ECSA ATTRIBTES PROJECT

QUALITY ASSURANCE - 344

Jacques Carl Huskisson 10/21/22 20962452@sun.ac.za

Table Of Contents:

Contents

List of figures:

List of tables:

Introduction

This report documents the transformation, exploration and statistical analysis of data provided in the file called "SalesTable2022.csv." The document comprises 180 000 entries consisting of ten features and an index column. This report aims to provide the business with meaningful insights and interpretations of identified trends and relationships within the dataset. The order of operations is as follows: the data is to be cleaned so that trends and relationships can be abstracted with the use of descriptive statistics. Statistical Process Control (SCP) methodology is then applied to the delivery times process and a more optimised average delivery time can be calculated. Finally, service and product delivery times are investigated and reported on.

Part 1: Data Wrangling

Figure A: Data Cleaning

```
d <- read.csv("C:/Users/user-pc/Documents/salesTable2022.csv")

dclean <- d[d$Price>=0,]
dclean <- dclean[complete.cases(dclean),]

ddirty <- d[rowSums(is.na(d)) > 0 | d$Price <=0,]

ddirty$Primary_Key <- 1:nrow(ddirty)
dclean$Primary_Key <- 1:nrow(dclean)

ddirty<- ddirty %>% select(Primary_Key, everything()) #Move key col to front

dclean <- dclean %>% select(Primary_Key, everything())

write.csv(ddirty,"C:/Users/user-pc/Documents/missing.csv", row.names = FALSE)
write.csv(dclean,"C:/Users/user-pc/Documents/valid.csv", row.names = FALSE)
```

```
df <- read.csv("C:/Users/user-pc/Documents/valid.csv")
df<- df[base::order(df$Year, df$Month, df$Day, df$X),]
str(df)
head(df)
summary(df)
```

The valid data is ordered in terms of year, month, day and key.

The data wrangling section of this project involves the importation and cleaning of a csv file that contains the firm's sales data. It quickly becomes clear that the data contains missing values and some observations with negative p rice indexes. It Is important to ensure that all the data at hand is present and valid before heading into the next phase (Figure A.) The data is therefore cleaned* and the following information is gained:

Valid records = 179978

Invalid Records = 22

NAs = 17

Negative Price Values = 5

The next phase involving descriptive statistics involves various ways of representing the data as well as some process capability calculations to better help us get familiar with the data. R allows us to extract and display many relationships between the features that can be found within the sales datasheet. This phase of the report aims to identify and analyse the most useful of these relationships.

Figure B provides us with a data quality summary of the valid data. Figures C and D provide us with data types and feature cardinality.

Figure B: Data Quality Summary

```
1st Qu.:   482.31   1st Qu.:2022   1st Qu.: 4.000
 Median : 89990   Median : 90005   Median :55081   Median : 53.00   Mode  :character
Median :  2259.63   Median :2025   Median : 7.000
 Mean   : 89990   Mean   : 90003   Mean   :55235   Mean   : 54.57
Mean   : 12294.10   Mean   :2025   Mean   : 6.521
 3rd Qu.:134984   3rd Qu.:135000   3rd Qu.:77637   3rd Qu.: 70.00
3rd Qu.: 15270.97   3rd Qu.:2027   3rd Qu.:10.000
 Max.   :179978   Max.   :180000   Max.   :99992   Max.   :108.00
Max.   :116618.97   Max.   :2029   Max.   :12.000
      Day          Delivery.time    Why.Bought
 Min.   : 1.00   Min.   : 0.5   Length:179978
 1st Qu.: 8.00   1st Qu.: 3.0   Class :character
 Median :16.00   Median :10.0   Mode  :character
 Mean   :15.54   Mean   :14.5
 3rd Qu.:23.00   3rd Qu.:18.5
 Max.   :30.00   Max.   :75.0
── Data Summary ────────────────────────
                           Values
Name                       Piped data
Number of rows             179978
Number of columns          11

_______________________
Column type frequency:
  character                2
  numeric                  9

________________________
Group variables            None
```

Figure C: Data Types and Cardinality

A tibble: 11 x 6

variables <chr>	types <chr>	missing_count <int>	missing_percent <dbl>	unique_count <int>	unique_rate <dbl>
Primary_Key	integer	0	0	179978	1.000000e+00
X	integer	0	0	179978	1.000000e+00
ID	integer	0	0	15000	8.334352e-02
AGE	integer	0	0	91	5.056174e-04
Class	character	0	0	7	3.889364e-05
Price	numeric	0	0	78832	4.380091e-01
Year	integer	0	0	9	5.000611e-05
Month	integer	0	0	12	6.667482e-05
Day	integer	0	0	30	1.666870e-04
Delivery.time	numeric	0	0	148	8.223227e-04
Why.Bought	character	0	0	6	3.333741e-05

Figure D: Categorical Data and Cardinality

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	min <int>	max <int>	empty <int>	n_unique <int>	whitespace <int>
1	Class	0	1	4	10	0	7	0
2	Why.Bought	0	1	4	11	0	6	0

2 rows

Part 2: Descriptive Statistics

The first relationship that is noteworthy to the firm is probably the most intuitive of the bunch and that is the cumulative relationship of class (SKU group) to purchases made (Figure E.) It is evident that gifts and technology have the highest demands and luxury items have the lowest demand. Additional demand related insights can be gained from Figures F and H. Figure F shows the total earnings per class. Even though luxury items have the least number of sales they are still responsible for a significant amount of total revenue – trumped only by the technology class. This "contradiction" can be explained by having a look at the average prices of items in these classes. The boxplot of sales price per category (Figure H) indicates that technology and luxury are the most valuable SKU groups. These classes are the firm's majority income.

