QUALITY ASSSURANCE 344 - PROJECT

Department of Industrial Engineering at Stellenbosch University

MH Geel – 20826923

25/09/2022

Table of Contents

List of Figures

List of Tables

1. Introduction

Client data for an online business was received and analysed. This analysis has been done by reporting and evaluating the data set given to adhere to all ECSA outcomes as expected. The data was cleaned, completed sets were formed, and necessary statistical analysis was conducted to draw multiple conclusions to give the client feedback on his business reliability, quality, and services.

2. Data Wrangling

The purpose of the first section was to identify any errors and incomplete data from the main data set. The area addressed here was that of missing values. The data set was separated into two smaller data sets, namely a data set seen as valid, and a data set seen as incomplete. The one data set contained all the observations with no missing values, whereas the other data set contained all the observations with missing values.

For easy identification, a new primary key was introduced. This was added to the left of the original primary key which was now taken as the secondary key in each of the new data sets created. In this manner the length of each of the new data sets could easily be determined, by the implementation of the new primary key added.

From Figure 1 and 2 the new added primary key can easily be seen. This was implemented by using the cbind() function after new vectors have been created to be added as the primary key columns to each of the newly formed data sets.

```
> head(valid_d)
  valid_dfull X    ID AGE      Class    Price Year Month Day Delivery.time  Why.Bought
1           1 1 19966  54     Sweets   246.21 2021     7   3           1.5 Recommended
2           2 2 34006  36  Household  1708.21 2026     4   1          58.5     Website
3           3 3 62566  41      Gifts  4050.53 2027     8  10          15.5 Recommended
4           4 4 70731  48 Technology 41843.21 2029    10  22          27.0 Recommended
5           5 5 92178  76  Household 19215.01 2027    11  26          61.5 Recommended
6           6 6 50586  78      Gifts  4929.82 2027     4  24          14.5      Random
```

Figure 1: Valid Data Set [head()]

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

Figure 2: Incomplete Data Set [head()]

After observing the incomplete data set in full it can be seen that all the missing values appeared in the 'Price' variable. This observation can be validated in Figure 3.

```
> incomp_d
       incompvd      X    ID AGE      Class Price Year Month Day Delivery.time  Why.Bought
12345         1  12345 18973  93      Gifts    NA 2026     6  11          15.5     Website
16320         2  16320 44142  82  Household    NA 2023    10   2          48.0       EMail
16321         3  16321 81959  43 Technology    NA 2029     9   6          22.0 Recommended
19540         4  19540 65689  96     Sweets    NA 2028     4   7           3.0      Random
19541         5  19541 71169  42 Technology    NA 2025     1  19          20.5 Recommended
19998         6  19998 68743  45  Household    NA 2024     7  16          45.5 Recommended
19999         7  19999 67228  89      Gifts    NA 2026     2   4          15.0 Recommended
23456         8  23456 88622  71       Food    NA 2027     4  18           2.5      Random
34567         9  34567 18748  48   Clothing    NA 2021     4   9           8.0 Recommended
45678        10  45678 89095  65     Sweets    NA 2029    11   6           2.0 Recommended
54321        11  54321 62209  34   Clothing    NA 2021     3  24           9.5 Recommended
56789        12  56789 63849  51      Gifts    NA 2024     5   3          10.5     Website
65432        13  65432 51904  31      Gifts    NA 2027     7  24          14.5 Recommended
76543        14  76543 79732  71       Food    NA 2028     9  24           2.5 Recommended
87654        15  87654 40983  33       Food    NA 2024     8  27           2.0 Recommended
98765        16  98765 64288  25   Clothing    NA 2021     1  24           8.5    Browsing
144443       17 144443 37737  81       Food    NA 2022    12  10           2.5 Recommended
144444       18 144444 70761  70       Food    NA 2027     9  28           2.5 Recommended
155554       19 155554 36599  29     Luxury    NA 2026     4  14           3.5 Recommended
155555       20 155555 33583  56      Gifts    NA 2022    12   9          10.0 Recommended
166666       21 166666 60188  37 Technology    NA 2024    10   9          21.5     Website
177777       22 177777 68698  30       Food    NA 2023     8  14           2.5 Recommended
```

Figure 3: Incomplete Data Set [Full]

3. Descriptive Statistics

The analysis of the statistical information wrapped up in the data set as a whole has been explored. The valid data has only been analysed due to accuracy and validation purposes. A detailed analysis has been done on both the categorical features as well as the continuous features of the valid data set.

The next step was to determine if any correlation and/or relationships existed between the different features.

3.1. Categorical Data

The categorical features identified in the data set were "Class" and "Why Bought". These two features were both futured in the form of a pie chart and a bar graph. The pie chart was used to represent the different proportions of the classes or the reasons why certain items were bought. By focusing on the bar graphs the frequency of each class could be identified and the reasons presented as to why certain items were bought.

From Figures 4 and 5 presented below, the pie chart and the bar graph for the Class feature can be observed.

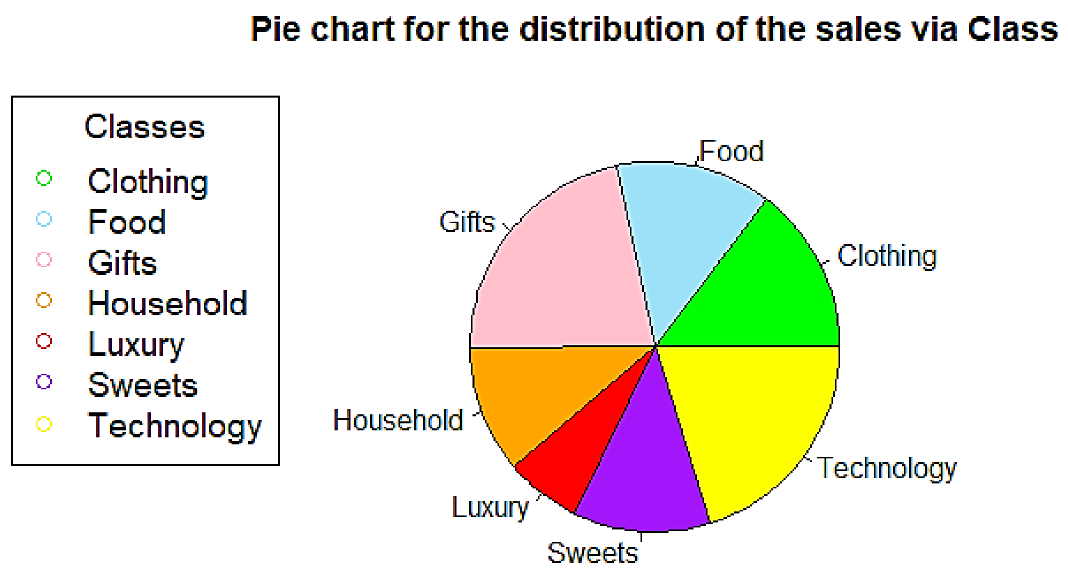


Figure 4: Pie Chart Class Distribution

As can be observed from Figure 4, the classes "Gifts" and "Technology" were the largest and "Luxury" the smallest.

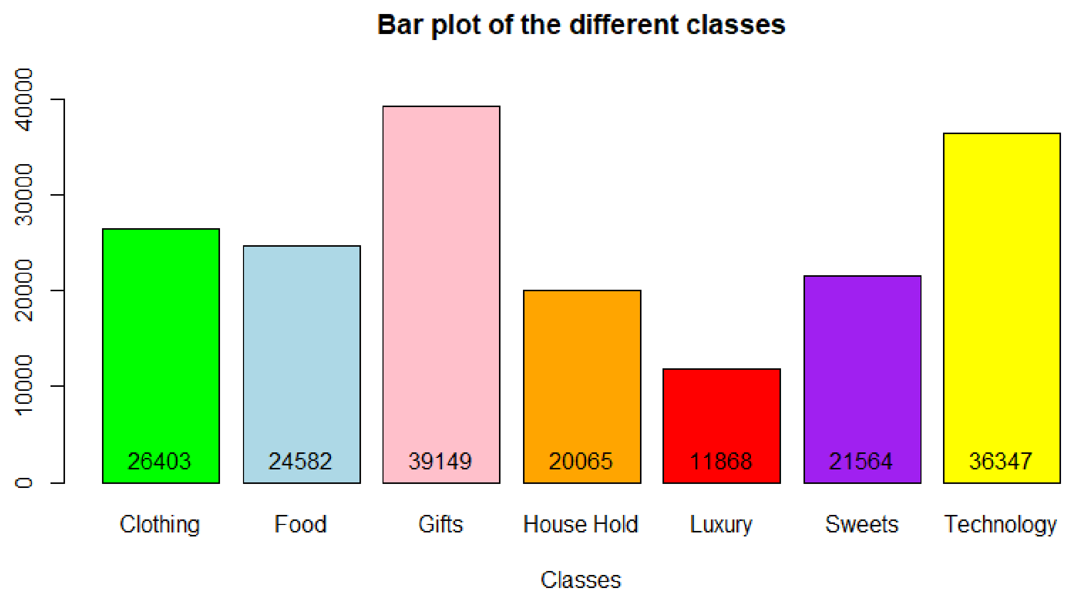


Figure 5: Bar Graph Class Distribution

From Figure 5 the frequencies of each class can easily be observed. From this the smallest, "Luxury" has a frequency of 11 868 observations, whereas the largest class, "Gifts" has a frequency of 39 149 observations.

From Figures 6 and 7 presented below, the pie chart and the bar graph for the Why Bough feature can be observed.

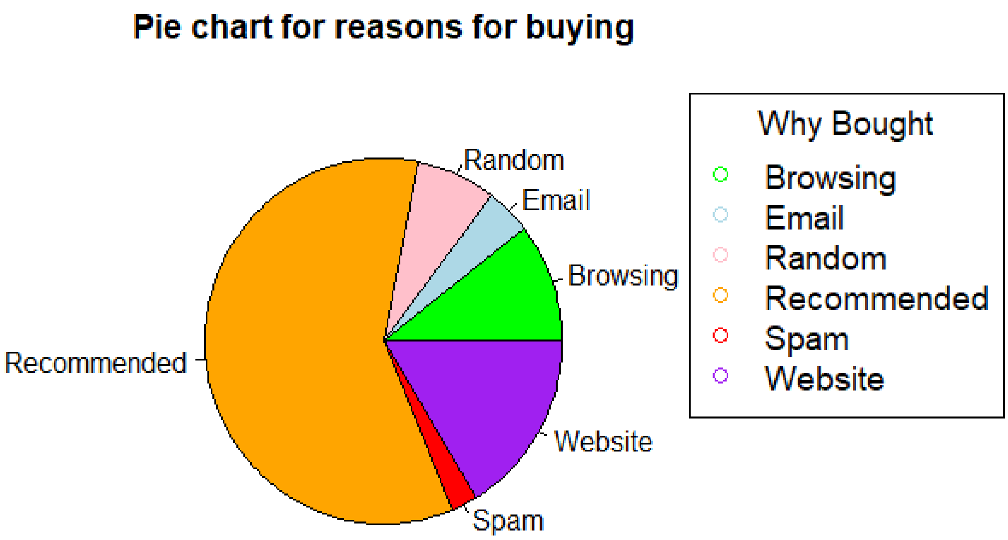


Figure 6: Pie Chart Reasons for Buying

From Figure 6 is can easily be observed that the biggest reason why certain products were bought was due to recommendation for it.

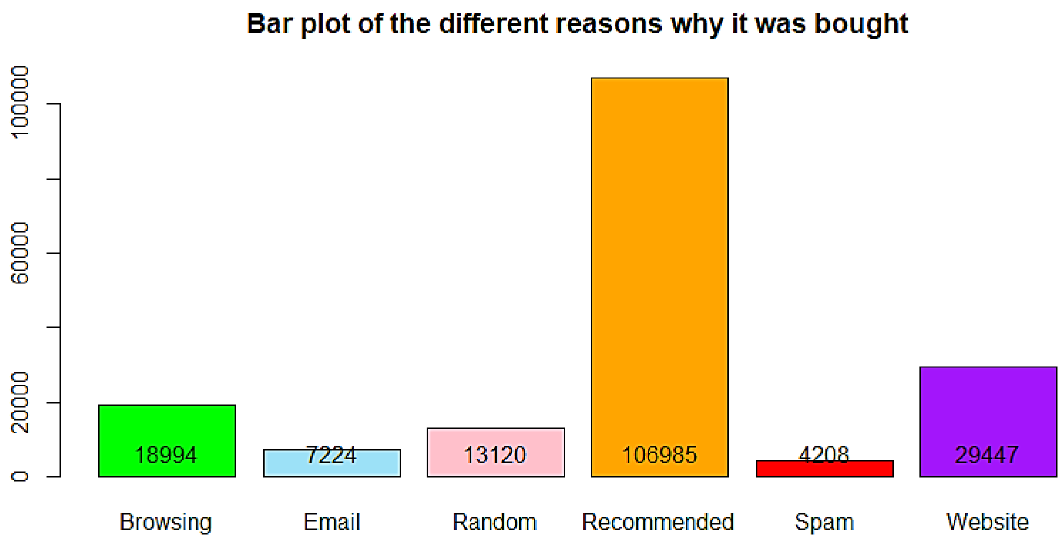


Figure 7: Bar Plot Reasons for Buying

From Figure 7 the frequencies for each different category ‘why’ products were bought can be Observed. From this the largest frequency is from the category “Recommended” which has a frequency of 106 985.

The following tool has been implemented to ensure that the data has been correctly explored. This is known as cardinality. Cardinality is known as the unique number of records that are found within a feature. This has been applied to ensure that the results obtained did line up with what was expected. The cardinality of the two features mentioned above can be found in Table 1: Cardinality – Categorical Features.

Table 1: Cardinality of Categorical Features

Feature	Cardinality
Class	7
Why Bought	6

3.2. Continuous Data

Continuous features are known to be able to take any value. These feature types are usually transformed to a numerical form due to the reason that most machine learning models require all input and output variables to be numerical. Thus, more analysis have been needed for this feature type compared to categorical feature. The continuous features identified were X, ID, Age, Price, Year, Month, Day, and Delivery Time.

Visualization methods applied for the continuous features were box plots and histogram graphs.

3.2.1. Box Plots

Box plots were plotted for the features Age, Price, Year, Month, Day, and Delivery Time to identify any outliers present in these features. This was a useful way to present the continuous data. The features X and ID were not plotted because they are only present for recognition purposes and not to obtain any output from the results. The respective box plots can be observed in Figure 8.

