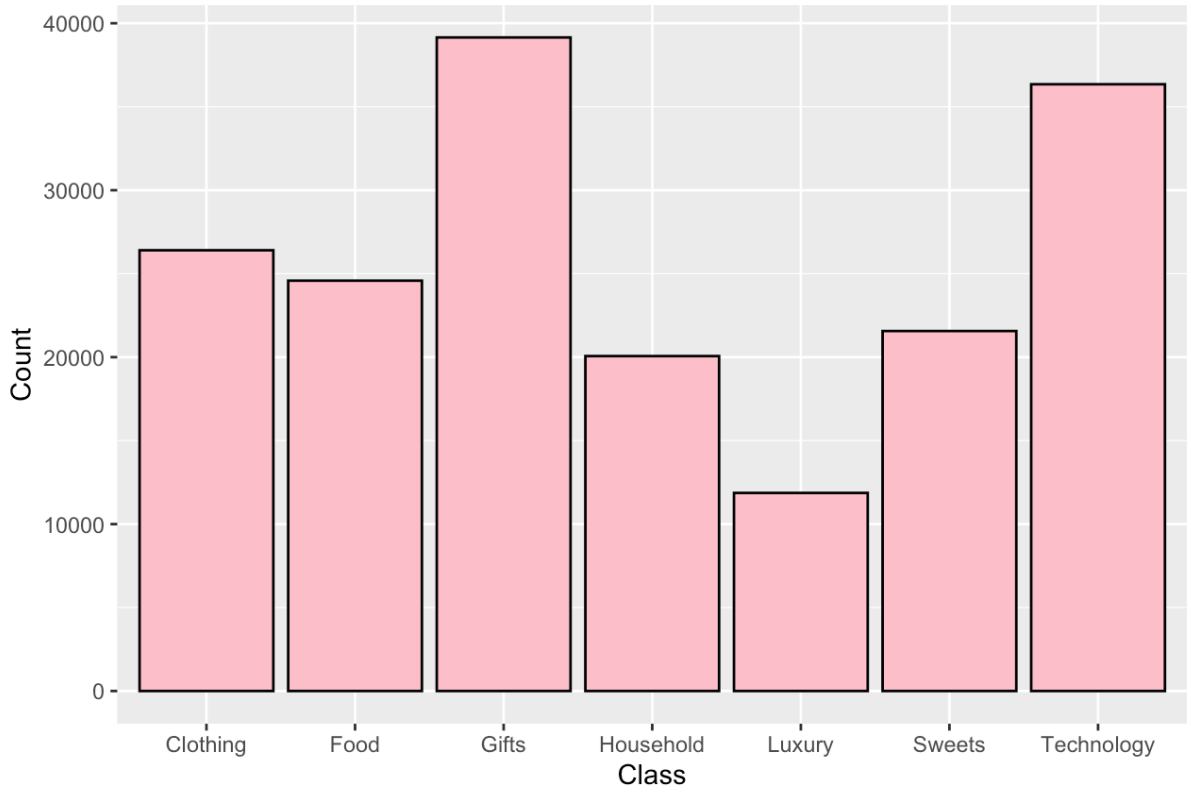

Figure 6: Distribution of sales by class

In terms of the distribution of sales by class, the plot above indicates that the gifts and technology class have the highest number of sales and the luxury class has the least number of sales. It can be expected that luxury products sell the least as the market for luxury goods is usually a smaller one.

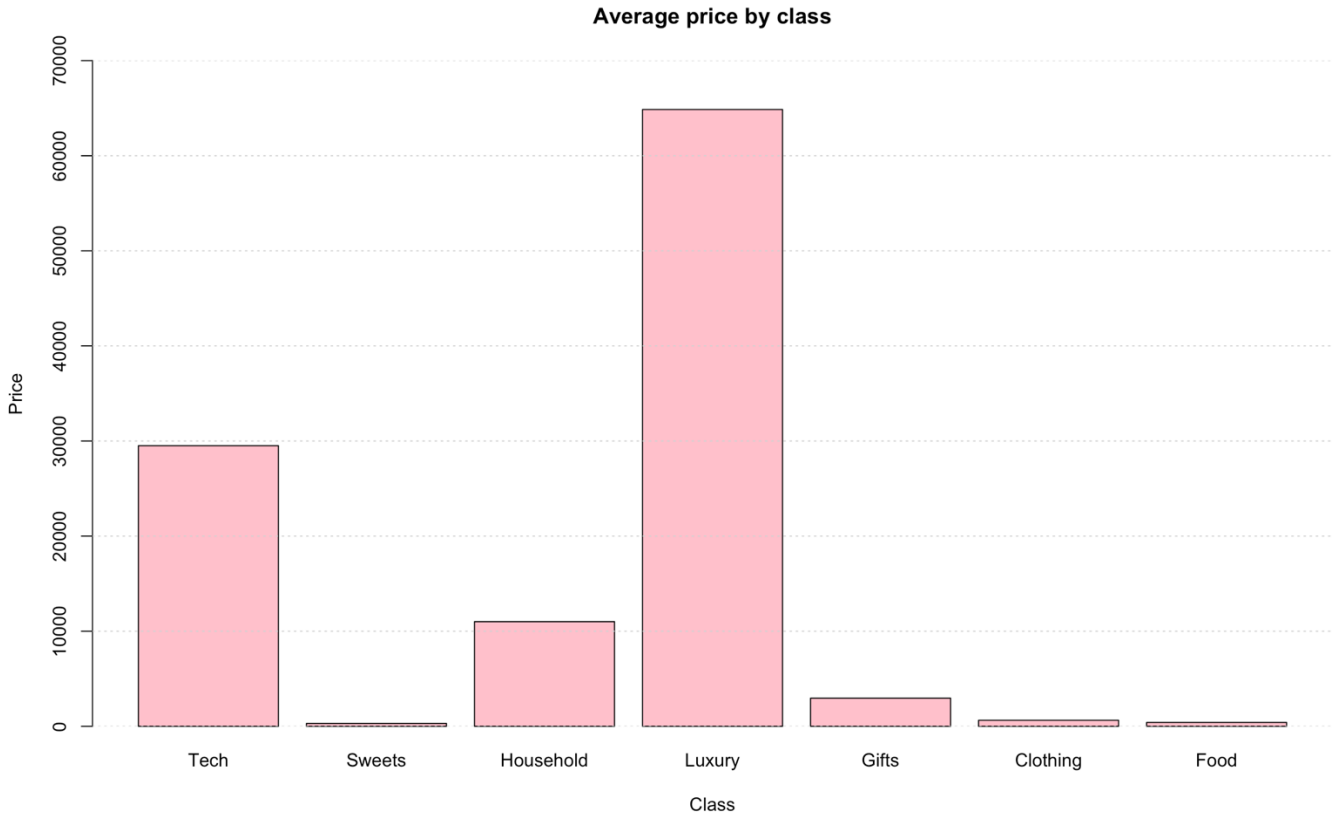

Figure 7: Average price by class

The plot above shows that the average price of technology and luxury sales are the highest. This makes sense and is in line with what would be expected, as technology items and luxury goods are generally at a higher price point than items such as food or clothing.