Figure E: Cumulative sales graph that is grouped by SKU

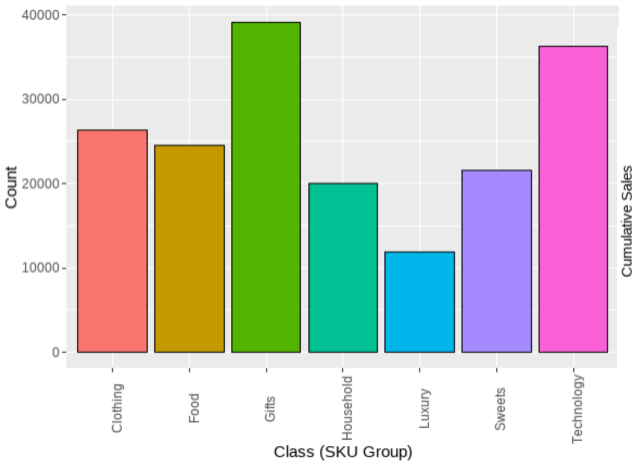
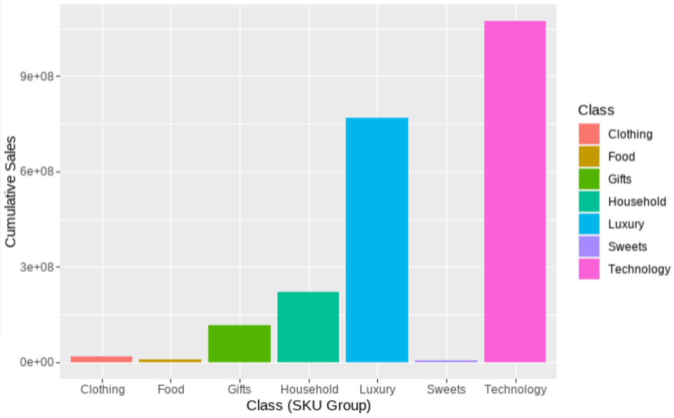


Figure F: Cumulative sales graph that is grouped by SKU groups (Total sales price.)



Customer identification should be an iterative process as is integral to demand forecasting and production planning. Figure G displays the effectiveness of different communication channels that the firm are using to entice potential customers into action. It is noted that 'recommended' is cited as the most popular customer motivation. This translates to word of mouth and emphasises the importance of product quality and customer experience. It follows that browsing and website advertisement are next most effective at winning over customers. The avenues should be further explored by the advertising department. Spam is least effective at gaining user attention. It is the writer's opinion that the firm should consider phasing out this strategy as it could damage customer relationships.

Figure G: Cumulative sales graph that is grouped by customer motivation.

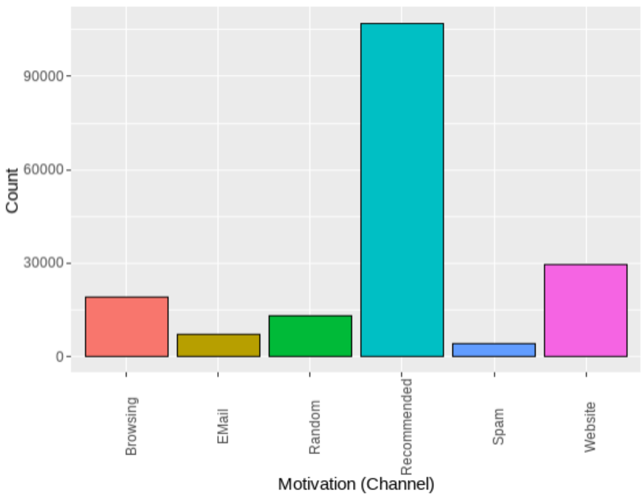
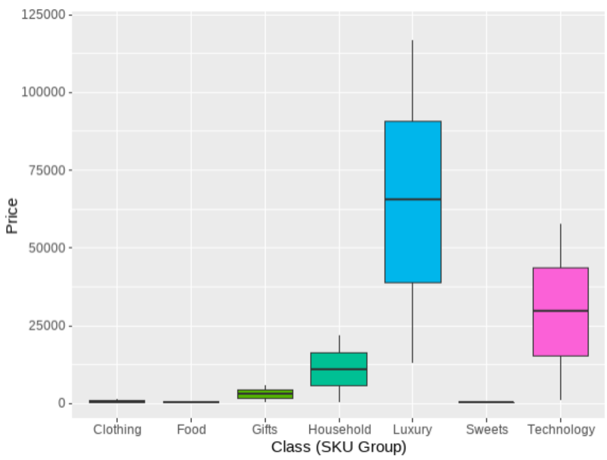
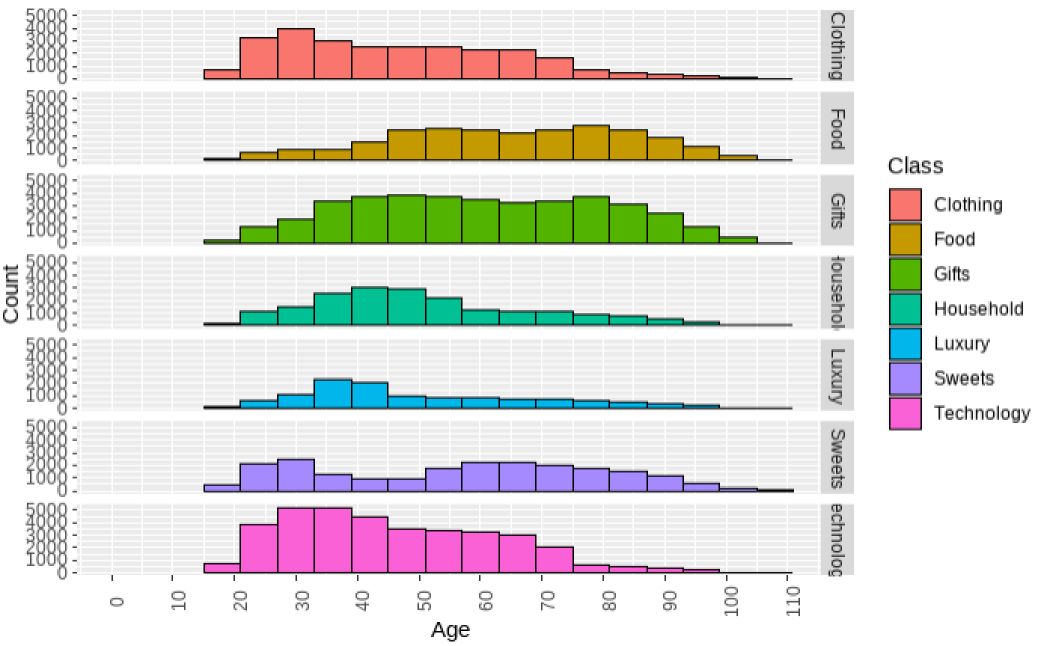