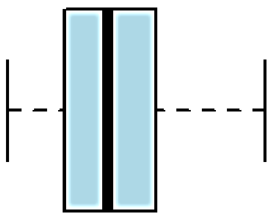


Figure 8a: Boxplot Age

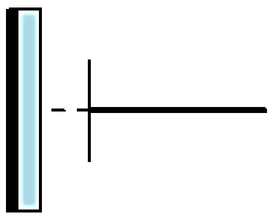


Figure 8b: Boxplot Price

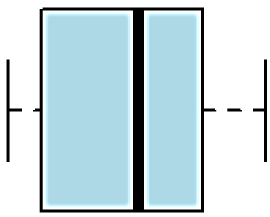


Figure 8c: Boxplot Year

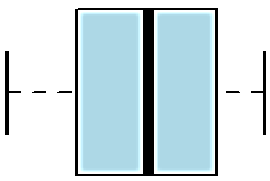


Figure 8d: Boxplot Month

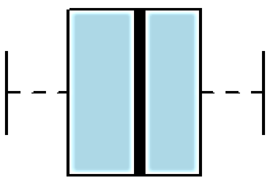


Figure 8e: Boxplot Day

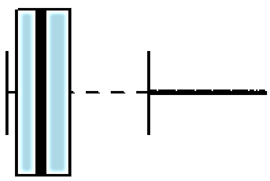


Figure 8f: Boxplot Delivery Time

From Figure 8 it can easily be seen that the features “Price” and “Delivery Time” contains multiple outliers.

For the comparison purposes of Class vs Price box plot. All classes are being plotted against the price in one graph. This created the opportunity to identify which class obtains the highest mean, interquartile range, minimum value, and maximum value. This approach has also been executed for Class vs Delivery Time box plot. Both these graphs can be seen in Figure 9 and Figure 10.

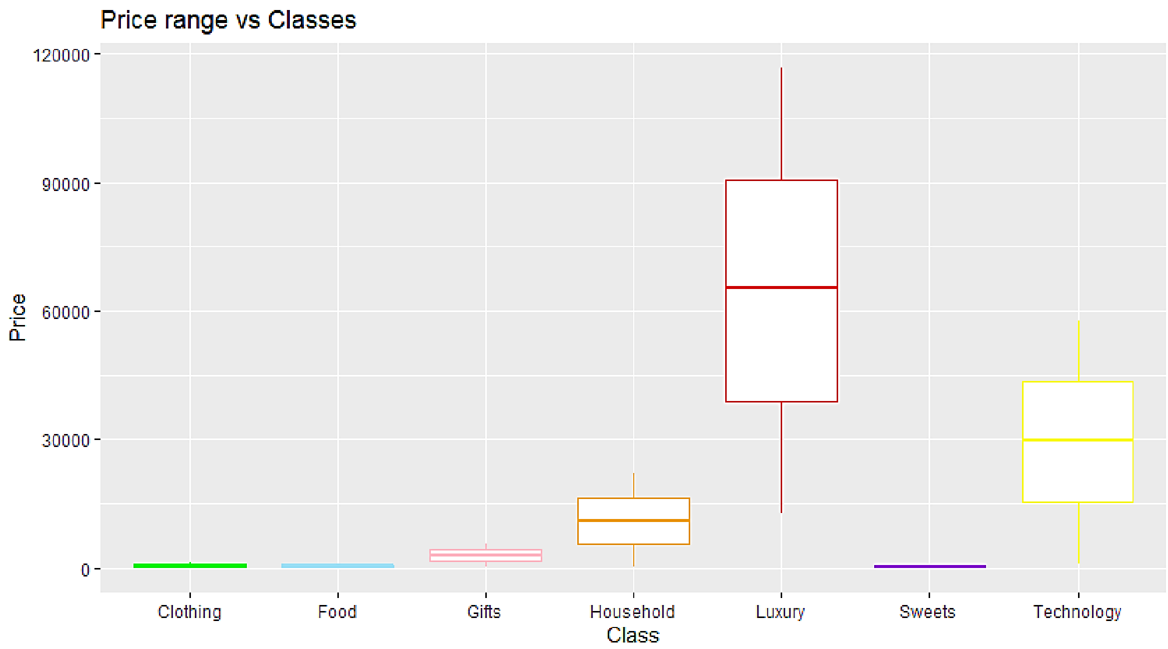


Figure 9: Boxplot Comparison Price

From the figure above it can easily be seen that the class feature “Luxury” has the highest mean price value. Thus, the assumptions can be made that luxury items are the most expensive. The class features “Clothing”, “Food”, and “Sweets” can be seen as the cheapest items.

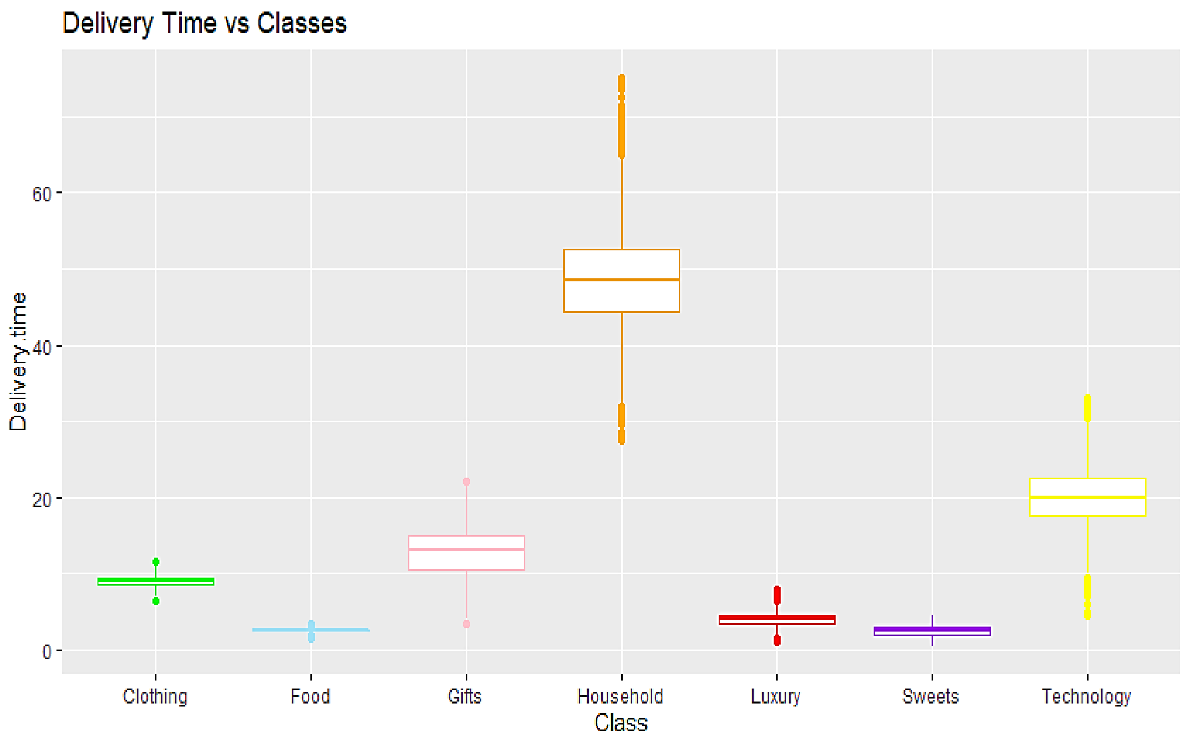


Figure 10: Boxplot Comparison Delivery Time

From the figure above a lot of outliers can be identified for the different classes over the delivery time. The class feature "Household" contains a lot of outliers which are far from the mean values which can make the data representation skew. These features were analysed in more detail later in the report.

3.2.2. Histograms

Histograms were plotted to present the frequency of each feature but also the density. The density distribution curves were plotted over the frequency graph. This contributed by easily determining which feature accumulated the most sales. The frequency histograms can be seen in Figure 11 where it can easily be seen within which price range the most sales fall into. The density distribution curves can be observed from Table 2.

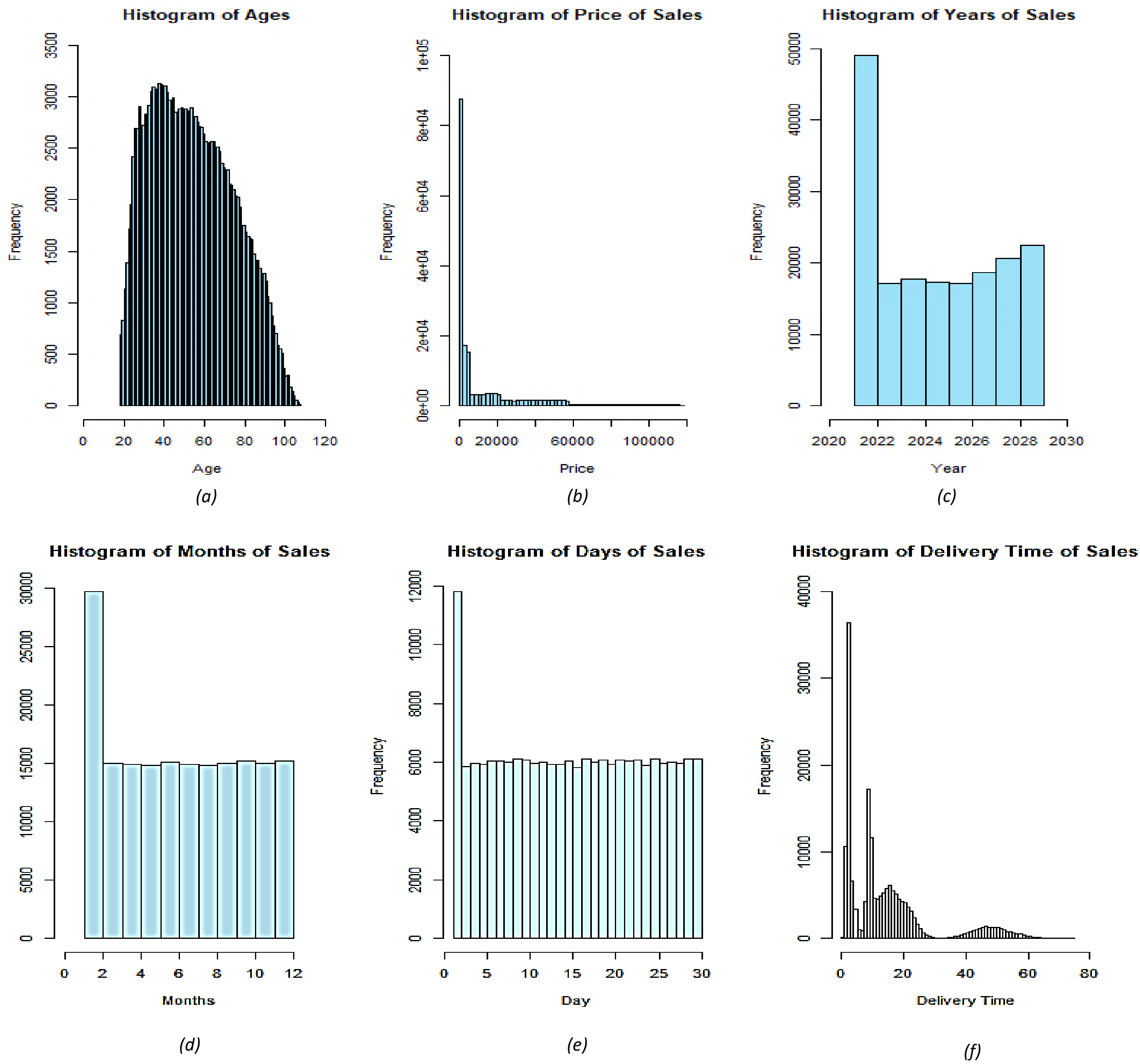


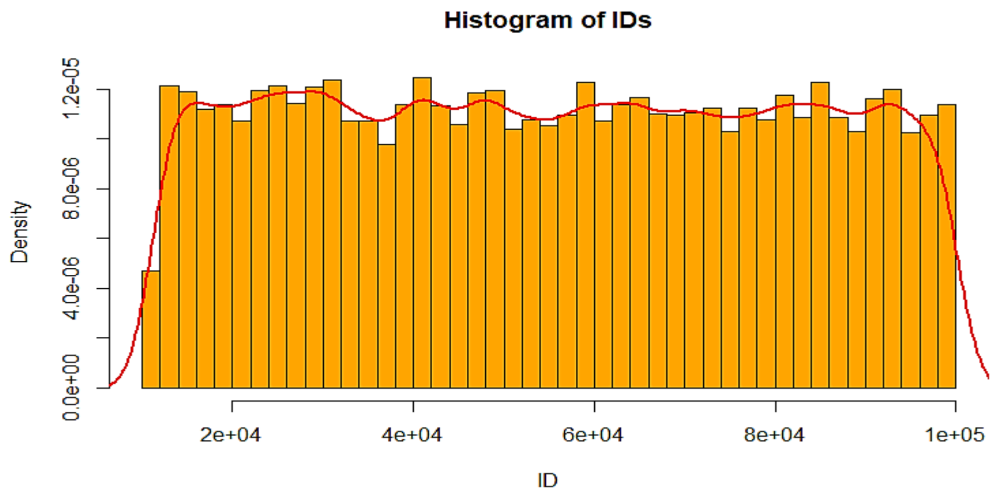
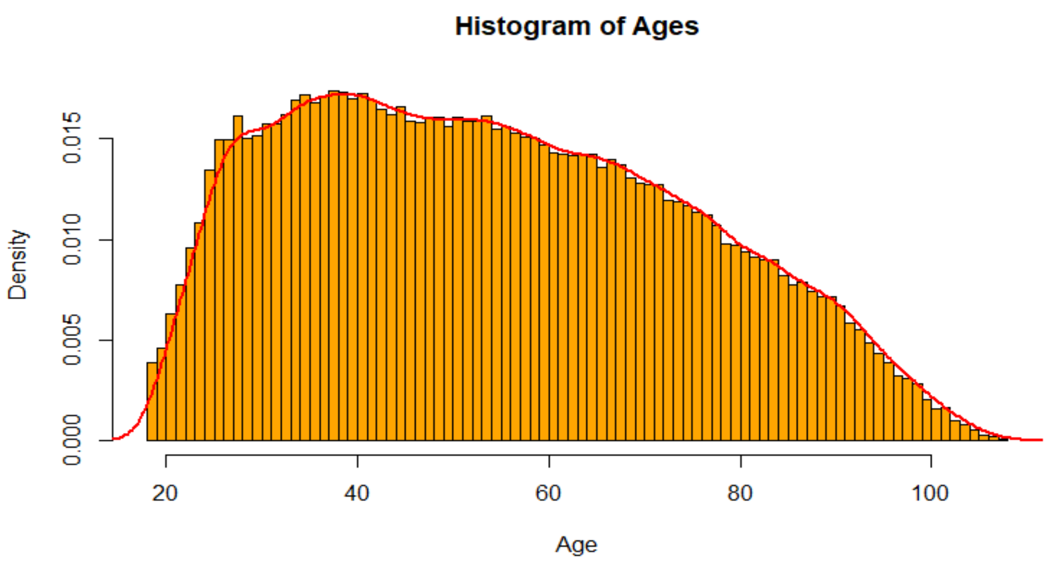
Figure 11: Histograms Frequency in each Feature