Reasons bought

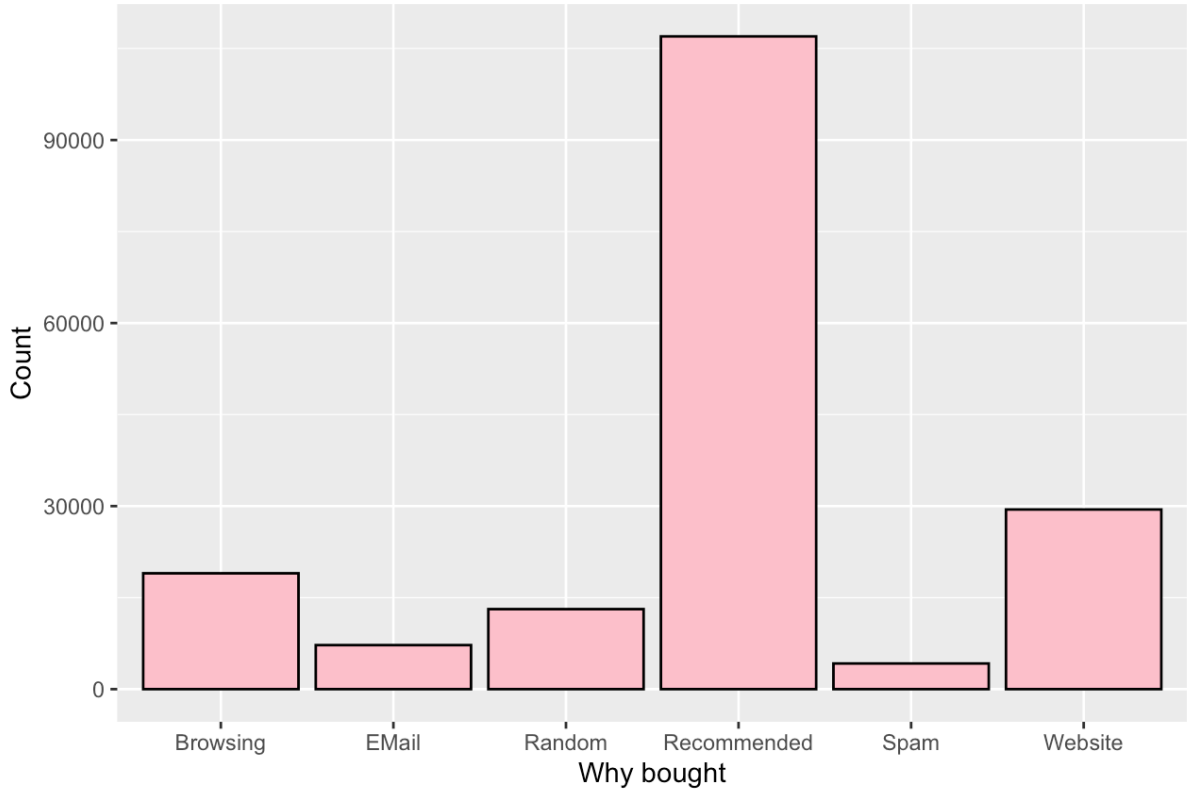

Figure 8: Distribution of sales by reasons for buying

Majority of the sales can be attributed to recommendations. This is positive and shows that the company is doing well and providing a good product (and service) if that many sales are being generated based on customers recommending the company to others. The high number of sales from the website, nearly 30 000, is also promising. It is worth trying to extend the range of customers and potential customers reached via e-mails and the website, to generate more sales and gain more traction through these channels.

Age

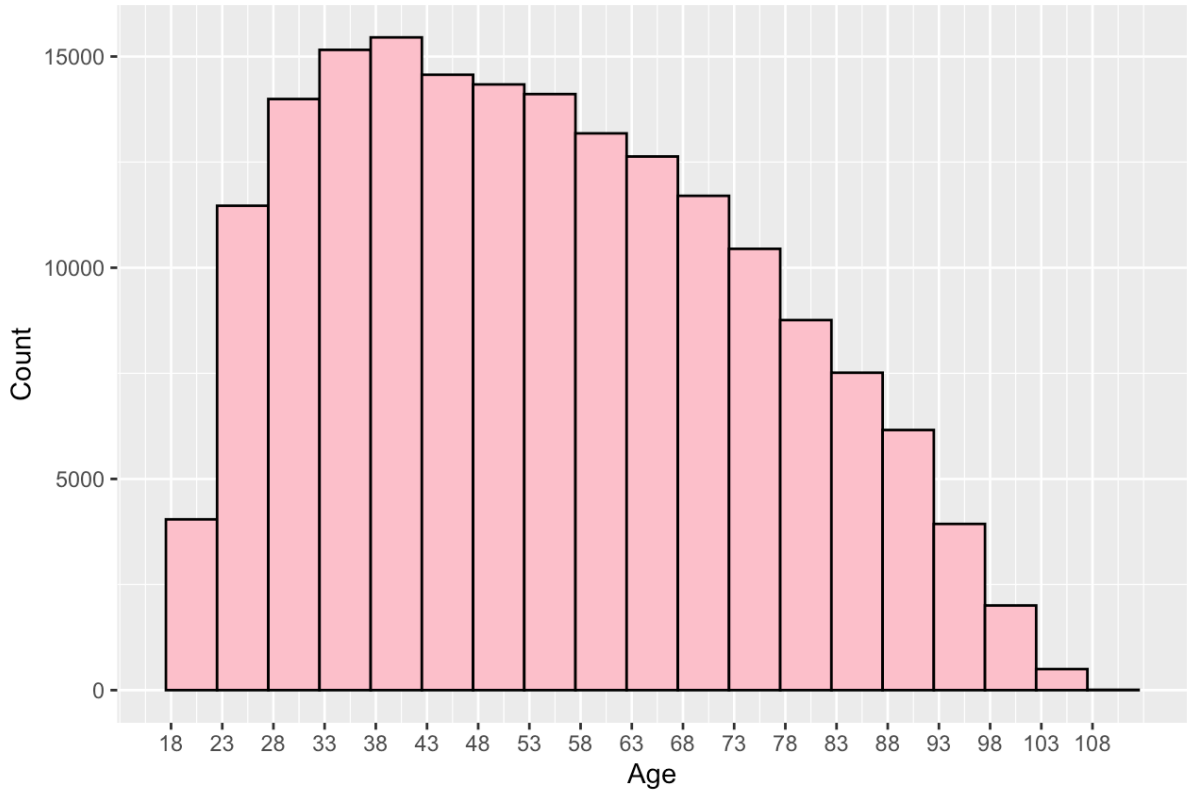

Figure 9: Distribution of sales by age

The data indicates the modal class in terms of age of customers is 38-43. The distribution of the age data is skewed to the right. There are some possible outliers in the data, as the maximum age recorded in the sales data is 108, which seems high. It could be due to incorrect data inputting, however, given the futuristic dates of the data set, it is possible that the age is correct.

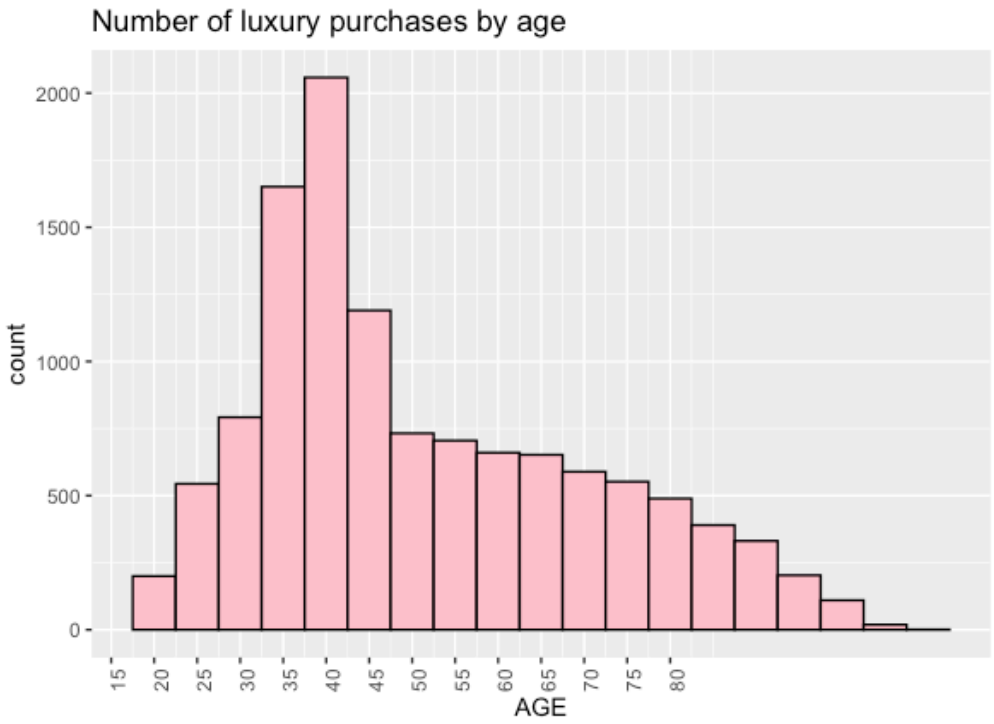

Figure 10: Distribution of luxury sales by age

As can be seen from the plot above, people between approximately age 30 and 45 purchase a large amount of luxury items.