Figure H: Cumulative sales graph that is grouped by SKU groups (Total sales price.)



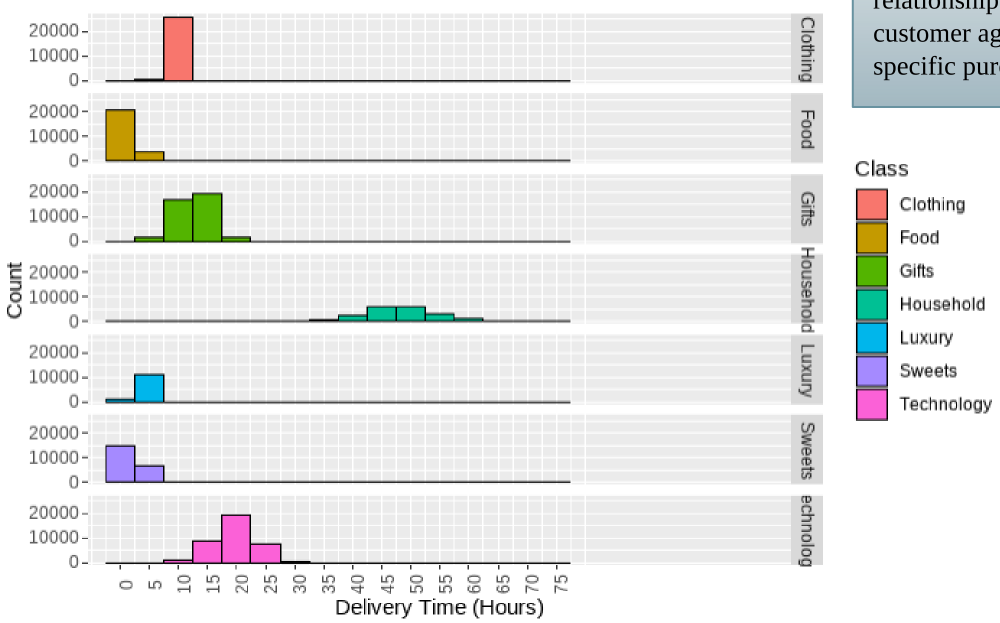
More useful information pertaining to customer identification can be abstracted from the relationship between customer age and the amount and type of products that they buy – such a relationship is on display in Figure I. From this information the firm will be able to effectively target customers. For example, technology and luxury item advertisements that directly reach the intended buyer could significantly increase sales rates.

Figure I: Histogram of relationship between customer age and class specific purchases.



Lastly Figure J provides a glimpse into the different delivery times of the various SKU groups. It follows that household items have the longest delivery time – this is probably because of item size and transportation method. It is noted that household, technology and gift item deliveries have larger distributions than their counterparts. Household deliveries range from 32 to 75 hours. This should be investigated by the firm. Correlation between delivery time and distance should be taken into account.

Figure I: Histogram of relationship between customer age and class specific purchases.



The lower specification limit, LSL, represents the lowest limit that a measurement can reach and still be considered acceptable by a customer. For the feature, delivery time, an LSL of zero makes sense since delivery time will not get minimized beyond the USL and thus a window is provided in which the product is to be delivered. See figures J and K for process capability calculations (JR Evans and WM Lindsay, 2013)

It is calculated that Cpk is equal to 0.38. This is alarming because a Cpk value of less than one is usually indicative of an incapable process. The business should aim for a Cpk of at least 1.3. Such a case would imply that the process is cantered and in bounds. Figure L provides an example visualisation of the relationship between Cp/Cpk values and the distribution of a process.