From Figure 11 the following can be observed:

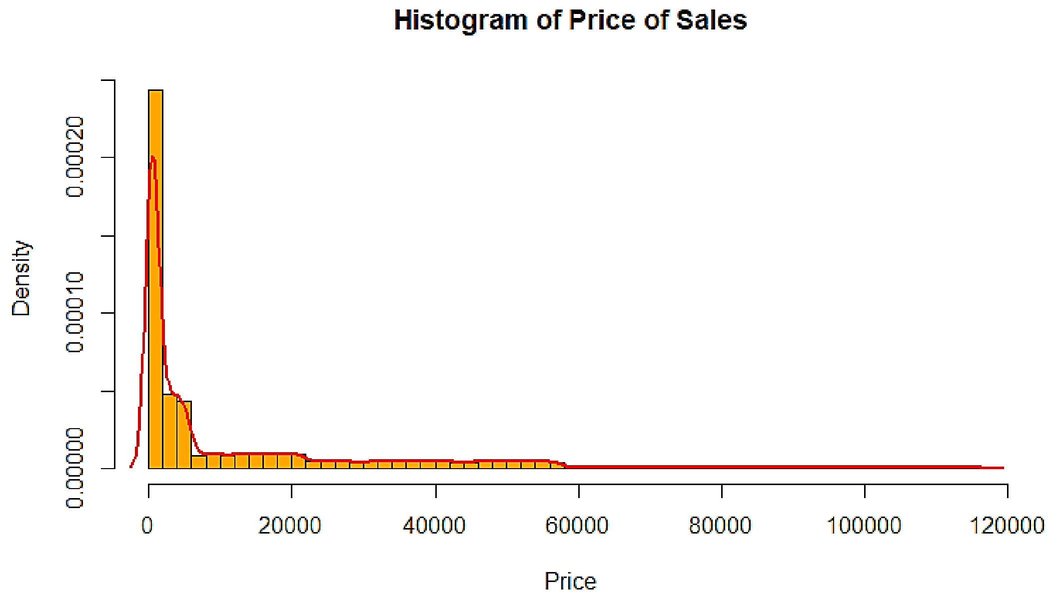
- Most sales occurred on the first day of the month.
- Most sales occurred at the beginning of the year.
- Most sales occurred at the start of the study period in terms of years. But as can be seen, sales started to increase in the last two years of the study (Figure 11.c).
- Lower prices for different items were more regularly bought.
- More deliveries have also been done in shorter time frames.

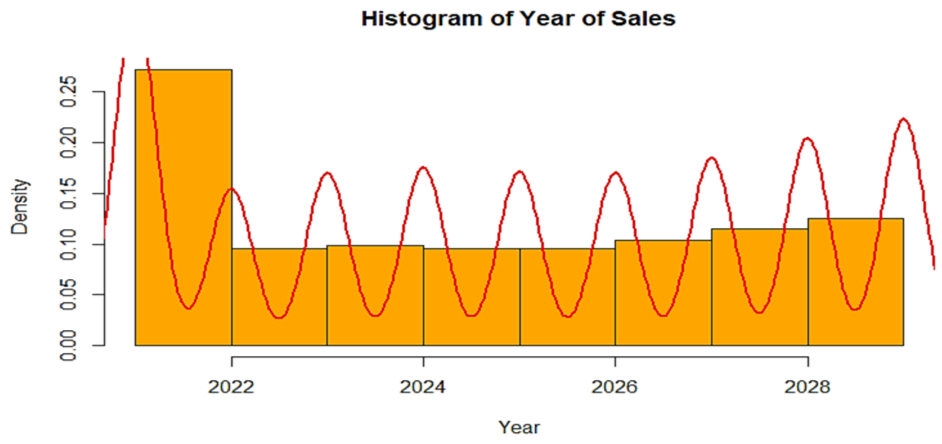
Table 2: Histogram Density Distribution

From Table 2 the density distribution of the different features can be observed over a continuous period interval as specified for each feature.

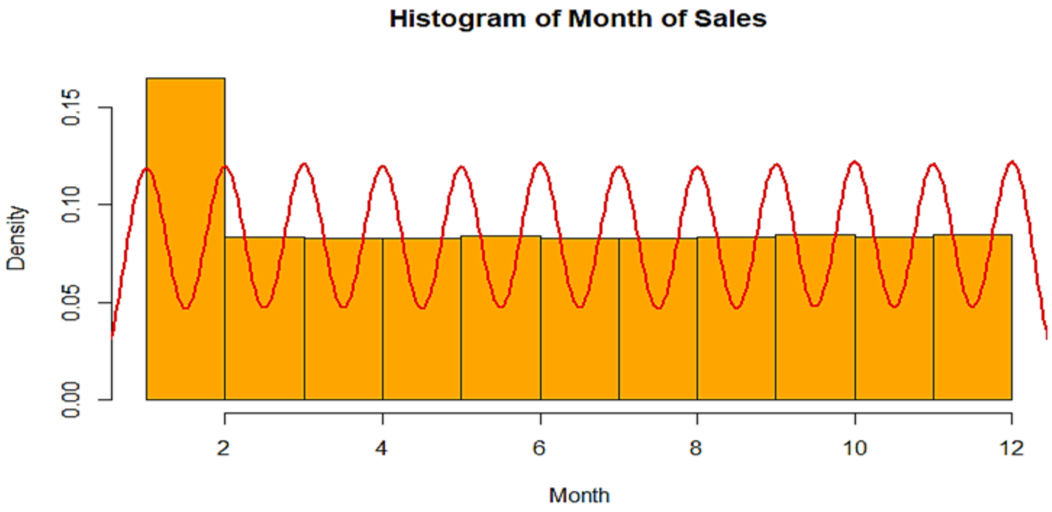
Distribution Type	**Feature**
Seen as a multimodal but can go also be seen as a uniform distribution.	
Skew left Interpretation can be done that younger age categories tend to spent/buy more than older age categories.	

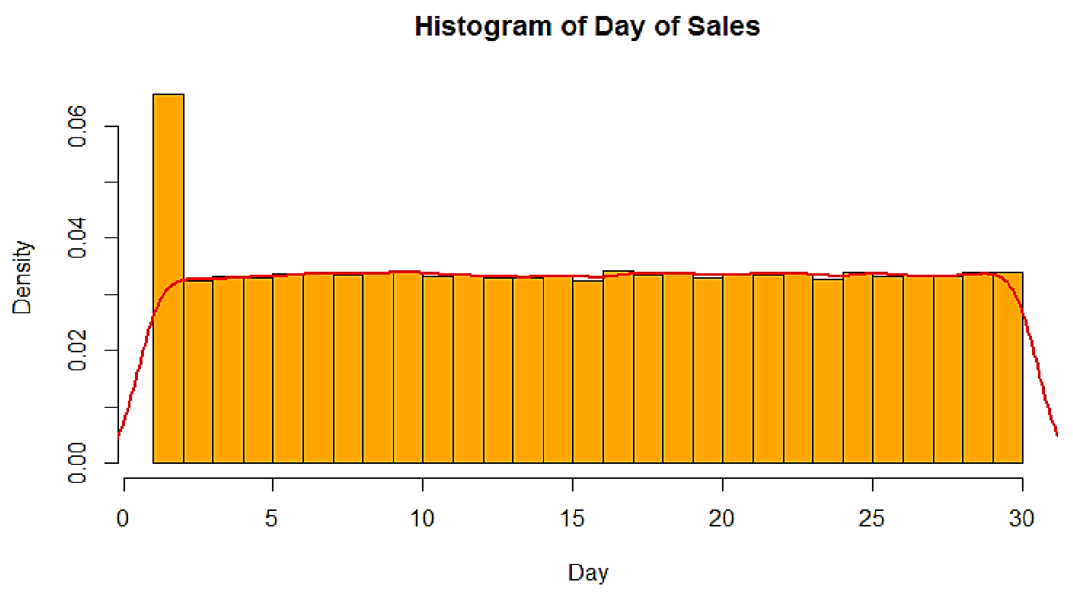
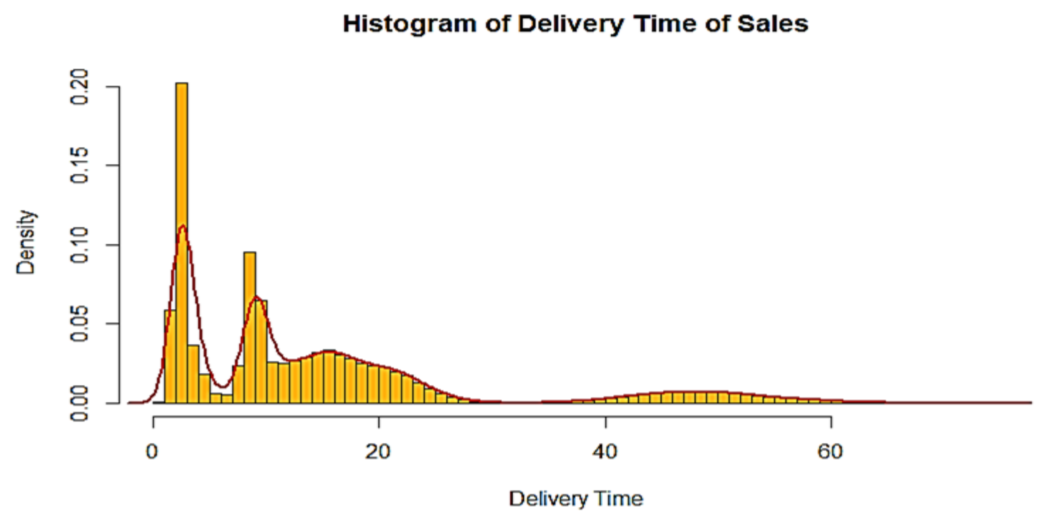
<table>
<tr><td>Exponential

The shape is most likely due to the category “luxury”. Very large values are present in this class compared to any of the values in the other classes.</td><td>
</td></tr>
<tr><td>Multimodal

Significant variation in the sales of the first year compared to more constant sale variation in the years 2022 through to year 2029.</td><td>
</td></tr>
</table>

<table>
<tr><td>Multimodal

Again, variation in sales between the months can actually be seen as quite constant except for the first month of the year (January).</td><td>
</td></tr>
</table>

Uniform, Although there is one day (first day of the month) where a lot of sales take place. The more overall distribution can be seen as uniform due to the constant density distribution over the other days.	
Multimodal Variation of delivery times between the different classes and prices can easily be seen in this distribution. Due to the interesting distribution seen, the feature delivery time will be observed in more detail later in the report. There may be trends present in the classes themselves.	

3.3. Relationships between different features

Relationship identification between features were seen as important because through this action there can be determined which features influence one another and which ones do not. Through this it can easily be determined which features contribute to business decisions and which ones are irrelevant. A feature is deemed irrelevant when it shows no correlation or relationship between any other feature.

3.3.1. Correlation Table

A correlation table helps to identify if different features have positive or negative correlation with any other feature in the table. A correlation value of 0 indicates there is no relationship between the current feature and the feature it is being compared to. The table below presents the correlation table of the valid data set evaluated.

Legend	
	Strong Positive Correlation
	Positive Correlation
	No Correlation
	Strong Negative Correlation

Table 3: Correlation Table

	ID	AGE	Price	Year	Month	Day	Delivery.time
ID	1.00	0.00	0.00	0.00	0.00	0.00	0.00
AGE	0.00	1.00	-0.14	0.13	0.00	0.00	-0.12
Price	0.00	-0.14	1.00	0.07	0.00	0.00	0.09
Year	0.00	0.13	0.07	1.00	0.00	0.00	-0.10
Month	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Day	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Delivery.time	0.00	-0.12	0.09	-0.10	0.00	0.00	1.00

From Table 3 it can be observed that there are multiple features which shows no correlation with another feature. The same number of positive and negative correlation pairs have been identified as can be seen in the table. A positive correlation is an indication that the two features operate in unison thus, when the one rises and/or falls the other feature will do the same. Whereas negative correlation is the opposite indication of unison between two features.

3.3.2. SPLOM (Scatter Plot of Matrix)

The main purpose of the SPLOM graph is to visually represent the relationship between pairs of features. The SPLOM plots scatter plots of each possible combination of the features. The SPLOM of the valid data set if shown below and the different classes are represented by the different colours.