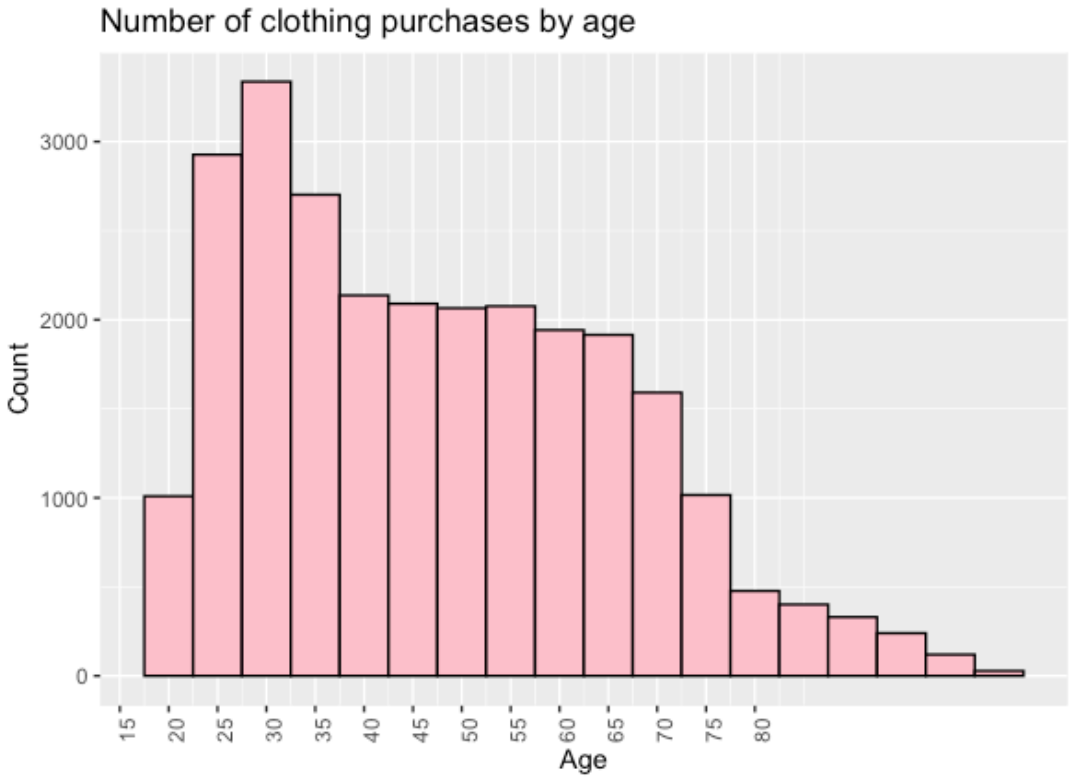

Figure 11: Distribution of clothing sales by age

The above plot indicates that the modal age group for clothing sales is the 25-30 age group.

Process capacity metrics

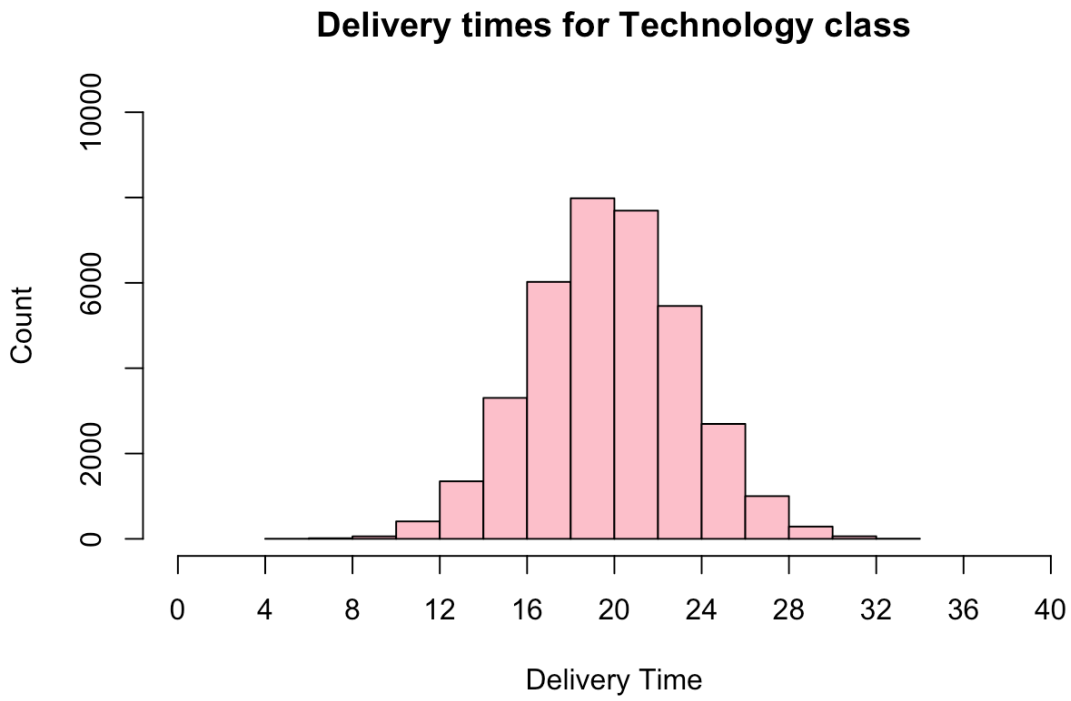

Figure 12: Distribution of delivery times for Technology class

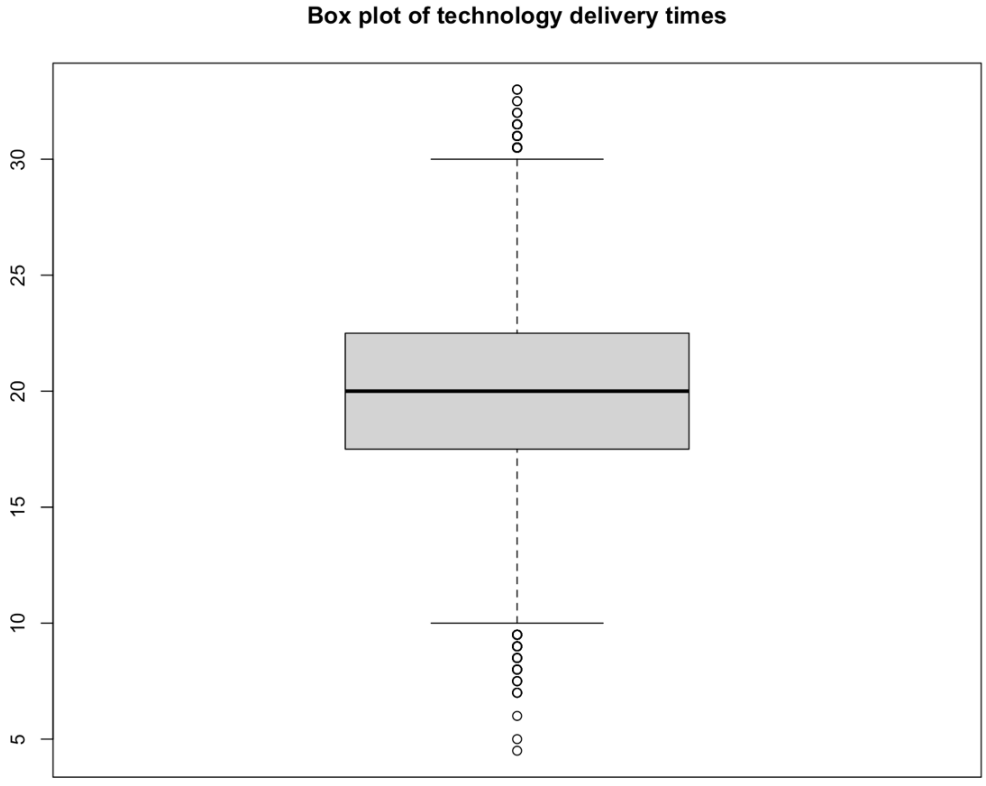

Figure 13: Box plot of Technology delivery times

The mean delivery time for sales in the technology class is 20.02 hours. For this data, the upper service level is 24 hours, and the lower service level is 0 hours. A mean delivery time of 20.02 hours indicates that the process is not centred, as the mean time is not in the middle of the upper and lower service levels (Hessing, no date).

The Cp metric indicates whether the process has the capability to fit within the specification of that process (Interpret Cp and Cpk, [S.a.]). However, it does not consider how centred a process is (Hessing, no date). For this specific process, the Cp metric is 1.142. Having a Cp metric of higher than 1 indicates that the process does have the capability to operate within the limits of the process (Hessing, no date). However, there are also other metrics that need to be taken into consideration.

The Cpk (capability index) is 0.38. Having a capability index of below 1 indicates that the process is not capable (Process Capability – Cp, Cpk, Pp, Ppk, [S.a.]). It is indicative of an issue with the delivery time of products as there is a big variation within the process.

The Cpu of this process is 0.38 and the Cpl is 1.904. A Cpu of less than 1 indicates that there will be delivery times above the upper service level and therefore out of specification.

As mentioned above, the lower service level (LSL) for this process is 0. This makes sense as the delivery time cannot be lower than 0 hours. The Cpl is greater than 1, so there should be no delivery times below the LSL of the process (Process Capability Part 2, [S.a.]). Again, this makes sense as there cannot be any delivery times below 0 (which is also the LSL). The plots above further support the abovementioned observations.

Part 3: Statistical Process Control

The X-bar chart and S-chart metrics are tabulated below for each class of sales. These were calculated using the first 450 instances of sales for each class. The first 450 instances were divided into 30 samples, containing 15 instances each. The S-chart metrics were computed in a similar manner.