Figure J: Process Capability values

Description: df [4 x 2]

Cp.Calculations <chr>	Process.Capabilities <dbl>
cp	1.1422068
cpu	0.3796933
cpl	1.9047203
cpk	0.3796933

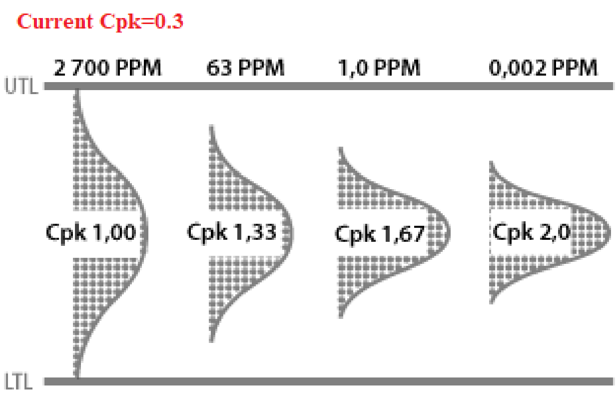
4 rows

Figure K: Process capability calculations

```
USL<-24
LSL<-0
dftech <- df[df$Class=="Technology",]
sdtech<- sd(dftech$Delivery.time)
sdmean<- mean(dftech$Delivery.time)


cp<-(USL-LSL)/(6*sdtech)
cpu<-(USL-sdmean)/(3*sdtech)
cpl<-(sdmean-LSL)/(3*sdtech)
cpk<-min(cpl,cpu)
```

Figure L: Example of correlation between distribution and Cpk values.



Part 3: Statistical Process Control

SPC is defined as the use of statistical techniques to control process or production method (Ipek and Ankara, 1999.) Here we will apply the SPC methodology to the Delivery Times of the dataset. This will aid in the discovery of previously unseen behaviours or problems within the process.

Firstly, the X- and S-Charts for the delivery process times are initialised with 30 samples of 15 sales each. These initial 30 samples are used to calculate the centrelines, outer control limits and the 1- and 2-sigma-control limits for the respective classes (SKU Groups.) These values are entered into the respective X- and S-Chart tables. See Table 1 and 2.

Table1: X-Chart

Class		UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	XCCL_Cloth	9.404934	9.259956	9.114978	8.970000	8.825022	8.680044	8.535066
Food	XCCL_Food	2.709458	2.636305	2.563153	2.490000	2.416847	2.343695	2.270542
Gifts	XCCL_Gifts	9.488565	9.112747	8.736929	8.361111	7.985293	7.609475	7.233658
Household	XCCL_House	50.248328	49.019626	47.790924	46.562222	45.333520	44.104818	42.876117
Luxury	XCCL_Lux	5.493965	5.241162	4.988359	4.735556	4.482752	4.229949	3.977146
Sweets	XCCL_Sweets	2.897042	2.757287	2.617532	2.477778	2.338023	2.198269	2.058514
Technology	XCCL_Tech	22.974616	22.107892	21.241168	20.374444	19.507721	18.640997	17.774273

Table2: S-Chart

Class		UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	SCCL_Cloth	0.8665596	0.7614552	0.6563509	0.5512465	0.4461422	0.3410379	0.2359335
Food	SCCL_Food	0.4372466	0.3842133	0.3311800	0.2781467	0.2251134	0.1720801	0.1190468
Gifts	SCCL_Gifts	2.2463333	1.9738773	1.7014213	1.4289652	1.1565092	0.8840532	0.6115971
Household	SCCL_House	7.3441801	6.4534101	5.5626402	4.6718703	3.7811003	2.8903304	1.9995605
Luxury	SCCL_Lux	1.5110518	1.3277775	1.1445032	0.9612289	0.7779546	0.5946803	0.4114060
Sweets	SCCL_Sweets	0.8353391	0.7340215	0.6327039	0.5313862	0.4300686	0.3287509	0.2274333
Technology	SCCL_Tech	5.1805697	4.5522224	3.9238751	3.2955278	2.6671805	2.0388332	1.4104859

X- and S- charts are visualised with samples comprising the entire dataset and control limits calculated from the first 30 samples. This implies that the green line visible on the plots is representative of the mean of the first 30 samples of each respective class.