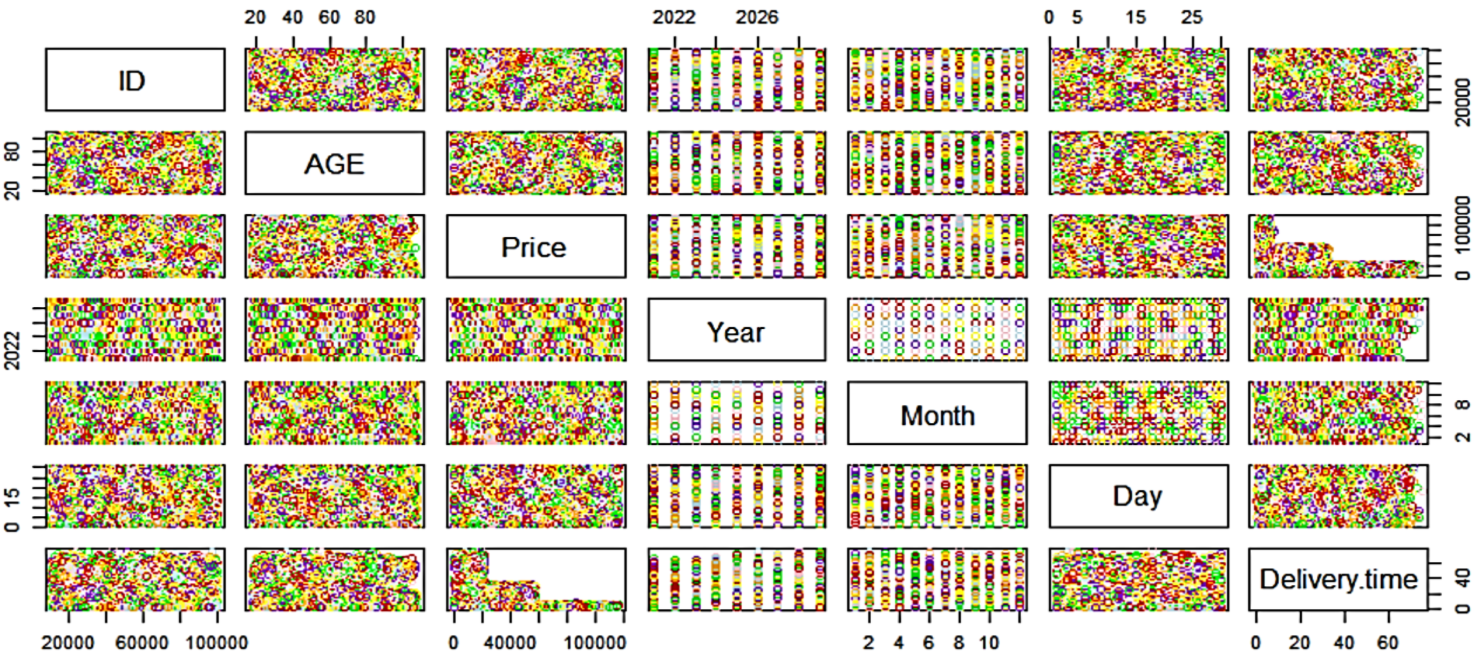


Figure 12: SPLOM Graph

The following graph methods look similar to the SPLOM presentation but represents different data results. The chart.correlation() function plots histograms of the features down the diagonal showing the distribution, scatter plots with lines joining each feature, and then the correlation in the upper panel. Three different charts have been plotted by making use of this function, three different methods of evaluation have been applied.

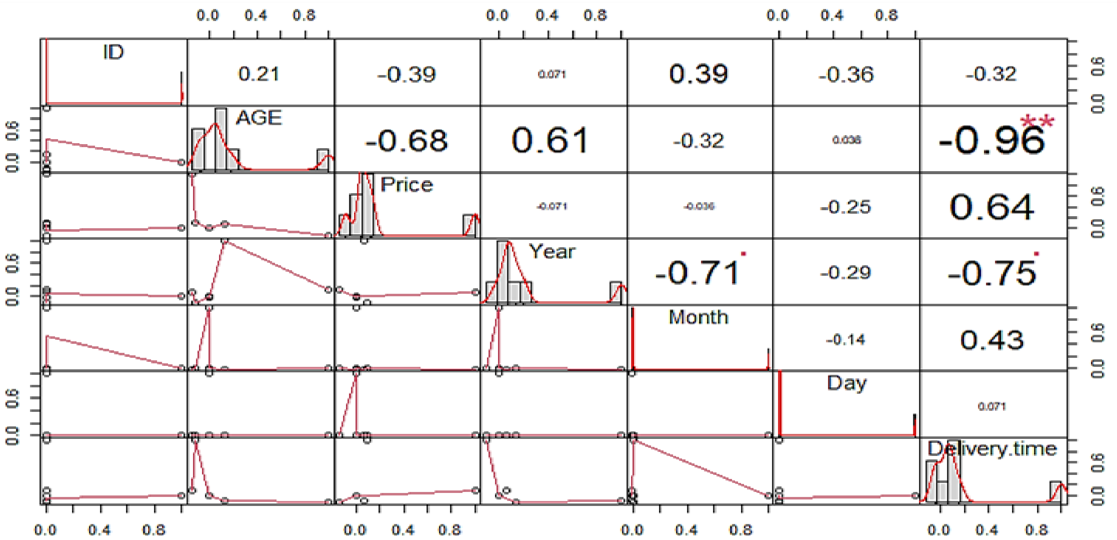


Figure 13: Spearman – representing correlated data which is normally distributed

From above the relationship between two variables is described by making use of a monotonic function. High correlation has been detected when observations have similar rank. Otherwise, dissimilar rank will be seen as low correlation (Anon., 2022).

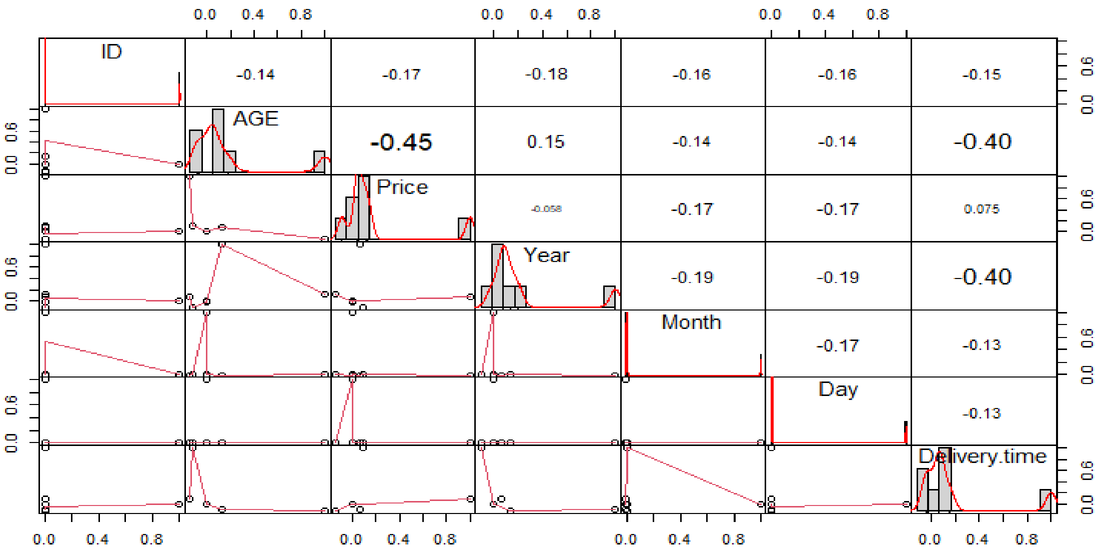


Figure 14: Pearson – representing correlated data which is normally distributed

The Pearson correlation method refers to the Pearson correlation factor which specifically refers to the linear relationship between two variables. This method can only be applied when the relationship between two features are seen as significant, this can be a negative or a positive linear relationship (J. Puranen, n.d.).

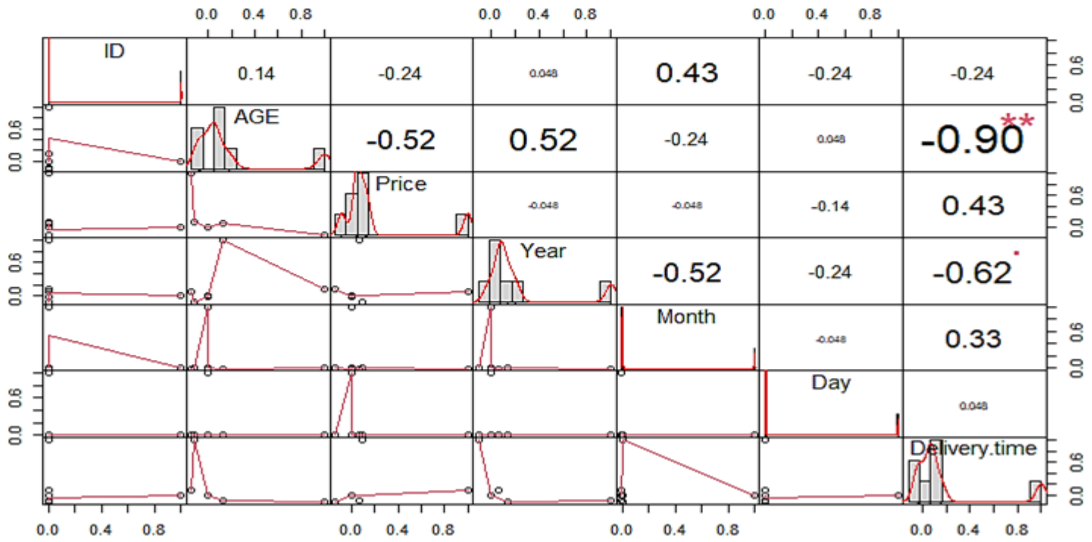


Figure 15:Kendall – representing a test of strength dependencies

The Kendall method is applied to measure the ordinal association between two measured features. This refer to the similarity of the different orderings of the data when ranked by each of the quantities. As can be seen from the figure above, the Kendall correlations are quite low in general. This means that the strength of the linear relationships between the different features are low (Nickolas, 2021) (Anon., 2022).

3.3.3. Correlation Plot

Another method to be applied to represent and visualize the correlation and relationship between different feature pairs are by the correlation plot function [corrplot()]. The correlation of the different features are being displayed by the function by different sized and coloured dots. The big and dark dots represents a stronger correlation between those features whereas a lighter and smaller dot represents less correlation between those features.

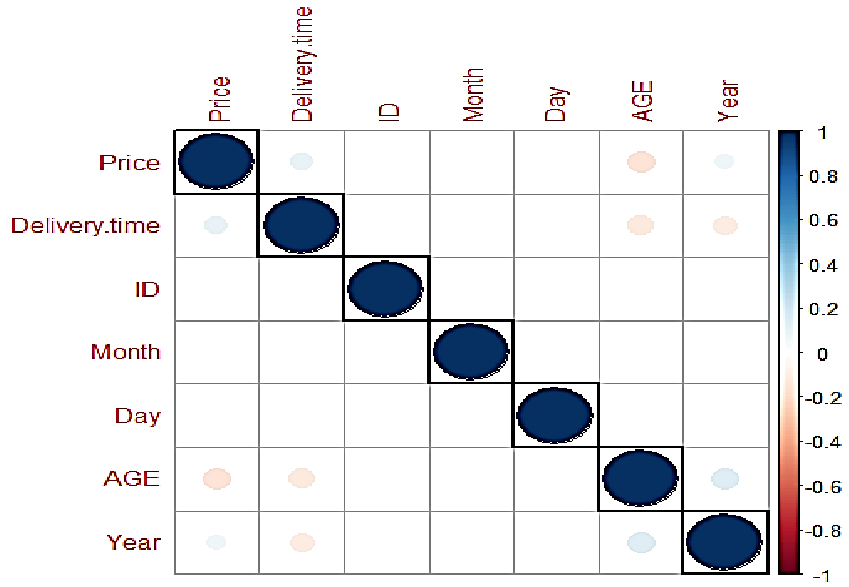


Figure 16: Correlation Plot

3.3.4. Heat Map

The last and final method to present correlation and relationships between different features and different pairs of features are Heat Maps. Heat Maps use different shades of colours to represent the difference in strength of the different relationships between the features. The strength is based on the darker the colour the stronger the correlation between those features.

Strong relationships have been observed between the following features:

- Price and Delivery Time (Dark green)
- Age and Year (Dark green)

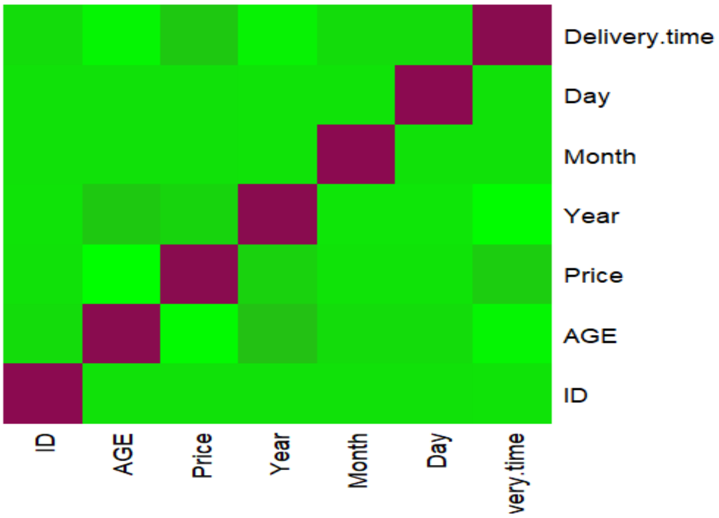


Figure 17: Heat Map

3.3.5. Process Capability Indices

From the project description and guidelines, it was necessary to determine the different process capability indices which are Cp, Cpu, Cpl, and also Cpk specifically for the delivery times of the technology class. The given value for the USL is 24 hours and LSL is 0.

To calculate these indices the mean and the standard deviation of the technology class and delivery time were determined.

SD = 3.501993
Mean = 20.01095

$$Cp = \frac{(USL - LSL)}{6 * SD} = \frac{(24 - 0)}{6 * 3.501993} = 1.142207$$

$$Cpl = \frac{(Mean - LSL)}{3 * SD} = \frac{(20.01095 - 0)}{3 * 3.501993} = 1.90472$$

$$Cpu = \frac{(USL - Mean)}{3 * SD} = \frac{(24 - 20.01095)}{3 * 3.501993} = 0.3796933$$

$$Cpk = \min(Cpl, Cpu) = 0.3796933$$

Figure 18: PCI Calculations

4. Statistical Process Control (SPC)

The main focus in this section was placed on the implementation of SPC (Statistical Process Control) for the X- and S charts. This was done by ordering the data from the oldest to the newest as requested by the question stated. The data was placed in samples, to be exact, 30 samples were taken from the ordered data which consisted of fifteen sales each. Thus, this made up the first 450 entries of each class. The samples obtained from each class was used to determine the centre lines [CL], outer control limits [OCL], 2-sigma, and 1-sigma control limits for both the X-charts and the S-charts.

4.1. X-Charts, S-Charts, Results from Charts [30 samples, 15 sales each]

From the tables below, Table- 4 and 5, the control limits of the X-chart and the S-chart for the delivery times for each class can be seen respectively. The feature delivery time has been inspected in more depth due to the big number of outliers detected in this feature.