X-bar chart							
Class	UCL	U2sig	U1sig	CL	L1sig	L2sig	LCL
Tech	22.927149	22.076248	21.225346	20.374444	19.523543	18.672641	17.821740
Sweets	9.467983	9.099026	8.730068	8.361111	7.992154	7.623197	7.254239
Household	5.480120	5.231932	4.983744	4.735556	4.487367	4.239179	3.990991
Luxury	2.889388	2.752184	2.614981	2.477778	2.340574	2.203371	2.066168
Gifts	50.181037	48.974766	47.768494	46.562222	45.355951	44.149679	42.943407
Clothing	9.396994	9.254662	9.112331	8.970000	8.827669	8.685338	8.543006
Food	2.705451	2.633634	2.561817	2.490000	2.418183	2.346366	2.274549

S-chart							
classes	UCL	U2sig	U1sig	CL	L1sig	L2sig	LCL
Technology	5.1805697	4.9973310	4.1464294	3.2955278	2.4446262	1.5937246	1.4104859
Sweets	0.8353391	0.8057929	0.6685896	0.5313862	0.3941829	0.2569796	0.2274333
Household	7.3441801	7.0844137	5.8781420	4.6718703	3.4655986	2.2593268	1.9995605
Luxury	1.5110518	1.4576053	1.2094171	0.9612289	0.7130406	0.4648524	0.4114060
Gifts	2.2463333	2.1668797	1.7979225	1.4289652	1.0600080	0.6910508	0.6115971
Clothing	0.8665596	0.8359090	0.6935778	0.5512465	0.4089153	0.2665841	0.2359335
Food	0.4372466	0.4217810	0.3499638	0.2781467	0.2063295	0.1345124	0.1190468

Below is an example of an X-chart initialized for the first 30 samples of the technology class.

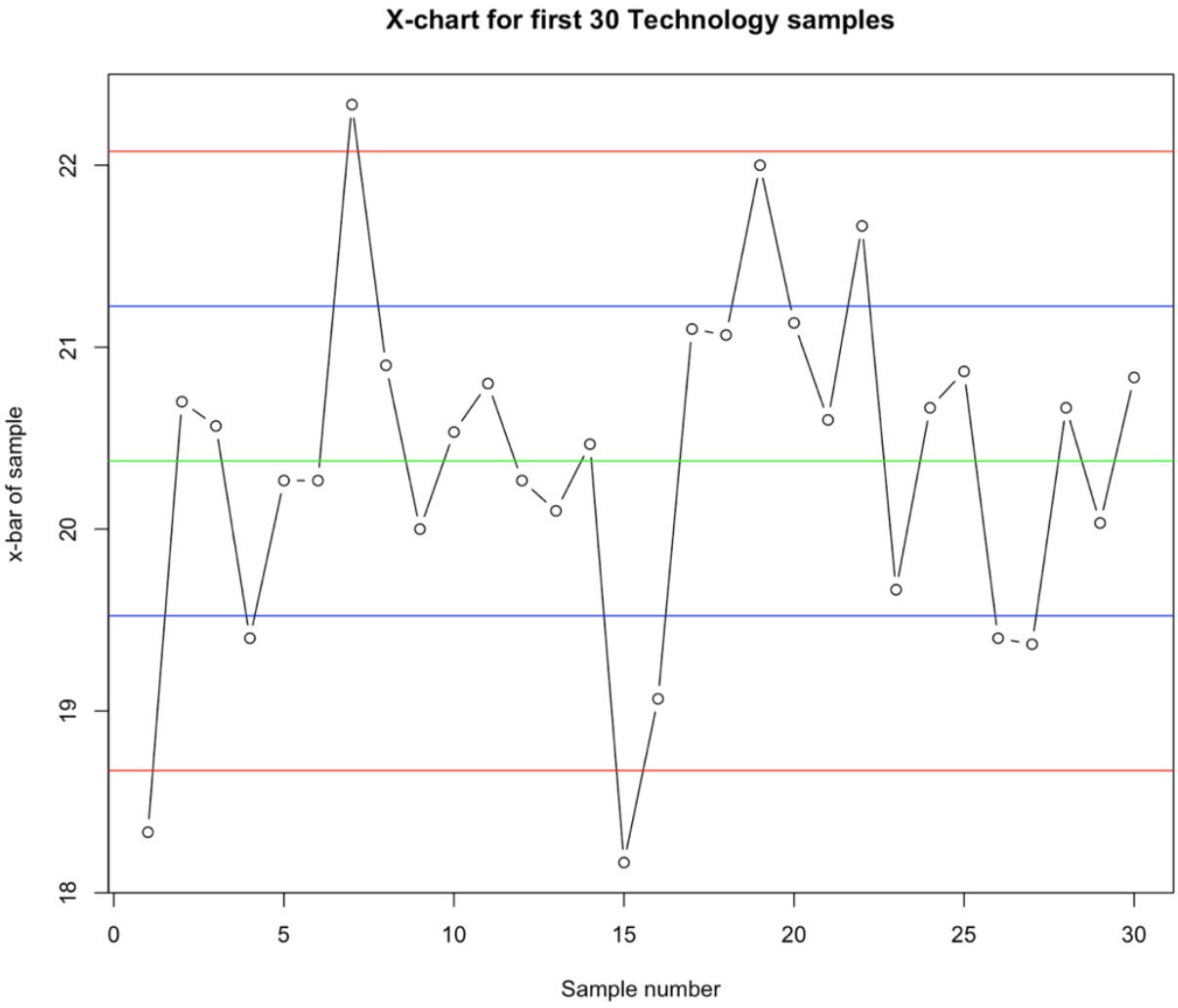

Figure 14: X-chart example

Below is an example of an S-chart, initialized for the first 30 samples of the Sweets class.

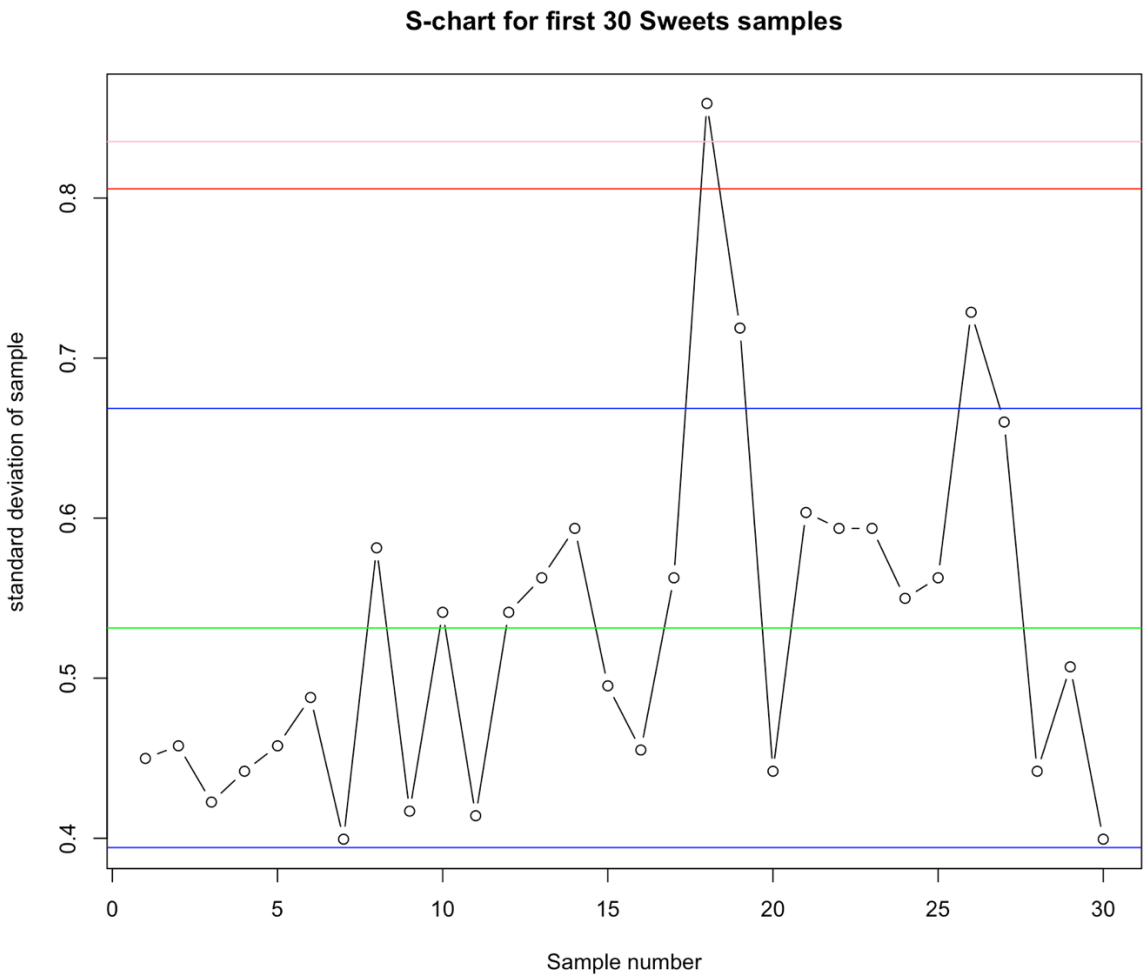

Figure 15: S-chart example

Part 4: Optimising the delivery process

Part 4.1

A

The samples which are out of the control limits of the process for each class of sales were computed and are tabulated below.