Figure M: Average Delivery Hours SPC: **Clothing**

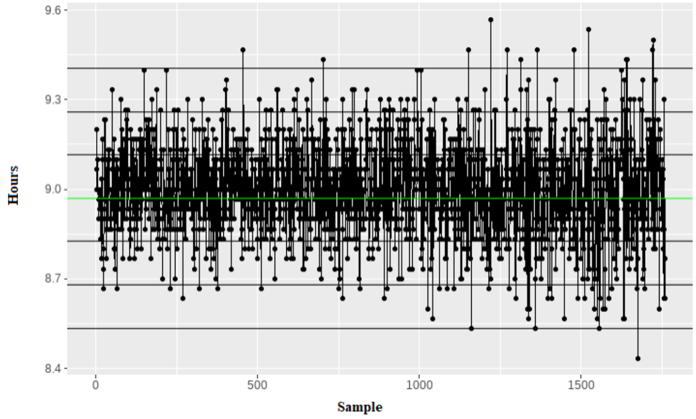


Figure N: Average Delivery Hours SPC: **Food**

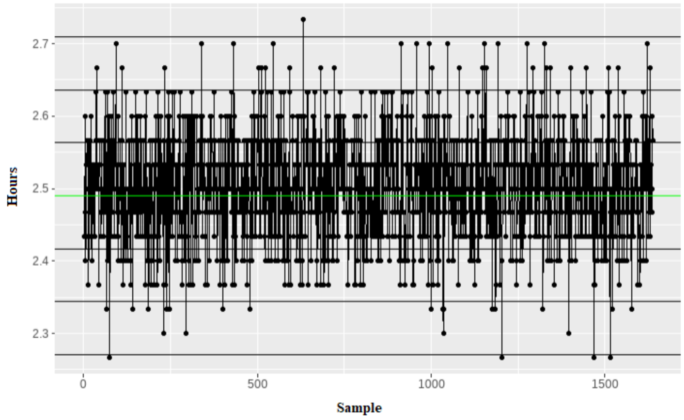


Figure O: Average Delivery Hours SPC: **Gifts**

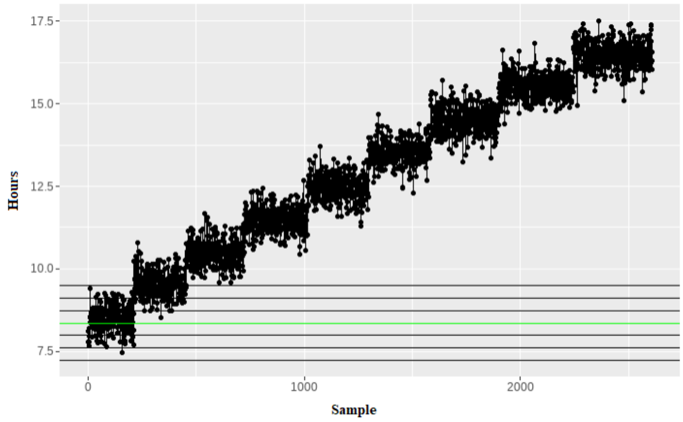


Figure P: Average Delivery Hours SPC: **Household**

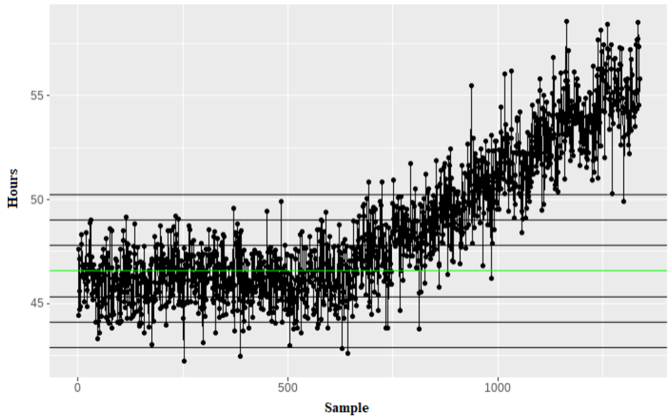


Figure Q: Average Delivery Hours SPC: **Luxury**

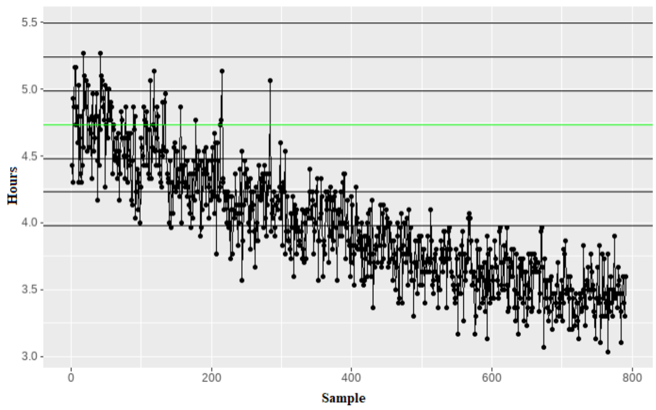
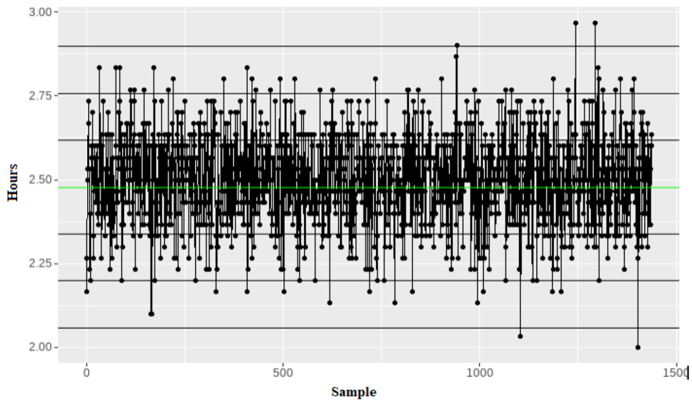
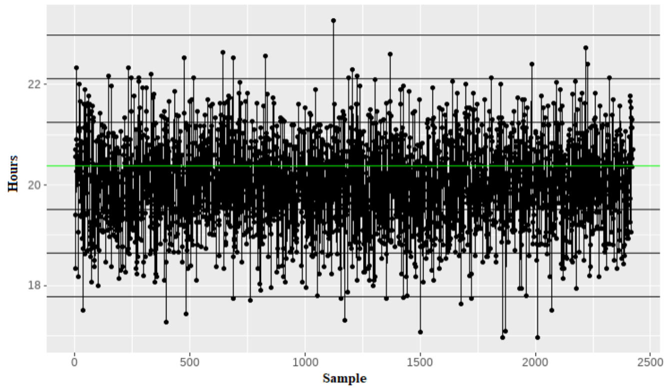


Figure R: Average Delivery Hours SPC: **Sweets**



S-Chart visualisations included in appendix A.