Table 4: X-Chart

	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	9.404681	9.259787	9.114894	8.97	8.825106	8.680213	8.535319
Food	2.70933	2.63622	2.56311	2.49	2.41689	2.34378	2.27067
Gifts	9.487909	9.11231	8.73671	8.361111	7.985512	7.609913	7.234313
Household	50.24618	49.0182	47.79021	46.56222	45.33423	44.10625	42.87826
Sweets	2.896798	2.757124	2.617451	2.477778	2.338104	2.198431	2.058758
Luxury	5.493524	5.240868	4.988212	4.735556	4.482899	4.230243	3.977587
Technology	22.9731	22.10688	21.24066	20.37444	19.50822	18.64201	17.77579

Table 5: S-Chart

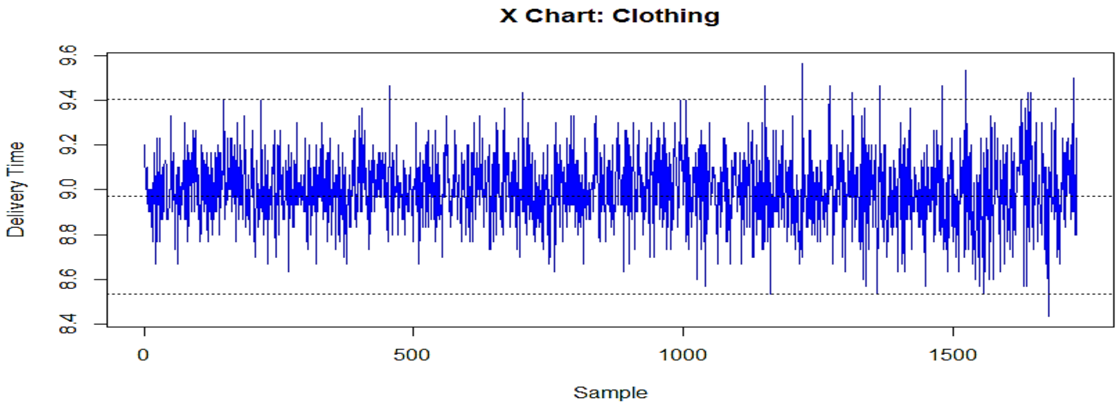
	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	0.8664	0.7613	0.6563	0.5512	0.4462	0.3411	0.2360
Food	0.4372	0.3842	0.3312	0.2781	0.2251	0.1721	0.1191
Gifts	2.2460	1.9737	1.7013	1.4290	1.1566	0.8842	0.6119
Household	7.3432	6.4528	5.5623	4.6719	3.7814	2.8910	2.0004
Sweets	0.8352	2.7955	2.6942	0.5314	2.2613	2.1600	0.2275
Luxury	1.5109	1.3277	1.1444	0.9612	0.7780	0.5948	0.4116
Technology	5.1799	4.5518	3.9237	3.2955	2.6674	2.0393	1.4111

The X-charts and S-charts are being used to monitor a business through statistical processing. For the X-chart the mean of the variables are being monitored over a period of time. The S-chart is based on the X-chart because the standard deviation is estimated from the mean.

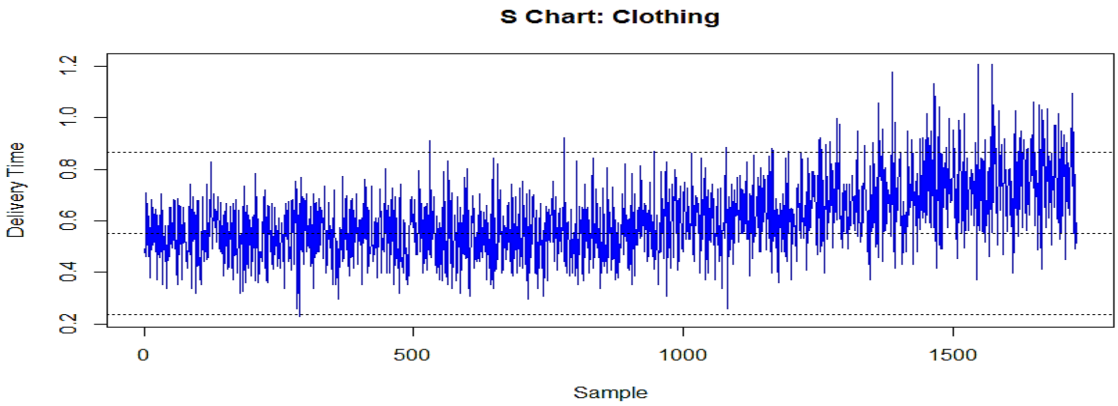
4.2. X-Charts, S-Charts, Results from Charts [Rest of the data]

This section focuses on the remaining of the data that did not form part of the first thirty samples taken. But this data was processed similarly as before, split into samples of fifteen sales in each.

4.2.1. Clothing



(a)

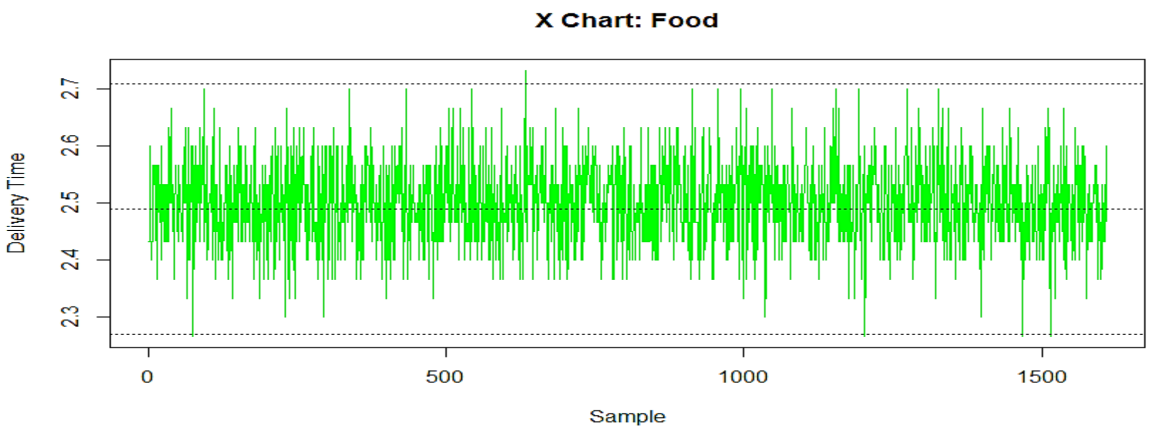


(b)

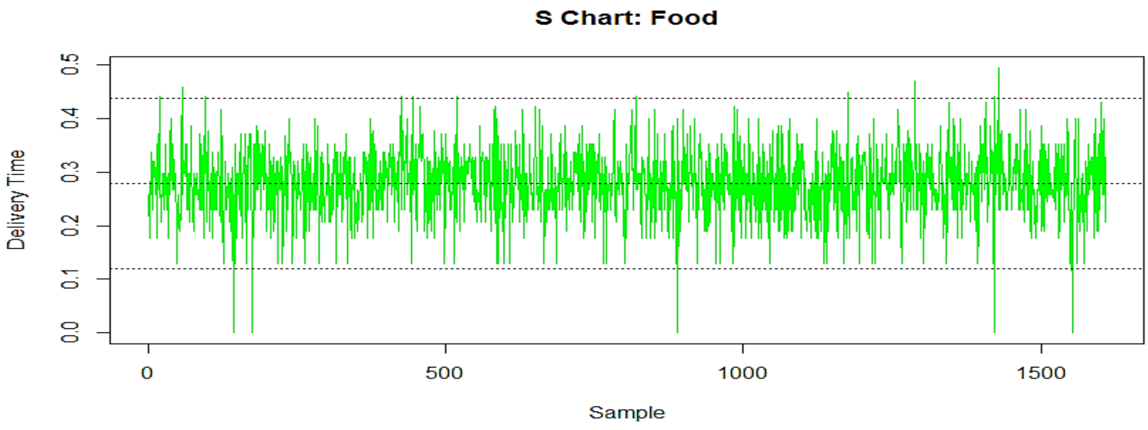
Figure 19: X-Chart and S-Chart for Clothing

The X- and S-Charts for the class Clothing can be observed in Figure 19 (a) and (b) respectively. Very little number to none of the samples fall outside the upper and lower control limits of the X-Chart. Whereas from the S-Chart quite a few samples exceed the upper control limit, this samples are observed closer to the end of the sample data set.

4.2.2. Food



(a)

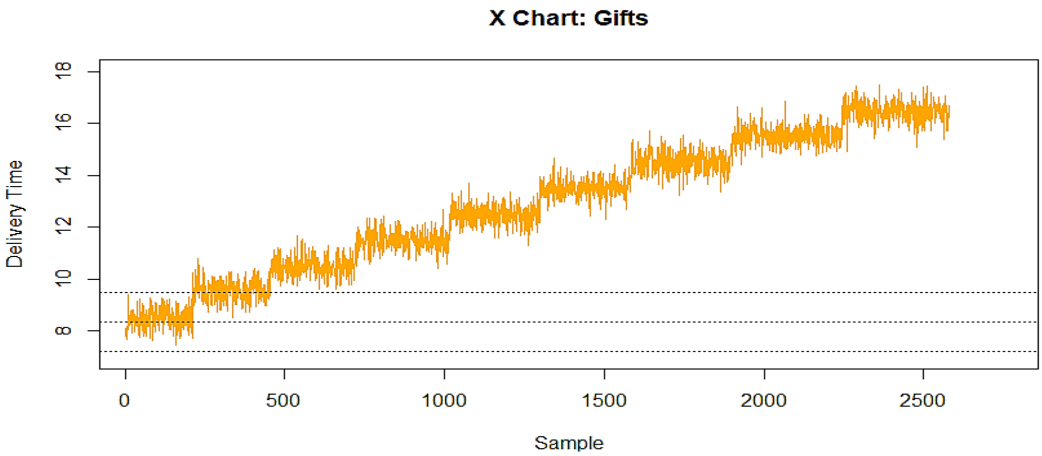


(b)

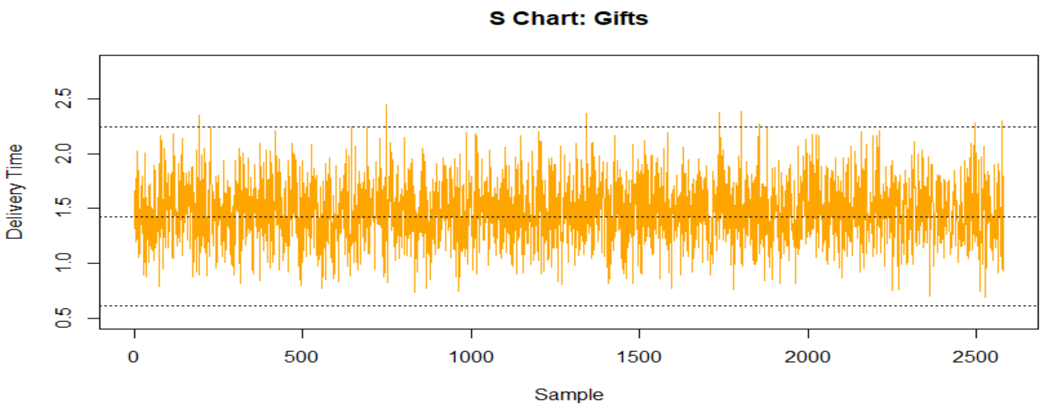
Figure 20: X-Chart and S-Chart for Food

The X- and S-Charts for the class Food can be observed in Figure 20 (a) and (b) respectively. Both the X-Chart and the S-Chart contain outlying samples. Number of outliers for the S-Chart is quite more than those observed for the X-Chart. Due to the little number of outliers in the X-Chart we can say the process is in-control.

4.2.3. Gifts



(a)

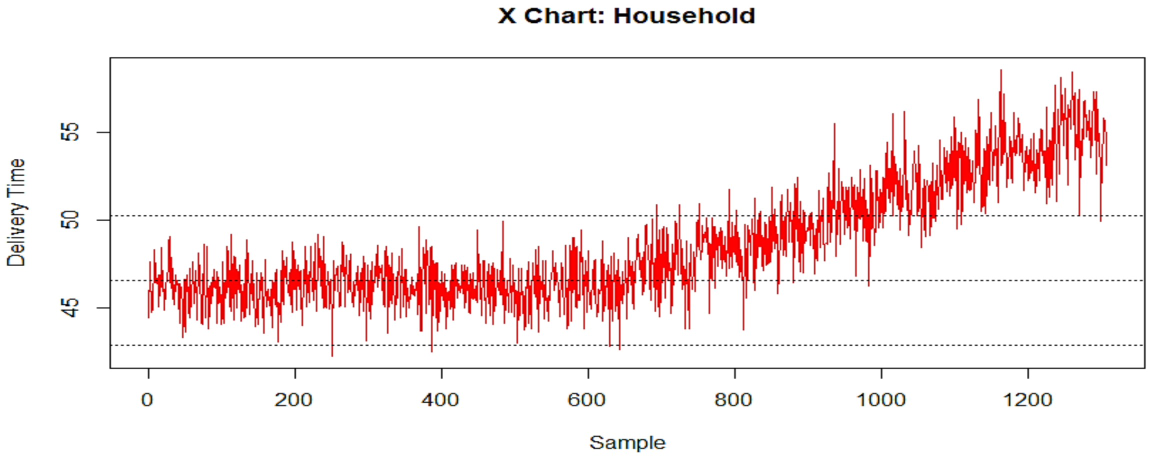


(b)

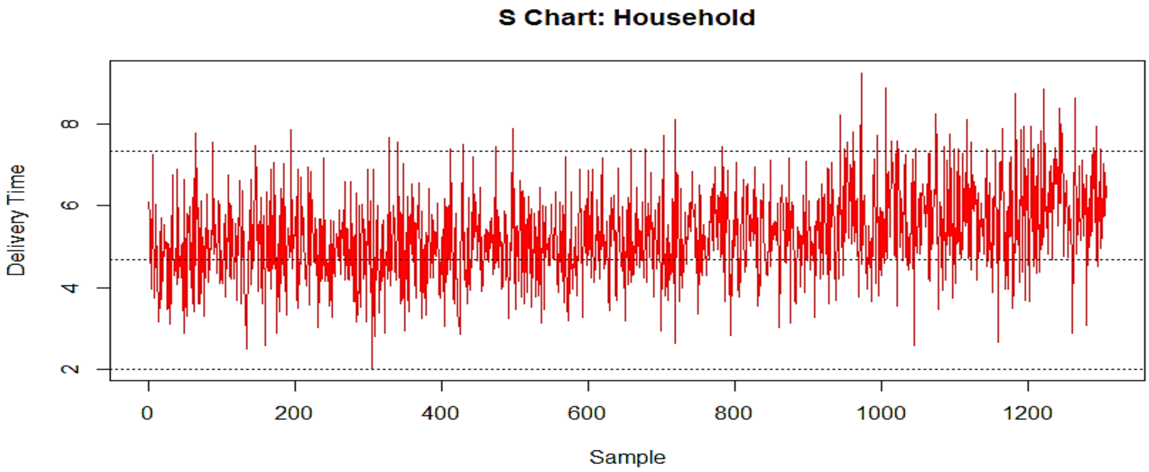
Figure 21: X-Chart and S-Chart for Gifts

The X- and S-Charts for the class Gifts can be observed in Figure 21 (a) and (b) respectively. From all the classes observed, the Gifts class has the results with the most samples which is out of control, especially the X-Chart. From the X-Chart of the "Gifts" class an upward trend can be identified. This trend can be confirmed by the visualisation that the further you go on in the samples, the greater the delivery time becomes. This is also seen as the class with the greatest number of samples which are seen as out of control, this can be observed from Table 6 as be 2261. Due to this big number of samples being out of control this sample is deemed to be out of control and a proposal will be made to the client of the online business that this class needs to be investigated to why there is so many outliers present in the samples taken.