Samples outside the control limits by class

Class	Samples with means above the UCL	Samples with means below the LCL	Number of samples outside of control limits
Technology	1122	head = 37 398 483 tail = 1961 2009 2071	20
Gifts	head = 213 216 218 tail = 2608 2609 2610	none	2291
Luxury	none	head = 142 171 184 tail = 789 790 791	434
Sweets	942 1243 1294	1104 1403	5
Household	head = 693 725 752 tail = 1336 1337 1338	252 387 629 643	405
Clothing	head = 148 217 455 tail = 1644 1723 1724	1161 1359 1557 1677 1761	23
Food	633	75 1203 1467 1515	5

Below are plots showing a sample from each class that has a mean delivery time outside of the upper or lower control limits for each respective class. It makes sense that these samples would have many instances outside of the control limits.

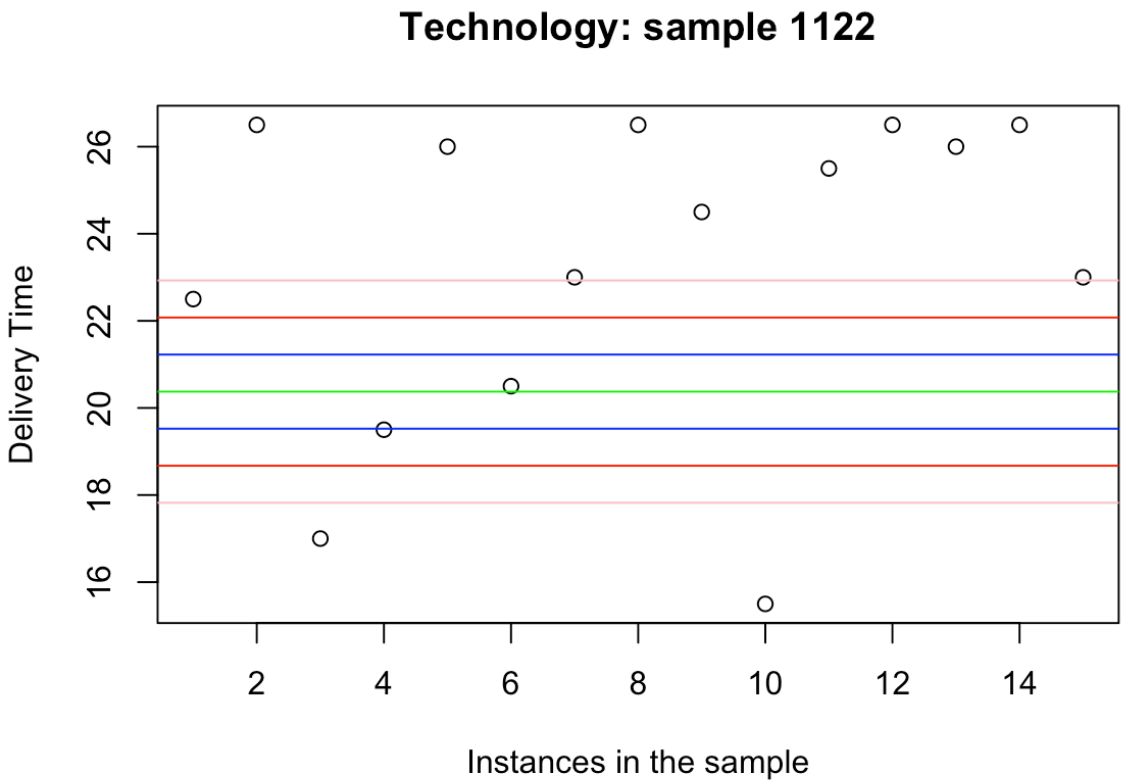

Figure 16: Technology sample

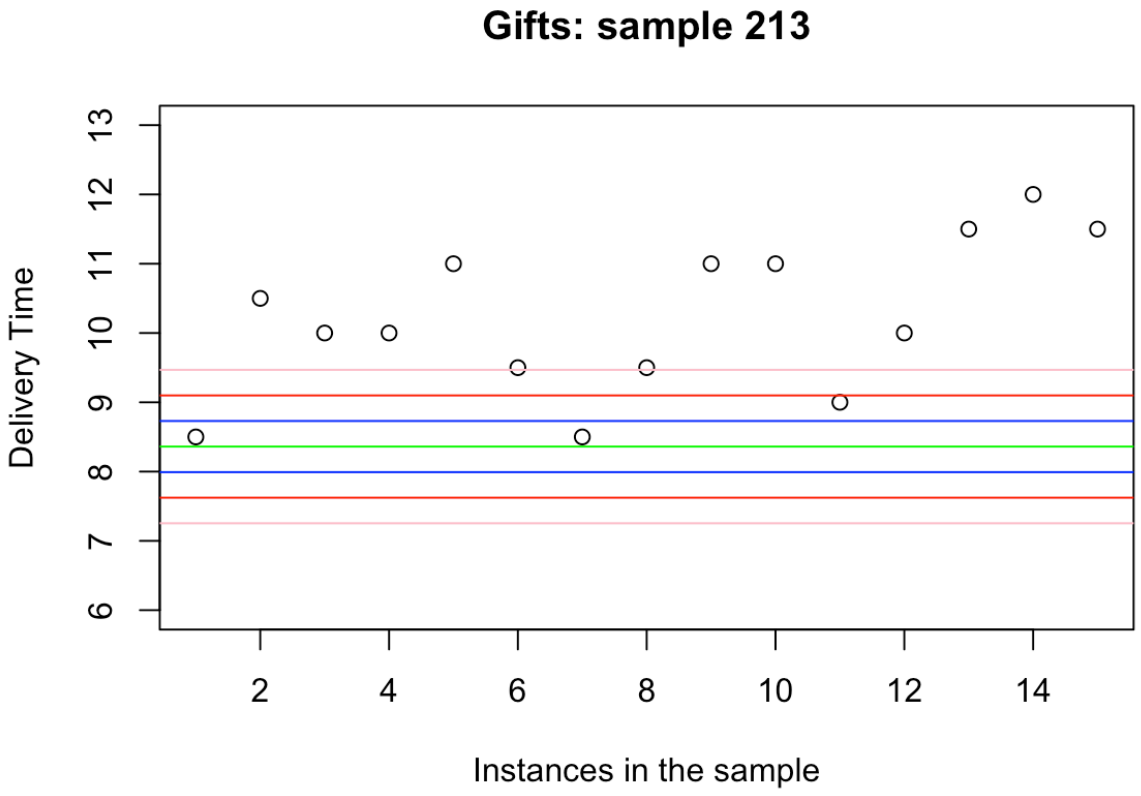

Figure 17: Gifts sample

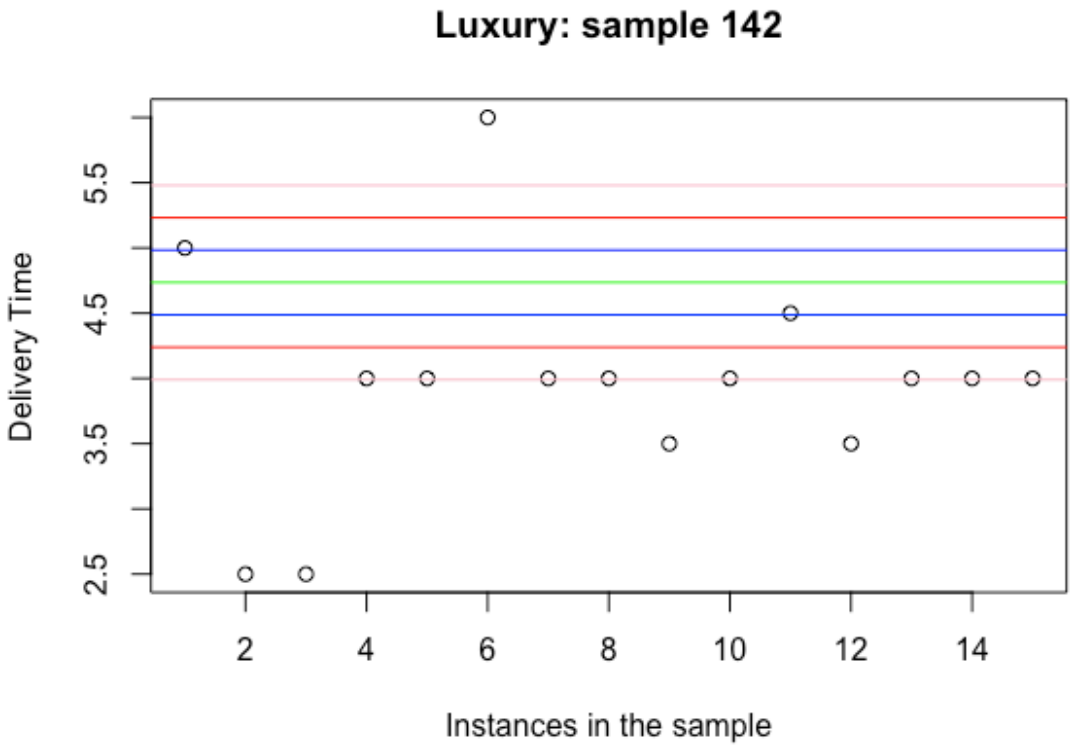

Figure 18: Luxury sample

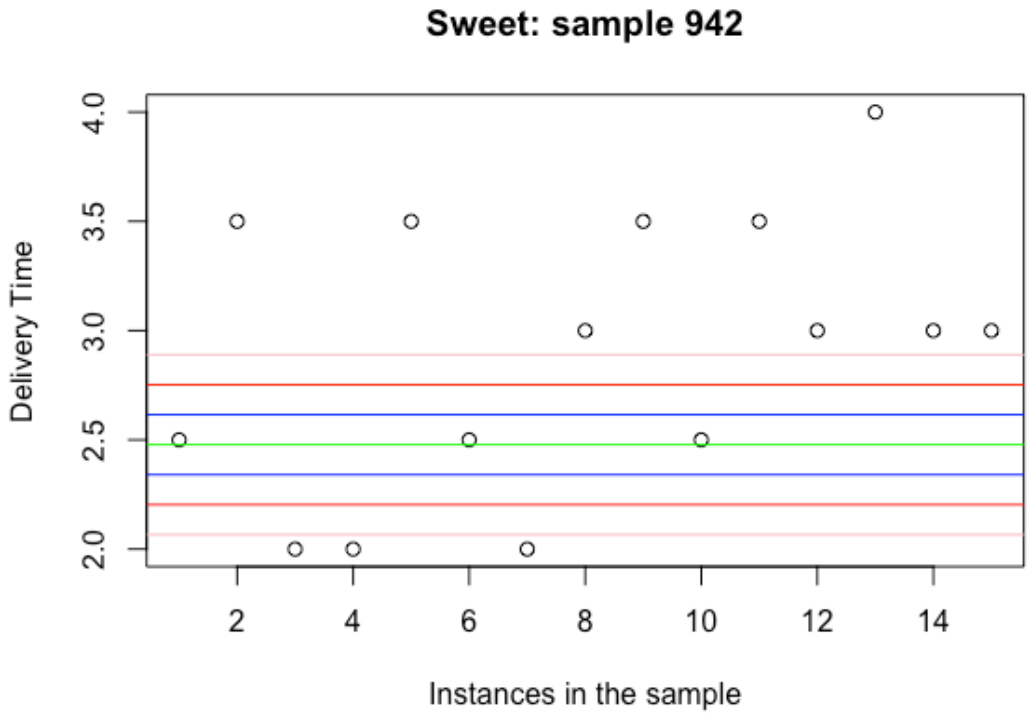

Figure 19: Sweet sample

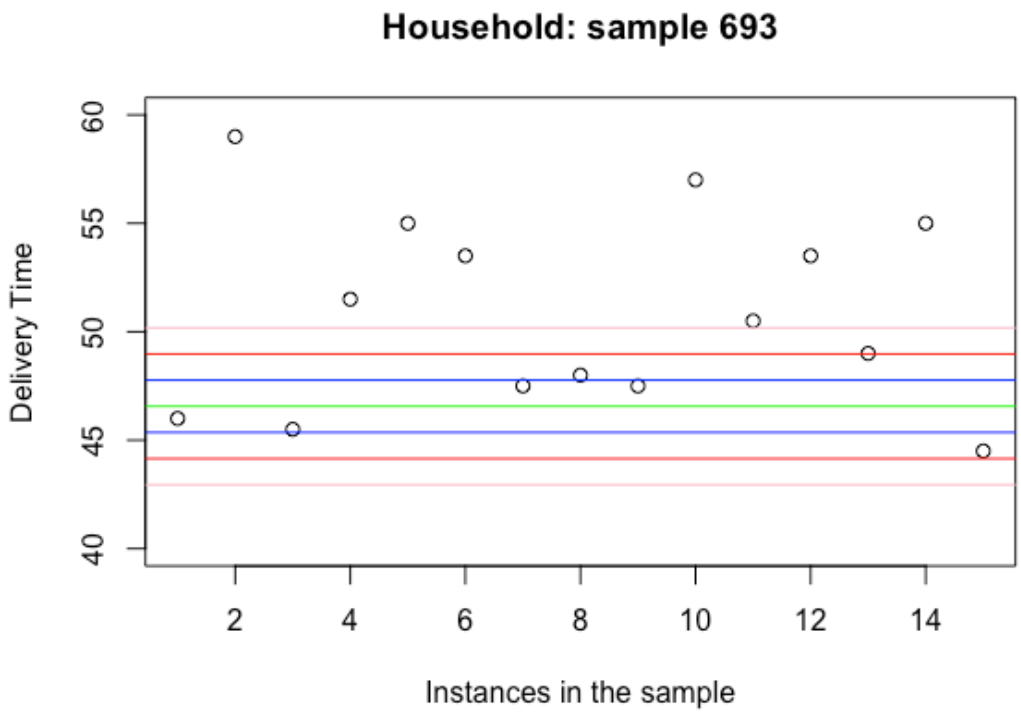

Figure 20: Household sample

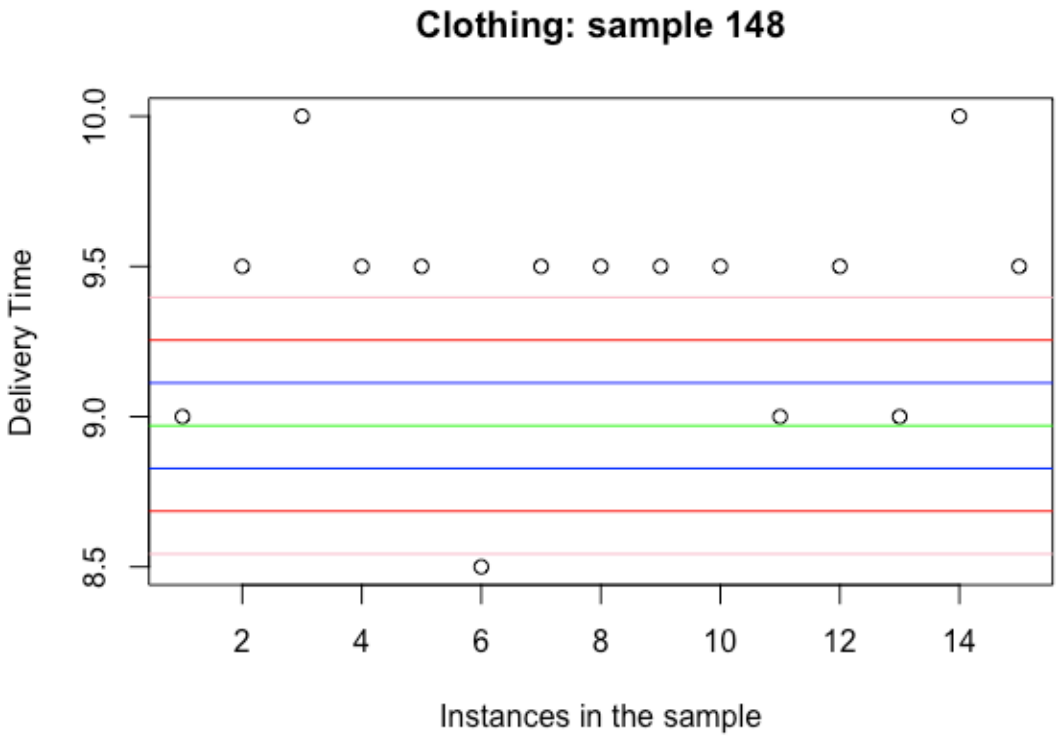

Figure 21: Clothing sample

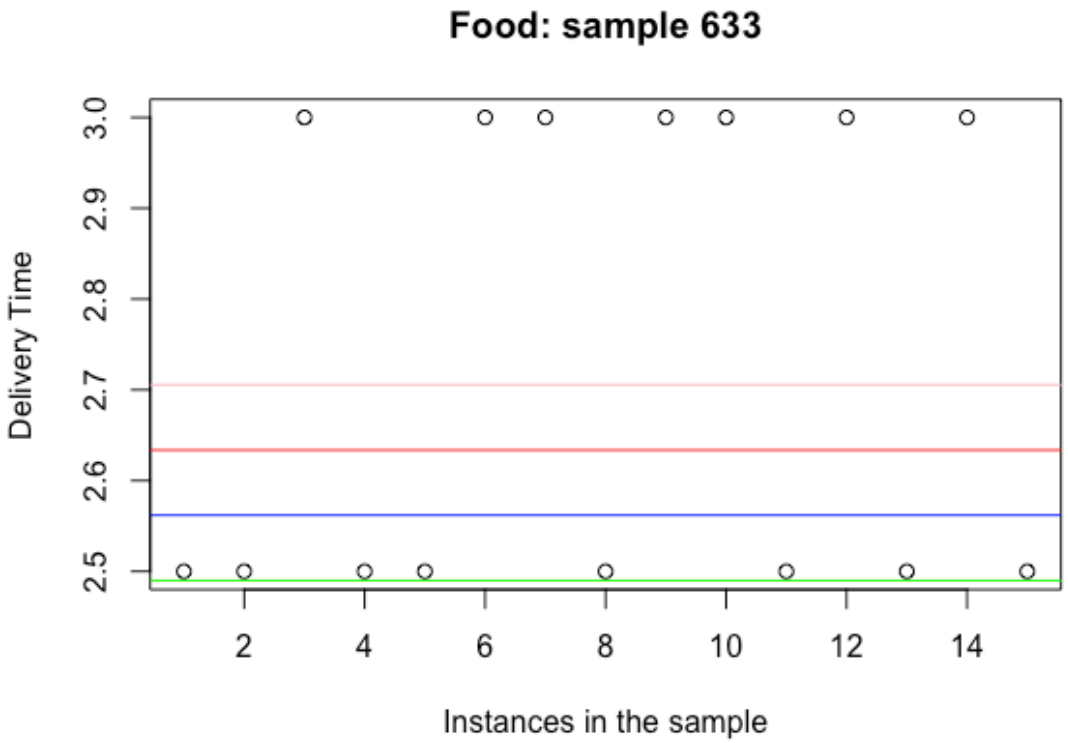

Figure 22: Food sample

B

The maximum number of consecutive samples within the range [-0.3σ, 0.4 σ] for the respective classes, as well as the ending sample number has been computed and tabulated below. In some cases, there is more than one string of consecutive samples within the range, hence the multiple ending sample numbers.