Figure S: Average Delivery Hours SPC: **Technology**



It can be deducted that three classes have averages that do not conform to the limits that were calculated with the first 30 samples.

Gifts:

The shape of the graph indicated that the delivery time is increasing in a step-like manner. This could be the result of increasing pressure on logistics. Increased demand could have resulted in a situation where there are too many orders to have them all delivered within the initial limits. This requires the business's immediate attention.

Household:

As more samples are taken into account, the number of outliers starts to increase and then once the +/- 700 threshold is reached, a positive linear trend becomes evident in the data. This could also be attributed to, too much stress on the current delivery process to be able to deliver items on time.

Luxury:

There is a negative linear trend evident in the data after the initial 30 samples were taken. This could be indicative of the business prioritising luxury item deliveries. This could ensure that buyers of luxury items remain satisfied but it is recommended that the business investigates the correlation between the decrease in delivery times for luxury items and the increase in delivery times for gifts and household items. A more optimal strategy could be possible.
Figure F, indicates that technology sales also contribute heavily to total earnings and therefore the business could experiment with prioritising technology delivery times in a similar manner.

Part 4.1A: $\bar{X}$ Samples Outside Outer Control Limits

Table 3: $\bar{X}$ Samples Outside Outer Control Limits

		Clothing	Food	Gifts	Household	Luxury	Sweets	Technology
First	1	455	75	213	252	142	942	37
Second	2	702	633	216	387	171	1104	398
Third	3	1152	1203	218	629	184	1243	483
Third Last	4	1677	1203	2607	1335	789	1243	1872
Second Last	5	1723	1467	2608	1336	790	1294	2009
Last	6	1724	1515	2609	1337	791	1403	2071
Total Identified	Num Outside LCL/UCL	17	5	2290	400	434	5	17

This section documents the identification of sample means outside of the outer control limits. The first three, last three and total number of out of bounds sample means are listed, per item category, in Table 3. From previous visualisations of the X-chart it is noted that the Class specific graphs do correlate with the numbers in the above table (Gifts, Household and Luxury have a higher number of outliers.)

Part 4.1B: Most Consecutive Samples of Sample Standard Deviations Between -0.3 and +0.4 σ-Control Limits

Table 4: Most Consecutive samples of s-bar between σ-Control Limits

	Clothing	Food	Gifts	Household	Luxury	Sweets	Technology
Max. Consecutives	4	7	5	3	4	4	7
Ending Sample #	1013	952	1651	908	63	1292	2419

The number and index of any consecutive samples of 's-bar' are obtainable through the implementation of for loops after the calculating the -0.3 and +0.4 σ-control limits. Table 4 lists the highest number of consecutive samples of 's-bar' for each class as well as the ending sample number/last sample number within that range. Although we obtained these values through the implementation of loops, a quick command into the console allows for confirmation that the applied methodology is sound. See Figures T and U.

Figure T: Sweets samples of s-bar within σ-Control Limits

```
> keptSamplingS$Sample
  [1]   10   12   13   17   24   25   29   32   36   38   43   44   48   53   54   65   66   71   81   91   92   93   94   96  100  105  106  111  114
 [30]  115  118  121  122  123  126  129  131  135  139  144  146  148  154  157  160  162  164  165  169  172  176  177  179  183  184  186  187  188
 [59]  189  194  201  207  209  214  224  230  232  235  238  239  244  248  250  255  256  259  260  262  269  284  293  295  296  298  300  304  308
 [88]  309  312  313  315  316  321  330  332  335  340  342  343  346  350  355  358  370  371  372  374  375  387  394  397  401  413  417  419  422
[117]  428  448  451  456  459  465  472  474  475  477  485  492  496  497  503  515  517  520  522  528  529  530  534  535  539  541  543  546  547
[146]  549  552  559  561  567  568  569  571  576  584  585  587  590  595  604  607  611  614  616  630  633  638  640  644  646  648  650  656  661
[175]  663  668  678  679  681  683  685  686  688  691  692  693  697  707  711  715  716  718  719  724  726  731  735  741  743  748  757  759  760
[204]  761  763  765  767  768  774  780  783  784  793  805  809  810  812  813  816  818  822  826  827  829  830  833  844  854  858  862  867  868
[233]  877  881  884  891  896  902  906  908  918  933  936  939  945  946  947  950  951  953  957  959  961  963  965  968  969  970  971  977  986
[262]  991  995  998 1004 1005 1006 1014 1017 1019 1030 1044 1048 1058 1065 1075 1077 1079 1082 1083 1084 1088 1101 1103 1107 1108 1119 1121 1124 1125
[291] 1128 1140 1146 1149 1150 1154 1158 1171 1174 1178 1180 1181 1188 1190 1196 1202 1204 1205 1208 1209 1212 1216 1217 1218 1230 1240 1241 1242 1255
[320] 1265 1267 1273 1276 1289 1290 1291 1292 1295 1296 1301 1302 1313 1315 1317 1319 1321 1328 1331 1356 1357 1358 1366 1370 1371 1378 1379 1382 1386
[349] 1394 1395 1396 1403 1409 1411 1412 1413 1419 1420 1424 1434 1437
```

The figure below the last section of a vector called *keptSamplingT$Sample* which contains all the samples of 's-bar' for technology delivery times that fall within the specified control limits.