4.2.4. Household



(a)

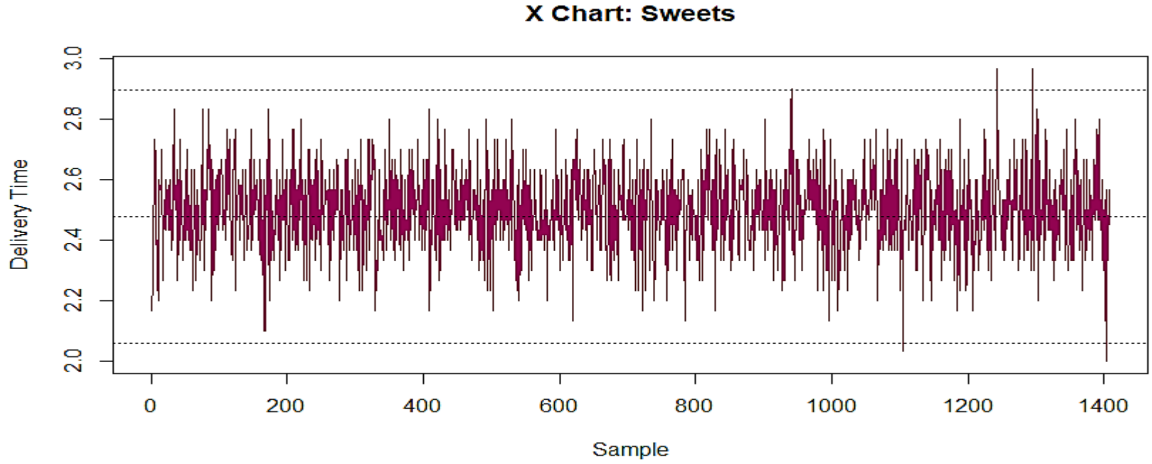


(b)

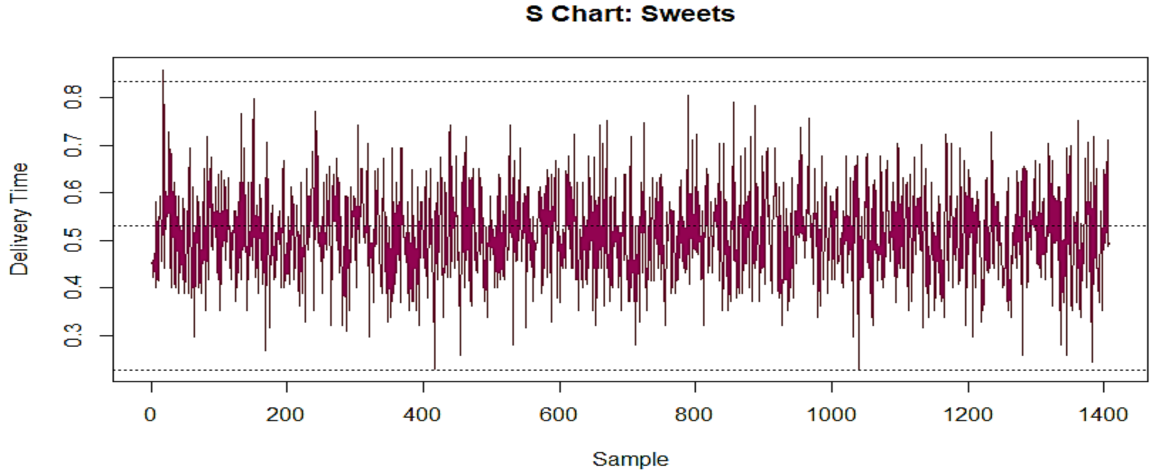
Figure 22: X-Chart and S-Chart for Household

The X- and S-Charts for the class Household can be observed in Figure 22 (a) and (b) respectively. From the X-Chart observation it can easily be seen that there is an increasing trend formed by the results of the samples. This is only more evident later on in the samples. Thus, again a large portion of the samples exceed the upper control limit. The number of observations seen as out of control can be observed from Table 6. Again, the statement is valid that the class household's delivery time increases over the period of time. Due to the large number of samples being out of control, being above the upper control limit, there will be a proposal made to the client of the online business that this class must be investigated why there are so many outliers present in the samples presented.

4.2.5. Sweets



(a)

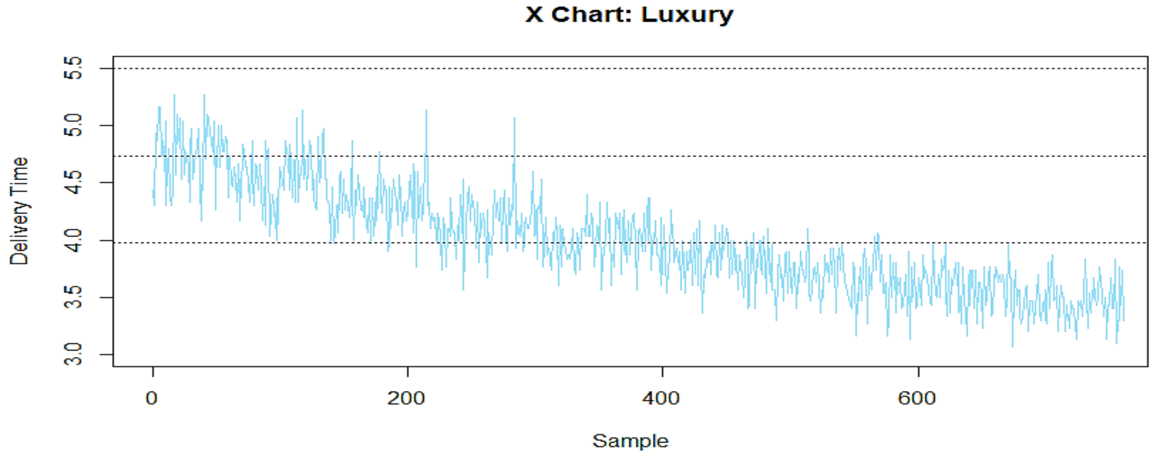


(b)

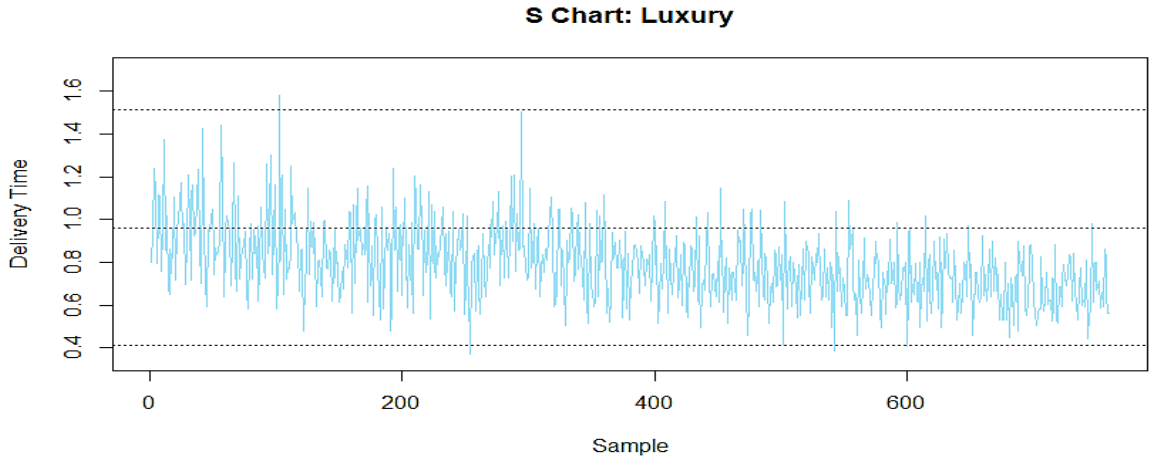
Figure 23: X-Chart and S-Chart for Sweets

The X- and S-Charts for the class Sweets can be observed in Figure 23 (a) and (b) respectively. Observation between the X-Chart and the S-Chart can be seen as very constant. No, extreme outliers identified above or below the upper and lower control limit values. This can indicate that the process is indeed in-control, compared to a process which contains a lot of outliers above or below the upper and lower control limit values.

4.2.6. Luxury



(a)

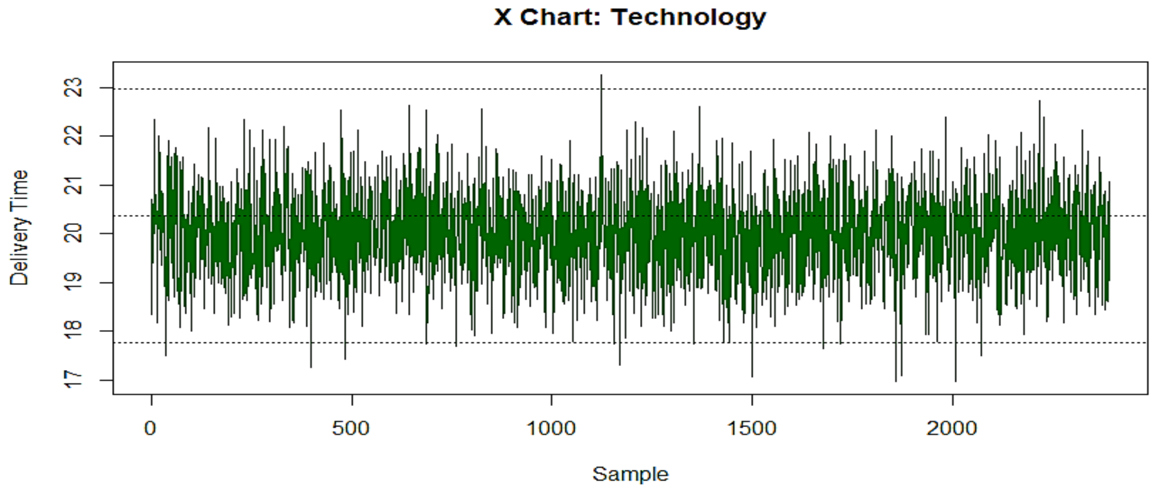


(b)

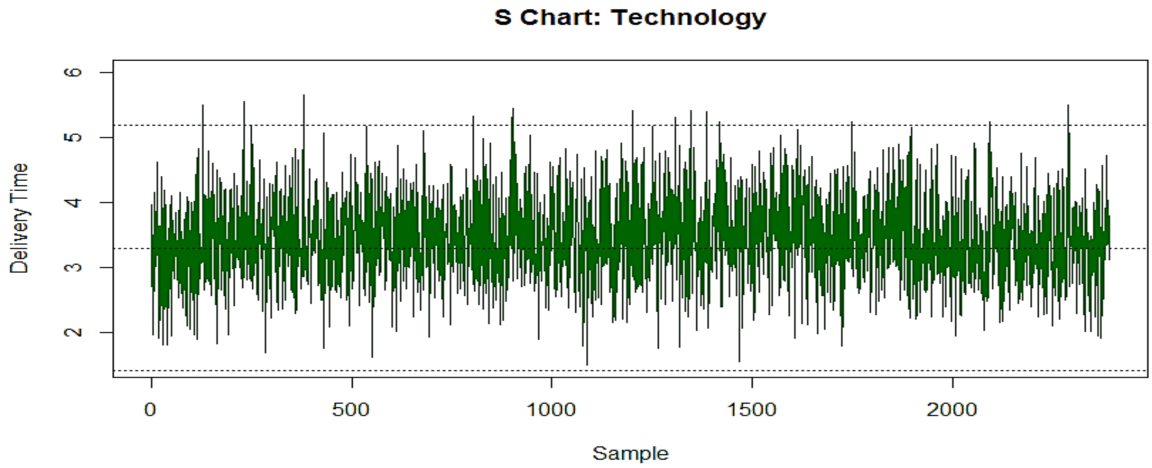
Figure 24: X-Chart and S-Chart for Luxury

The X- and S-Charts for the class Luxury can be observed in Figure 24 (a) and (b) respectively. From both the X-Chart and the S-Chart a continuous downward trend can be observed. A lot of samples fall outside of the lower control limit for the X-Chart as can be seen. These observations can be confirmed by Table 6 and Table 7 where the number of outliers for both the X-Chart and the S-Chart within the "Luxury" class can be found. Although very little of the values on the S-Chart fall outside of the upper and lower control limits, the process as a whole can be seen as out of control. Due to the large number of samples being out of control, being under the lower control limit, there will be a proposal made to the client of the online business that this class must be investigated why there are so many outliers present in the samples presented.

4.2.7. Technology



(a)



(b)

Figure 25: X-Chart and S-Chart for Technology

The X- and S-Charts for the class Technology can be observed in Figure 25 (a) and (b) respectively. Both the X-Chart and the S-Chart for the class "Technology" can be seen as in-control due to the little number of samples that lie beyond the upper and lower control limits. Again, these observations can be confirmed by the number of outliers observed from Table 6 and Table 7 for the X-Chart and S-Chart respectively.

5. Optimising Delivery Process

The focus for this part of the project is based on the whole data set.

5.1. Sample numbers out of control

5.1.1. Samples outside of outer control limits

From the tables below [Table 6 and Table 7], the different classes are shown with the total number of samples that are out of control in each class. The classes that showed numerous samples being out of control, the first three instances and the last three instances have been given as requested.