Class	Maximum consecutive samples	Ending sample number
Technology	6	372
Sweets	4	94 189 971 1292
Household	3	45 198 545 588 647 843 900 908
Luxury	4	63
Gifts	5	254 307 603 1651
Clothing	4	1013
Food	7	952

Part 4.2

A type I error is also commonly referred to as a manufacturer's error or a false positive. In the context of process control, a type I error is the probability of thinking that a process is out of control when it is actually within the control limits of the process (Swamidass, 2000). This is represented by an alpha (α) value and is the significance level chosen and set at the beginning of a study when assessing the probability of being within the capability of the process. α is usually around 5% (0.05) (Bhandari, 2021).

Part 4.3

To compute the optimal cost solution for delivery of technology items, the total costs were computed if the average delivery decreased by a range of hours. The minimum delivery time is 4.5 hours. Therefore, the most that the delivery time of every item can be reduced was selected as 4 hours for practicality purposes. The cost per item late per hour was added to the cost of reducing the average delivery time by a specific number of hours to find the total cost. The minimum total cost can be found if the average delivery time is reduced by 3 hours.

The original cost for items delivered after 26 hours is R 758674. This is accounting for the 1356 deliveries that have a delivery time of longer than 26 hours. It costs R329 per item per hour that a specific item is late.

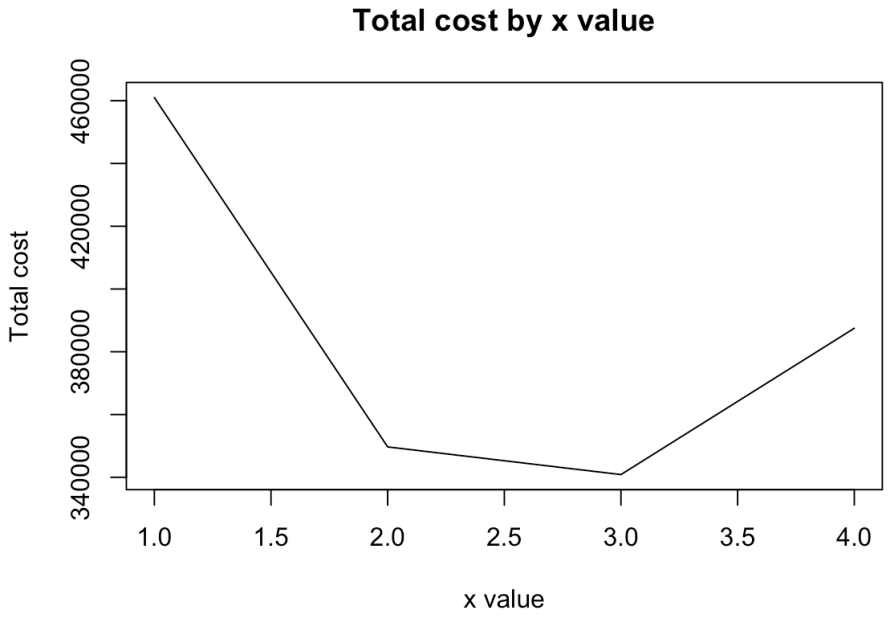

Figure 23: Total cost by x value

After plotting the total costs as the total number of hours that the delivery costs are delivered by is increased from 1 to 4, it is found that the optimal number of hours to reduce the delivery times by is 3 hours. This results in the lowest total cost, R 340870.

By reducing the delivery times of all technology items by 3 hours, the new best average delivery time is 17.01 hours, and the process should be centred on this time.

Part 4.4

A Type II error refers to the probability that a process will be out of control, but we fail to identify it as such. We can find the probability of making a Type II error by finding the probability that the process will still be within the upper control limit and therefore we do not notice that the centre of the process has moved.

This can be done by calculating a z-value based on the values provided in the question.

$$z = \frac{23 - 22.9}{UCL - LCL/6} = 0.085$$

This gives a z-value of 0.085. Therefore, the probability is approximately 53.2%.

Part 5: MANOVA

Running a MANOVA with class as the independent variable and age, price and delivery time as dependent variables revealed that price, delivery time and age are affected by the class of the item sold.

This can be seen in the plots below, as the mean price of the different classes is different, which is in line with what can be expected. The same can be seen for the mean delivery time of the different classes. This is further confirmed by the small p-value obtained when running the MANOVA in R.