Figure U: Technology samples of s-bar within σ-Control Limits

```
[378] 1160 1161 1165 1166 1168 1177 1181 1182 1189 1190 1193 1199 1201 1203 1204 1207 1209 1212 1214 1215 1219 1221 1223 1225 1226 1236 1238 1242 1245
[407] 1252 1263 1264 1265 1271 1274 1283 1285 1290 1292 1297 1298 1301 1311 1312 1313 1319 1320 1322 1326 1327 1334 1335 1341 1344 1345 1351 1360 1361
[436] 1363 1365 1370 1372 1376 1377 1378 1380 1383 1385 1387 1389 1391 1394 1397 1398 1401 1407 1413 1423 1425 1441 1442 1443 1444 1448 1449 1453 1458
[465] 1460 1461 1468 1474 1478 1480 1488 1489 1491 1495 1500 1503 1507 1517 1518 1528 1539 1540 1543 1544 1550 1568 1570 1575 1577 1578 1583 1588 1589
[494] 1592 1598 1600 1607 1624 1625 1634 1637 1638 1642 1648 1655 1657 1659 1660 1663 1665 1669 1672 1674 1679 1682 1683 1689 1690 1695 1698 1702 1705
[523] 1708 1716 1719 1721 1724 1725 1726 1727 1730 1732 1740 1743 1748 1749 1754 1758 1764 1765 1769 1770 1776 1779 1780 1787 1788 1795 1796 1797 1800
[552] 1809 1811 1815 1822 1825 1829 1831 1837 1842 1845 1853 1859 1864 1865 1870 1872 1873 1881 1882 1886 1887 1892 1893 1896 1900 1902 1903 1906 1907
[581] 1909 1912 1915 1918 1921 1923 1924 1925 1931 1934 1937 1939 1945 1949 1952 1953 1955 1956 1963 1964 1965 1968 1970 1971 1975 1978 1981 1985 1992
[610] 1994 1995 1998 2001 2002 2005 2007 2008 2009 2011 2012 2013 2014 2018 2020 2022 2026 2028 2029 2031 2034 2041 2045 2046 2051 2053 2055 2058 2060
[639] 2064 2065 2068 2072 2073 2074 2076 2078 2079 2080 2081 2082 2084 2087 2089 2091 2094 2101 2103 2115 2119 2120 2121 2124 2128 2129 2133 2138 2141
[668] 2142 2158 2159 2161 2162 2166 2169 2171 2173 2175 2176 2177 2179 2185 2187 2193 2195 2198 2201 2209 2210 2211 2212 2216 2221 2226 2227 2228 2230
[697] 2231 2232 2234 2237 2241 2249 2250 2251 2252 2255 2257 2258 2259 2264 2266 2267 2273 2275 2277 2279 2281 2283 2284 2291 2293 2300 2304 2309 2311
[726] 2315 2317 2324 2325 2327 2330 2331 2332 2339 2340 2341 2343 2345 2346 2350 2351 2355 2361 2362 2363 2369 2372 2376 2378 2379 2394 2401 2402 2405
[755] 2407 2409 2413 2414 2415 2416 2417 2418 2419 2421 2422
> 
```

Part 4.2: Likelihood of making a Type I (Manufacturer's) Error for A and B

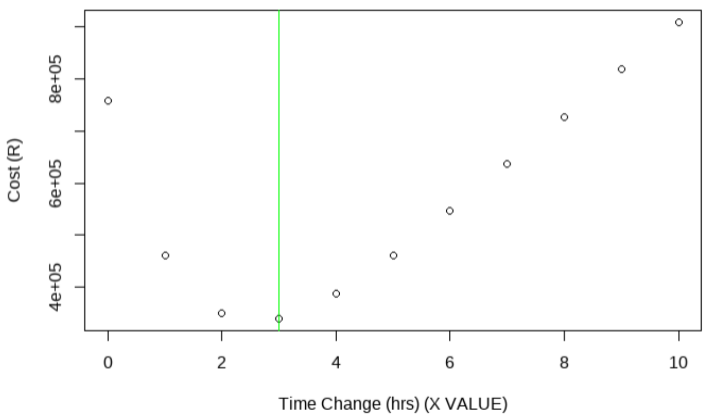
A type 1 error is a false positive conclusion or the probability of rejecting a true null hypothesis (Bhandari, 2022.)

The probability of making a type 1 error for 1 standard deviation is estimated: **P(typeI_A1) = pnorm(-1)*2 = 0.3173105**

The probability of making a type 1 error for 3 standard deviations is estimated: **P(typeI_A3) = pnorm(-3)*2 = 0.002699796**

The probability of making a type 1 error for B is estimated: **P(typeI_B) = pnorm(0.4) - pnorm(-0.3) = 0.2733332**

Part 4.3: Optimisation of Delivery Cost



The **X** Variable is indicative of the decision to decrease average delivery time. To find a brute forced solution to the below listed problem, it is necessary to keep track of the effect of various **X** values on the costs accumulated by the delivery process. Figure S provides a visual representation of the scenario. This problem is similar to obtaining Taguchi loss as this is also an example of finding an optimal x value. It is dissimilar in the sense that there is only and LCL in the form 26 hours.