Table 6: Samples out of Control for X-Charts based on the first 30 samples

Class	Total samples out of control	1st	2nd	3rd	3rd last	2nd last	1st last
Clothing	17	455	702	1152	1677	1723	1724
Food	5	75	633	1203	NA	1467	1515
Gifts	2261	213	216	218	2578	2579	2580
Household	371	252	387	629	1306	1307	1308
Sweets	5	942	1104	1243	NA	1294	1403
Luxury	404	142	171	184	759	760	761
Technology	17	37	398	483	1872	2009	2071

Table 7: Samples out of Control for S-Charts based on the first 30 samples

Class	Total samples out of control	1st	2nd	3rd	3rd last	2nd last	1st last
Clothing	89	289	530	780	1720	1722	1725
Food	16	19	57	96	1422	1429	1553
Gifts	8	193	746	1342	1855	2493	2576
Household	53	65	89	147	1290	1294	1299
Sweets	1	18	NA	NA	NA	NA	NA
Luxury	4	103	254	543	NA	NA	600
Technology	15	129	230	251	1750	2095	2290

5.1.2. Longest consecutive pattern in S-bar values between +0.4 Sigma and -0.3 Sigma

The table below [Table 8] shows the longest consecutive S-bar values that falls between the +0.4-sigma line and the -0.3-sigma line. Although a small range are being used to make this observation, the longest consecutive pattern identified is 7 being in the class "Food".

Table 8: Longest consecutive run of S-bar values as specified

Class	Longest pattern between +0.4 sigma and -0.3 sigma
Clothing	4
Food	7
Gifts	5
Household	3
Sweets	4
Luxury	4
Technology	6

5.2. Estimate Type I Error

5.2.1. Referring to section 5.1.1.

Looking into section 5.1.1. it refers to the values that fall between the upper and the lower control limits. These upper and lower control limits can be taken as +3.0-sigma values away and -3.0-sigma values away from the central control limit [theoretical wise]. Thus, to determine the chance for a type I error to occur we obtain the probability between the values that corresponds to the Z-value of +3.0 and -3.0. The results obtained are as follow:

$\left(1 - pnorm(3)\right) * 2 = 0.002699796$ which correspond to a value for the probability of 0.2699796 %.

5.2.2. Referring to section 5.1.2.

Looking into section 5.1.2. refers to the values that fall between an upper control limit of +0.4-sigma value and a lower control limit of -0.3-sigma value away from the central control limit. To determine the chance for a type I error to occur, the same procedure has been followed as in section 5.2.1. to obtain the probability as stated. The results obtained are:

$pnorm(0.4) - pnorm(-0.3) = 0.2733332$ which corresponds to a value for the probability of 27.33332 %.

5.3. Optimizing Delivery Time for Technology Items relating to Cost

The focus for section was directed towards the individual delivery times of the class "Technology" and not the samples created originally. The following information was given regarding the calculations needed to be done.

The money lost per item which was delivered late per hour is given as R329/ item. This was applied for each item delivered slower than 26 hours. To reduce this average delivery time of a single item by one hour incurs a fee of R2.5/item/hour. The task was given to find an optimum solution for the delivery hours which produces the best profit for the online business. This was addressed assuming that the process output distribution holds the same shape when the centre was moved and will cost you less (-R 2.5/item/hour) if the delivery time was increased.

Based on the calculations done, the current average delivery time for the class "Technology" is 20.01095 hours which costs the company a total amount of R 329 063,50. The advice obtained from the calculations done is that the company should reduce the delivery time by 4 hours to gain a new average of 16.01095 hours. This reduced delivery time will cost the company a total amount of R 298 201. By the reduction in delivery time for the class "Technology", the online business of the client will save R 30 862.50.

The following figure (Figure 26) shows the minimum cost obtained by incorporating the new average of 16.01 hours.

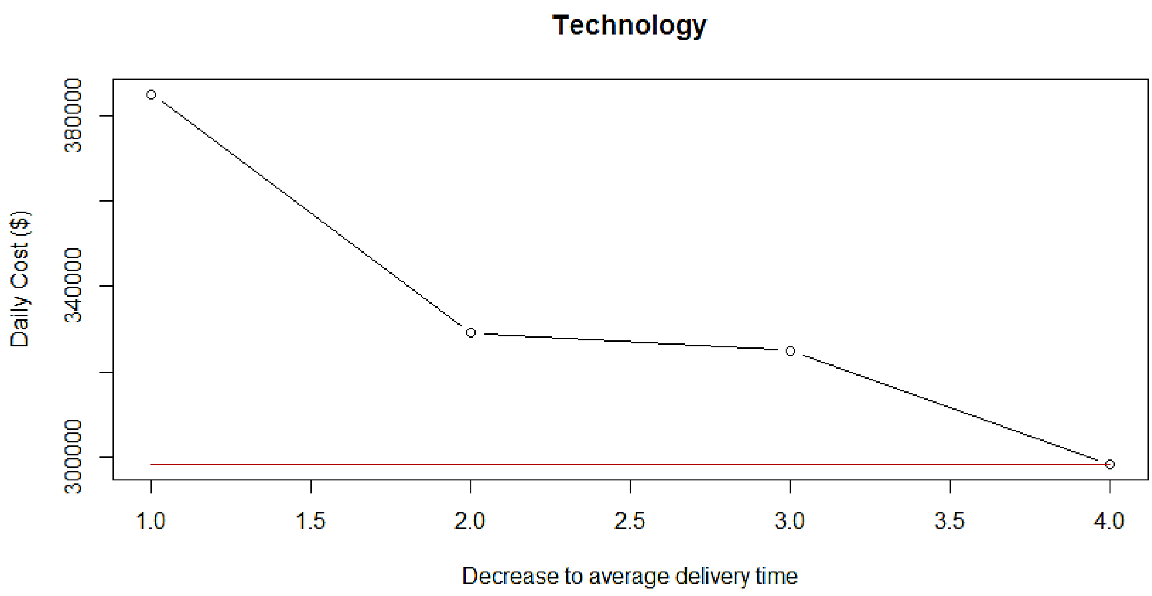


Figure 26: Minimum cost per hour obtained for the class Technology Delivery Time

Addressing the concepts of performance excellence (productivity, cost, quality) will ensure Customer and Producers satisfaction. This is one of the company goals to be achieved. From the calculations above the costs have been addressed successfully for the class Technology vs its Delivery Time. With this the delivery time has been minimised which in return satisfied the customers. The delivery time has been minimised by reducing the costs of all expenses and thus, ensures a good profit for the company. This in return will satisfy the producers.

5.3.1. Estimate Type II Error

A type II error is known as a false positive which states that a value to be in-control (within the calculated upper and lower control limits), when it exceeds the limits in reality as stated, and is then seen to be actually out-of-control.

The class technology was examined for the rule A and can be seen in the figure (Figure 27) below. The upper control limit (UCLT) has been found as 22.9731, and the lower control limit (LCLT) has been found as 17.77579. The mean is seen as 23, and the standard deviation is taken as 0.8662197.

The chance of a type II error occurring was estimated by making use of the following formula:

$$pnorm(UCLT, \mu, \sigma) - pnorm(LCLT, \mu, \sigma) = \ \ 48.76147$$

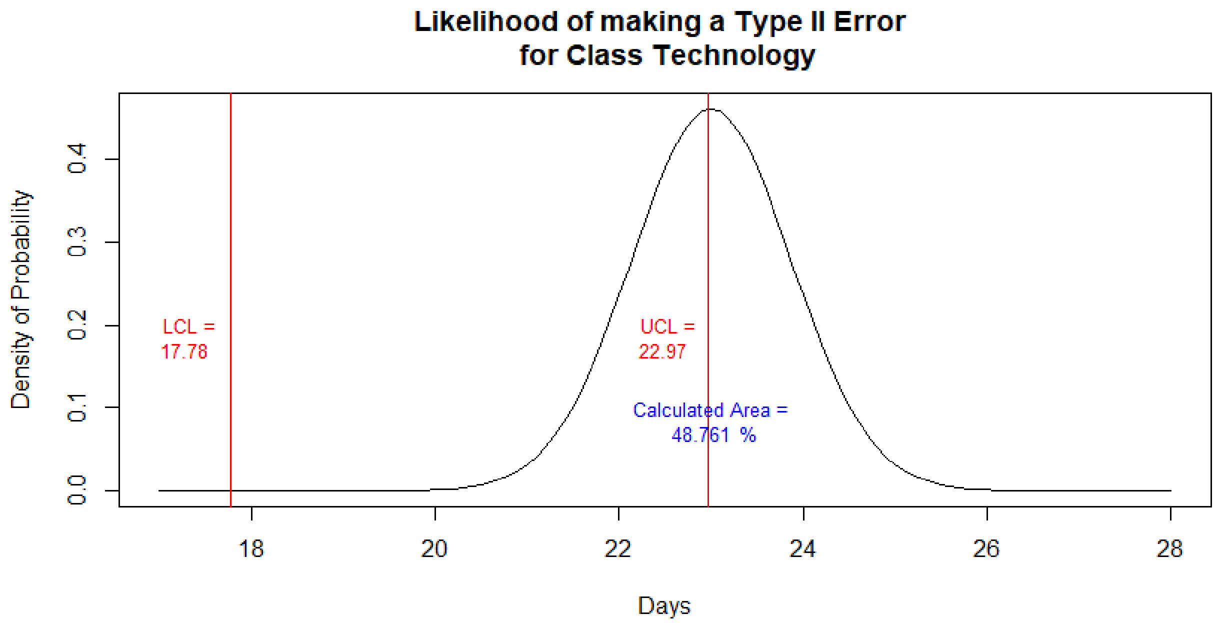


Figure 27: Visual Interpretation of a Type II Error

6. DOE and MANOVA

The focus of this part of the project is to simultaneously test multiple response variables and receive it in one output form. Thus, comparison between variables can be done easily.

This comparison will specifically be done between the categorical features ["Class" and "Why.Bought"] to the continuous features which were discovered to be relevant. Features found to be relevant by observing Part 2, Part 3, and Part 4 of the project were the following: "AGE", "Price", and "Delivery.time".

A MANOVA test was performed to compare these features with one another.

From tests conducted it can easily be seen that the default test has been used, Pillai.

6.1. Categorical Feature 1: Class

The figure [Figure 28] below reflects the summary of the categorical feature Class vs the different continuous features.

```
                Df Pillai approx F num Df den Df                Pr(>F)
(Intercept)      1 0.9701  1946689      3 179969 < 0.00000000000000022 ***
Class            6 1.7577    42440     18 539913 < 0.00000000000000022 ***
Residuals   179971
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 28: Summary on MANOVA test for the categorical feature Class based on all continuous variables

In the following three figures [Figure 29, Figure 30, Figure 31], the results obtained from the MANOVA test is shown in more detail. The relationships between the following features can be observed: "AGE" and "Class", "Price" and "Class", and lastly "Delivery.time" and "Class". All these relationships are presented separately.

```
 Response AGE :
                Df   Sum Sq Mean Sq F value                Pr(>F)    
Class            6  8422401 1403733    3805 < 0.00000000000000022 ***
Residuals   179971 66394669     369                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 29: MANOVA for Age vs Class

```
 Response Price :
                Df         Sum Sq       Mean Sq F value                Pr(>F)    
Class            6 57168427663229 9528071277205   80258 < 0.00000000000000022 ***
Residuals   179971 21365723828547     118717592                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 30: MANOVA for Price vs Class

```
 Response Delivery.time :
                Df   Sum Sq Mean Sq F value                Pr(>F)    
Class            6 33458565 5576427  629429 < 0.00000000000000022 ***
Residuals   179971  1594452       9                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 31: MANOVA for Delivery Time vs Class

As can be seen from the three figures above, all the p-values (last column on each figure) of the comparisons are smaller than 0.05. This is an indication that the differences between the means of the categorical variables being compared with the continuous variables are statistically significantly different from one another.

6.2. Why Bought

The figure [Figure 32] below reflects the summary of the categorical feature Why.Bought vs the different continuous features.

```
                Df  Pillai approx F num Df den Df                Pr(>F)    
(Intercept)      1 0.90735   587480      3 179970 < 0.00000000000000022 ***
Why.Bought       5 0.04414      538     15 539916 < 0.00000000000000022 ***
Residuals   179972                                                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 32: Summary on MANOVA test for the categorical feature Why.Bought based on all continuous variables

In the following three figures [Figure 33, Figure 34, Figure 35], the results obtained from the MANOVA test is shown in more detail. The relationship between the following features can be observed: “AGE” and “Why.Bought”, “Price” and “Why.Bought”, and lastly “Delivery.time” and “Why.Bought”. All these relationships are presented separately.

```
 Response AGE :
                  Df   Sum Sq Mean Sq F value                Pr(>F)    
Why.Bought         5   106542 21308.4   51.33 < 0.0000000000000022 ***
Residuals   179972 74710528   415.1                                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 33: MANOVA for Age vs Why.Bought

```
 Response Price :
                Df         Sum Sq      Mean Sq F value                Pr(>F)    
Why.Bought       5  1574211183514 314842236703  736.26 < 0.0000000000000022 ***
Residuals   179972 76959940308351    427621743                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 34: MANOVA for Price vs Why.Bought

```
 Response Delivery.time :
                  Df   Sum Sq Mean Sq F value                Pr(>F)    
Why.Bought         5   783320  156664  822.74 < 0.0000000000000022 ***
Residuals   179972 34269697     190                                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 35: MANOVA for Delivery.time vs Why.Bought

As has been seen in the previous section, all the p-values of the comparisons are smaller than 0.05. Again, this is an indication that the differences between the means of the categorical variables being compared with the continuous variables are statistically significantly different from one another.

7. Reliability of Service and Products

7.1. Problem 6, Chapter 7

The next section focus on Taguchi Loss Function method addressing problems from the textbook.

Problem 6: A blueprint specification for the thickness of a refrigerator part at Cool Food, Inc. is 0.06 +- 0.04 centimetres (cm). It costs $45 to scrap a part that is outside the specifications. Determine the Taguchi loss function for this situation.

From the problem statement it can easily be stated that T = 0.06 cm and L = $ 45. Making use of this given values the Taguchi Loss Function can be calculated by deducing the value of k.

Taguchi Loss Function is: $L(x) = k * (x - T)^2$

Where $\$45 = k * (0.04)^2 \quad \therefore k = 28\,125$

$L(x) = 28\,125 * (x - 0.06)^2$

From the figure (Figure 36) below the Taguchi Loss Function is plotted for the calculated values above. The x-variable was passed as a sequence from 0.0025 to 0.12 incrementing it by 0.005. The figure also represents the boundaries (in red) of the particular function set-up as specified by specifications.