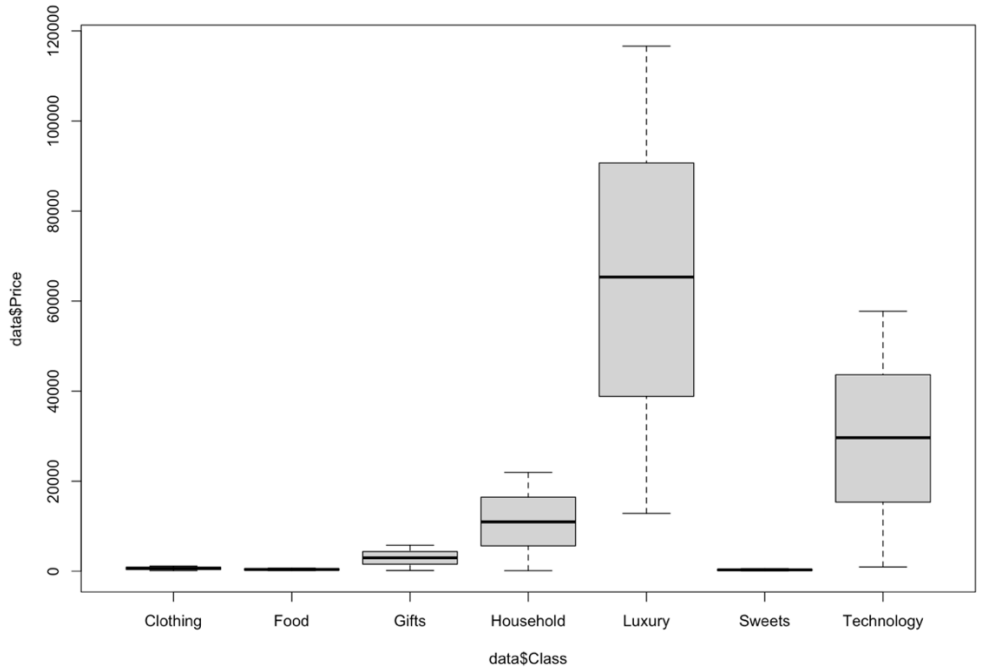

Figure 24: Box plots of price by class

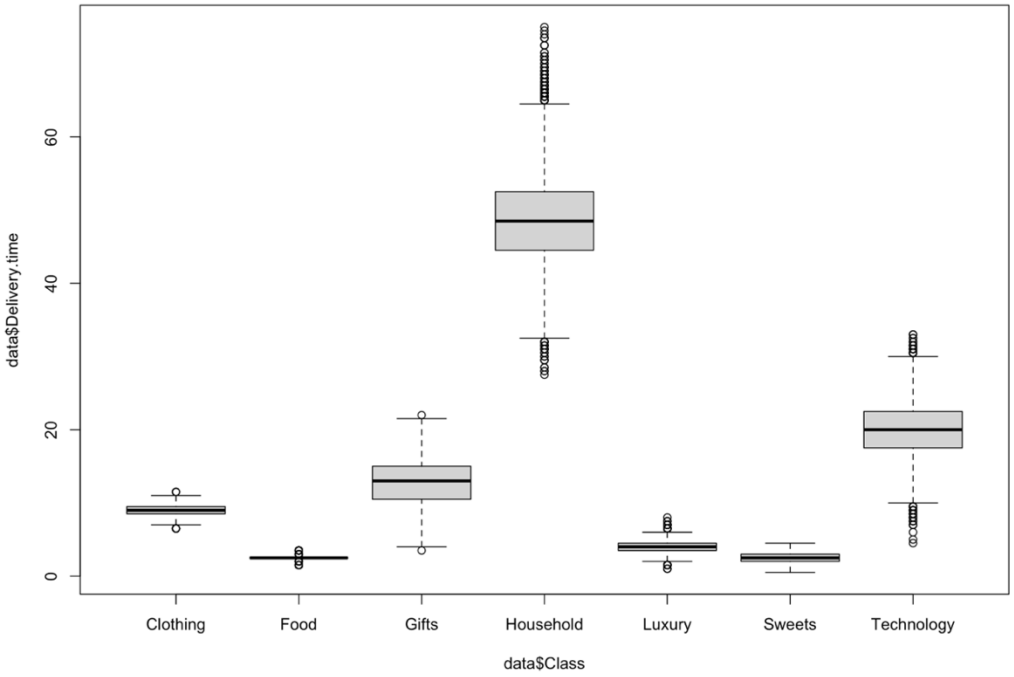

Figure 25: Box plots of delivery time by class

It can be seen from the box plots below that there is less of an effect of class on age when compared with the other two dependent variables under consideration (MANOVA Test in R: Multivariate Analysis of Variance, [S.a.]). However, there is still an effect, which is also further confirmed by the small p-value obtained when carrying out the MANOVA.

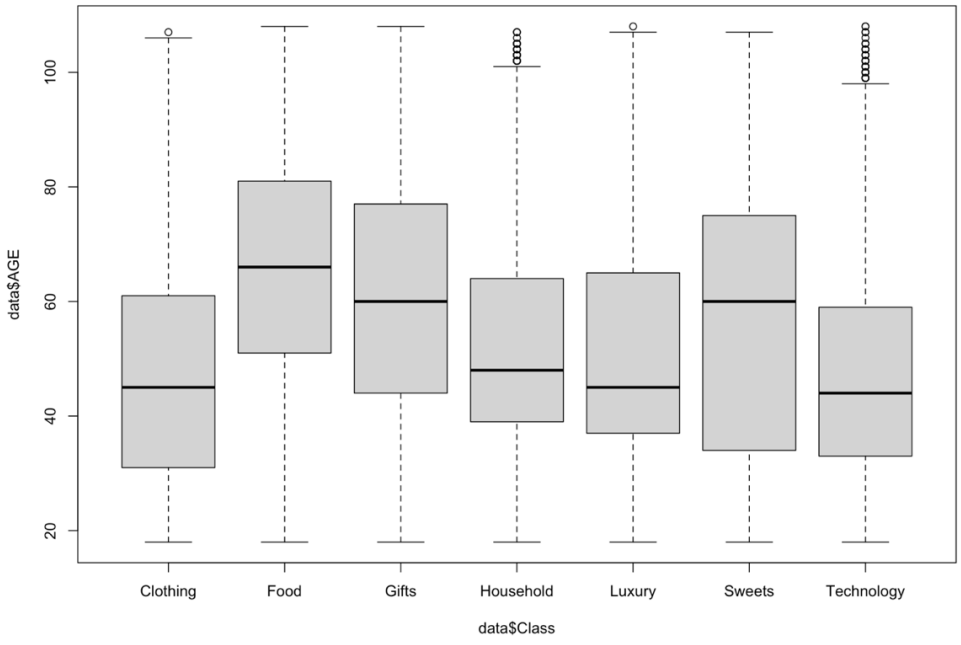

Figure 26: Box plots of age by class

Part 6: Reliability of the service and products

6.1

Question 6

$\text{LSL} = 0.06 - 0.04$

$\text{USL} = 0.06 + 0.04$

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

$\therefore L(x) = 28125(x - T)^2$

Question 7

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

$\therefore L(x) = 21875(x - T)^2$

$21875(0.06 - 0.027)^2 = \$23.82188$

6.2

a. $0.85 \times 0.92 \times 0.9 = 0.7038$

b. There is added reliability by adding an extra machine at each stage. The probability of both A machines not working is 0.0225. The probability of both B machines not working is 0.0064. The probability of both C machines not working is 0.0100. Therefore, the new system reliability with 2 machines at each station is 0.9615316 which is 96.15%. This increases the system reliability by 0.2577316 or 25.77%.

6.3

Binomial probabilities were calculated for each of the 7 cases (1 vehicle not working, 2 drivers not working, etc.) using the information given. Using these binomial probabilities, a weighted average was used to find the approximate probability of a vehicle not working on a given day and the approximate probability of a driver not working on a given day.

Using these probabilities, it can be computed that reliable delivery (days where there are at least 19 drivers and 19 vehicles working) can be expected on 355.1623 days in a year, which is 97.3% of the year. Increasing the number of vehicles from 21 to 22 increases the reliability to 362.0934 days in the year, which equates to 99.2% of the year.

Conclusion

Thorough data analysis of the sales data set under consideration has been presented and discussed. Interesting findings from the data have been discussed, along with useful data-driven recommendations to management, in order to improve and optimise the delivery process of sales. Multivariate analysis of variation has also been carried out and presented to find relationships between the different features within the data set. Lastly, binomial distribution problems have been solved and presented which tie into service and reliability of the delivery process.

References

Process Capability – Cp, Cpk, Pp, Ppk. [S.a.]. [Online]. Available: https://www.presentationeze.com/presentations/statistical-process-control/statistical-process-control-full-details/process-capability-cp-cpk-pk-ppk/ [2022, September]

*Interpret Cp and Cpk. [*S.a.]. [Online]. Available: https://www.pqsystems.com/qualityadvisor/DataAnalysisTools/capability_4.6.3.php [2022, September]

Hessing, T. *Process Capability (Cp & Cpk)* [Online]. Available: https://sixsigmastudyguide.com/process-capability-cp-cpk/ [2022, September]

Hessing, T. *X Bar S Control Chart* [Online]. Available: https://sixsigmastudyguide.com/x-bar-s-chart/ [2022, September]

Bhandari, P. *Type I & Type II Errors | Differences, Examples, Visualizations* [Online]. Available: https://www.scribbr.com/statistics/type-i-and-type-ii-errors/#:~:text=The%20risk%20of%20making%20a%20Type%20I%20error%20is%20the,set%20at%200.05%20or%205%25 [2022, October]

MANOVA Test in R: Multivariate Analysis of Variance. [S.a.]. [Online]. Available: http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance [2022, October]

Process Capability Part 2. [S.a.]. [Online]. Available: https://www.spcforexcel.com/knowledge/process-capability/process-capability-part-2#:~:text=One%20process%20capability%20is%20Cpu,where%20the%20process%20is%20centered [2022, October]

Swamidass, P.M. *Type I Error (Alpha Error)*. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/1-4020-0612-8_1011#:~:text=In%20making%20decisions%20using%20control,of%20a%20process%20control%20chart. [2022, October]