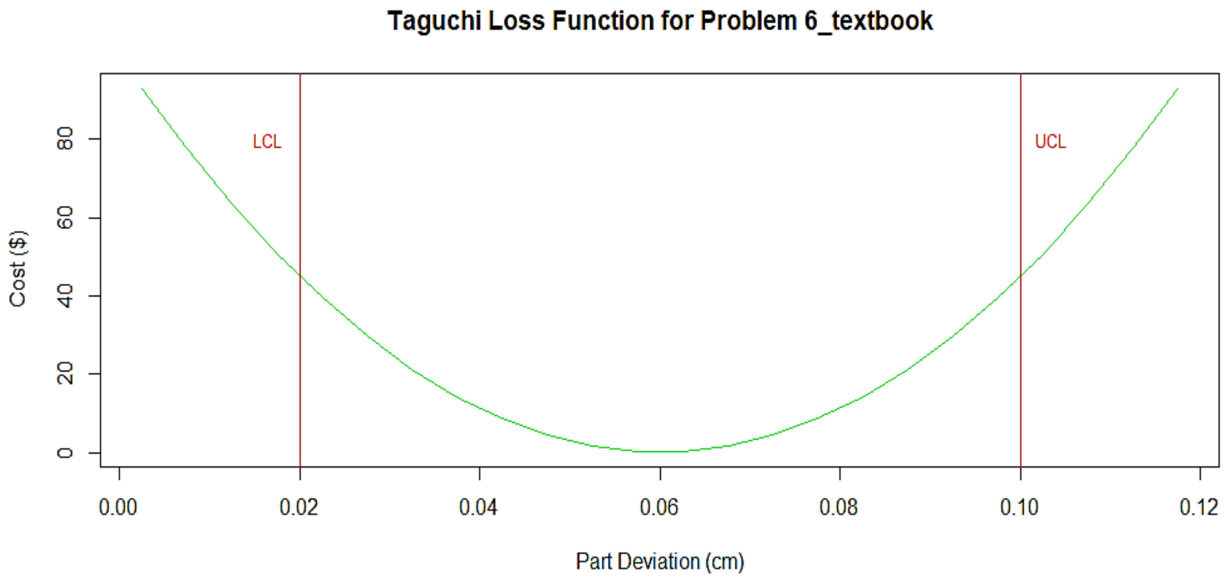


Figure 36: Taguchi Loss Function for Problem 6 (textbook)

7.1. Problem 7.a, Chapter 7

Problem 7: A team was formed to study the refrigerator part at Cool Food, Inc. described in Problem 6. While continuing to work to find the root cause of scrap, they found a way to reduce the scrap cost to $35 per part.

a. Determine the Taguchi loss function for this situation.

Same process will be followed as in the previous question. The T value is still T = 0.06 but with a scrap cost of the part of $35.

Taguchi Loss Function is: $L(x) = k * (x - T)^2$

Where $\$35 = k * (0.04)^2 \quad \therefore k = 21\,875$

$L(x) = 21\,875 * (x - 0.06)^2$

From the figure (Figure 37) below the Taguchi Loss Function is plotted for the calculated values above. The x-variable was again passed as a sequence from 0.0025 to 0.12 incrementing it by 0.005. The figure also represents the boundaries (in red) of the particular function set-up as specified by specifications.

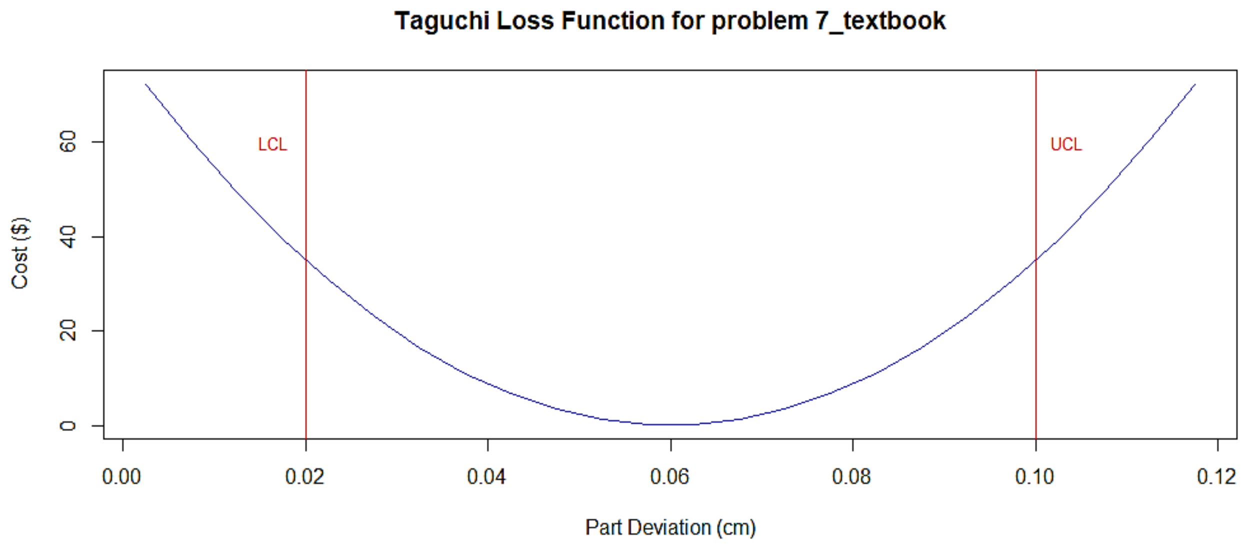


Figure 37: Taguchi Loss Function for Problem 7 (textbook)

7.1. Problem 7.b, Chapter 7

b. Deviation reduction to a value of 0.027 cm

For the following problem the only difference between Problem 7.a and Problem 7.b is the deviation is reduced to 0.027 cm instead of 0.04cm.

Thus, the Taguchi Loss Function is calculated as follows:

$$L(0.027) = 21\,875 * (0.027)^2 = \$\,15.946875/part$$

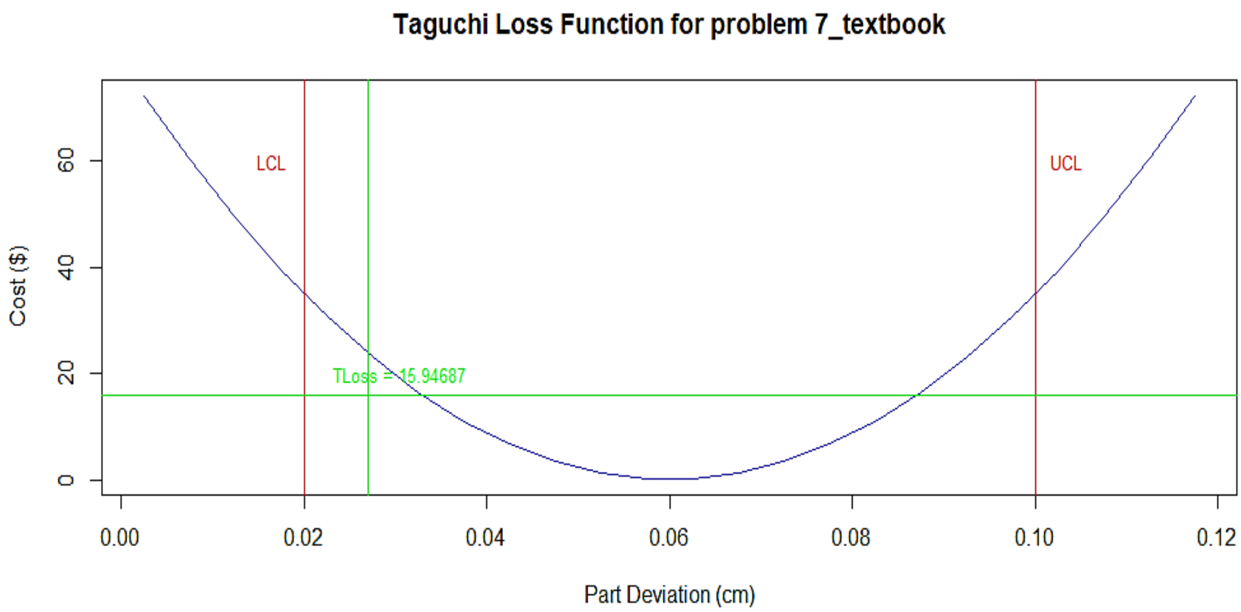


Figure 38: Taguchi Loss Function for Problem 7 - Reduced Deviation (textbook)

7.2. Problem 27, Chapter 7

Reliability for Problem 27 from Chapter 7.

Problem 27: Magnaplex, Inc. has a complex manufacturing process, with three operations that are performed in series. Because of the nature of the process, machines frequently fall out of adjustment and must be repaired. To keep the system going, two identical machines are used at each stage; thus, if one fails, the other can be used while the first is repaired (see accompanying figure).

The reliabilities of the machines are as follows:

Machine Reliability

P(A) = 0.85

P(B) = 0.92

P(C) = 0.90

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

$$P(A) * P(B) * P(C) = 0.85 * 0.92 * 0.90 = 70.38\ \%$$

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

A parallel method of application will be applied:

$$Raa * Rbb * Rcc = (1 - (1 - Ra)^2) * (1 - \ (1 - Rb)^2) * (1 - \ (1 - Rc)^2)$$

$$Raa * Rbb * Rcc = (1 - (1 - 0.85)^2) * (1 - \ (1 - 0.92)^2) * (1 - \ (1 - 0.90)^2)$$

$$Raa * Rbb * Rcc = \ 0.9615316 \quad \therefore 96.15316\ \%$$

7.3. Part 1 and Part 2

For the delivery process, there are 21 delivery vehicles available, of which 19 is required to be operating at any time to give reliable service. During the past 1560 days, the number of days that there was only 20 vehicles available was 190 days, only 19 vehicles available was 22 days, only 18 vehicles available was 3 days and 17 vehicles available only once. There are also 21 drivers, who each work an 8-hour shift per day. During the past 1560 days, the number of days that there were only 20 drivers available was 95 days, only 19 drivers available was 6 days and only 18 drivers available, once only.

Part 1: 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 22, how many days per year we should expect reliable delivery times?

Take note: When increasing the number of vehicles to 22, we must assume the probability of failure per vehicle stays constant.

From the table (Table 9) below the different number of days of reliability can easily be observed. For Part 2's increase in vehicle numbers from 21 to 22 there was not a big increase in reliability observed, the actual increase has been seen as very small 0.1108.

Table 9: Reliability of Vehicles and Drivers

	21 Vehicles	21 Drivers	Overall, 21 Drivers 21 Vehicles	22 Vehicles	Overall, 22 Drivers, 22 Vehicles
Availability of all	314.46252	341.06583	NA	312.23875	NA
Availability with less 1	314.46252	23.16911	NA	49.92299	NA
Availability with less 2	47.03181	0.74948	NA	3.65852	NA
Availability with less 3	NA	0.01535	NA	0.17371	NA
Availability with less 4	NA	0.00022	NA	0.00588	NA
Reliability Total Days	364.84395	364.999998	364.8439	365.11076	364.9547

8. Conclusion

From an analysis perspective it was noticed that certain features provided as part of the data set was not relevant and showed no correlation or any kind of relationship to another feature or between any other feature. This can be observed from the Correlation Table, SPLOM graph, Correlation Plot, and the Heat Map to just mention a few. However, the features ID, Year, Month, and Day have been used to sort the data set, from the oldest to the newest data collected. These features have been seen as redundant after this action has been implemented. For the rest of the analysis (thus after the sorting of the data) these features have not been used.

From the X-Chart and S-Chart results some outliers were observed where in certain features there was an extreme number of outliers observed. From Table 6 and Table 7 the exact number of samples seen as out of control for each feature within the X-Chart and the S-Chart can be found. This gives an indication that the delivery time has not been very constant from the past to the recent years. Certain classes reflected different kind of trends; upward, downward, and continuous trends have been identified to be present. It has been recommended that these classes need more investigation to ensure good quality, cost, and productivity in the long term.

Identifying an optimal delivery time for the class feature "Technology" was obtained by reducing the penalty costs which led to an increase in profit. If an item delivered took more than 26 hours a penalty cost was incurred. The new optimal average delivery time was found to be 16 hours which is 4 hours less than the original average delivery time of 20 hours.

The results obtained from the MANOVA tests by comparing Age, Price, and Delivery Time to Class and also Why Bought features concluded by comparing the differences obtained between the means that these features are statistically significantly different from one another.

The reliability tests conducted for the drivers and the vehicles showed that the difference between 21 vehicles and 21 drivers compared to 22 vehicles and 22 drivers were very small. Thus, implementing 22 vehicles and 22 drivers compared to only implementing 21 vehicles and 21 drivers will not necessarily bring a big improvement. A more in-depth cost analysis will be recommended to finalize the choice between implementing 21 vehicles and 21 drivers compared to 22 vehicles and 22 drivers. The reliability for both scenarios was around 365 days.

9. References

1. Anon., 2022. *Spearman's rank correlation coefficient - Wikipedia.* [Online]
Available at: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient
[Accessed 13 October 2022].

2. Anon., n.d. *Statistical tools for high-throughput data analysis.* [Online]
Available at: http://www.sthda.com/english/wiki/ggplot2-colors-how-to-change-colors-automatically-and-manually
[Accessed 17 October 2022].

3. Anon., n.d. *Taguchi's Quality Loss Function.* [Online]
Available at: https://www.brainkart.com/article/Taguchi-s-Quality-Loss-Function_5287/#:~:text=The%20quadratic%20loss%20function%20is,quality%20deviates%20from%20the%20target.&text=k%20%3D%20Quality%20loss%20coefficient.
[Accessed 27 September 2022].

4. Evans, J. R. & Lindsay, W. M., 2017. *Managing for Quality Performance Excellence.* 11 ed. United States of America: Cengage.

5. J. Puranen, G. P., n.d. *Finding order in randomness: correlation.* [Online]
Available at: http://nagysandor.eu/AsimovTeka/correlation_en/index.html
[Accessed 14 October 2022].

6. Nickolas, S., 2021. *Investopedia: What Do Correlation Coefficients Positive, Negative, and Zero Mean?.* [Online]
Available at: https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp#:~:text=Correlation%20coefficients%20are%20indicators%20of,zero%20signifies%20a%20negative%20relationship.
[Accessed 17 October 2022].

7. Schalkwyk, T. G. v., 2022. Statistical Methods in Auality Assurance. 22 August, p. 28.