Find optimal mean hours for delivery process given:

- **Lose** R329/item-late-hour in lost sales if technology items delivered slower than 26 hours.
- **Costs** R2.5/item/hour to reduce the average time by one hour.
- **Costs** less (-R2.5/item/hour) if the delivery time is increased.
- Assume the **current cost** for comparison purposes = 329* (item-late-hours).
- LOGIC: **New Delivery Time** <- Current Delivery Time – **X**
- Brute force solution with lowest cost **X**.

It is concluded that a value of **X** = 3 would result in the lowest cost for the process. This implies that the mean delivery time for the process should be centred at 16.01095 hours.

*20.01095 (Process Avg.) – 3 (**X**) = 16.01095 Hours*

Part 4.4: Likelihood of making a type II (Consumer's) Error for A in Class=Technology

A type 2 error for the delivery time process of the technology class implies that the business is under the impression that a delivery is being made on time but in reality, it is delivered late (False negative conclusion.) This could negatively impact customer relationships and at the same time provide the business with false confidence in the process capability. Type 2 errors are usually caused by the statistical power of a test being low. Thus, it follows that the probability of making a type 2 error can be reduced but not removed from hypothesis testing. This probability can be minimised by increasing the sampling size or increasing the significance level.

The probability of making a type 2 error for delivery times of class Technology is estimated:

sd = (XCCL[7,1] - XCCL[7,7])/6   ((UCL TECH - LCL TECH)/6)

P(type2) = pnorm(XCCL[7,1], mean = 23, sd=p)-pnorm(XCCL[7,7], mean = 23, sd=p)   = 0.4883177

Part 5: DOE & MANOVA

MANOVA or multivariate analysis of variance is an extension of regular ANOVA in the sense that the MANOVA test includes a minimum of two dependant variables to analyse differences between factors of the independent variable. To follow is a multivariate analysis of variance of the provided dataset. The target feature for this analysis year of purchase and the dependant variables are price and delivery time.

Take: H0: Group mean vectors for all groups do not differ significantly.

H1: At least one group mean vectors is different from the others.

alpha: 0.05

Assume: Dependant variables are normally distributed within groups.

Homogeneity of variances across the range of predictors.

Linearity between all pairs of dependent variables, covariates, and dependent variable-covariate pairs in each cell.

It is necessary for the sample size to be sufficiently large that there are more observations per group in the independent variable (Year) than a number of dependant variables. This implies that more than 2 observations per year of purchase (Radečić , 2022.)

Figures V and W provide a visualisation of the 250 samples that have been taken for this analysis.

Figure V: Price vs independent variable (Buyer motivation)

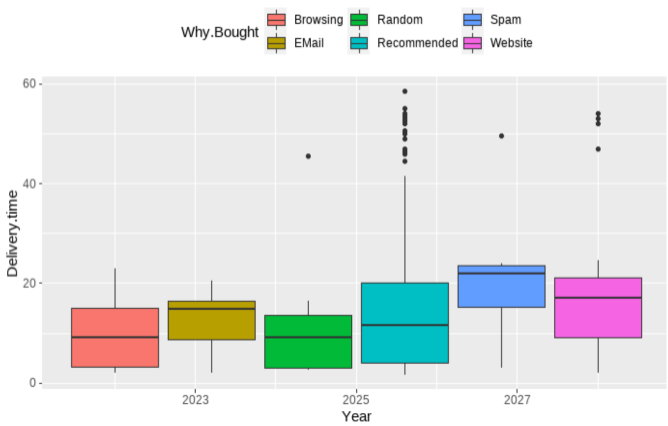
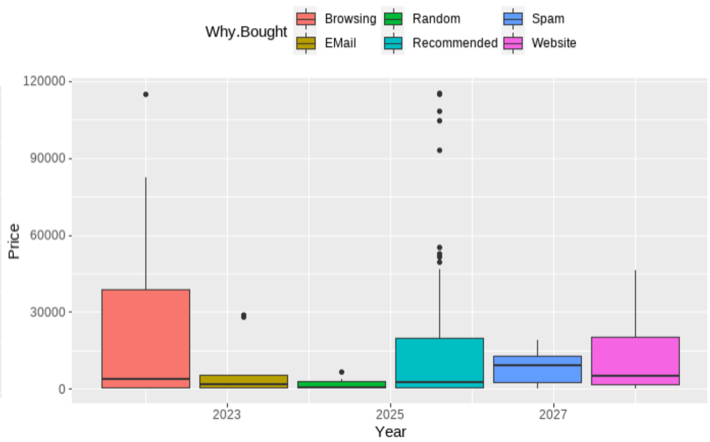


Figure W: Price vs independent variable (Buyer motivation)



From the above figures an estimate can be made that years 2025 and 2026 are more separated than the rest but we will have to confirm these suspicions by performing one-way MANOVA using the manova() function in r. Figure X provides its summary and figure Y provides its initialisation.

Figure X: MANOVA summary

```
> summary(manModel2)
           Df   Pillai approx F num Df den Df  Pr(>F)  
indvar2     1 0.028361   3.6048      2    247 0.02863 *
Residuals 248                                          
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure Y: MANOVA initialisation

```
mandSamp <- sample_n(df,250)
depvars2 <- cbind(mandSamp$Delivery.time,mandSamp$Price)
indvar2 <- mandSamp$Year
manModel2 <- manova(depvars2 ~ indvar2, data = mandSamp)
```

Pr > F in this case is the p-value associated with the default test statistic for this function which is the Pillai's Trace statistic. Therefore, with for alpha level of 0.05, the data is deemed statistically significant and the null hypothesis is rejected.

MANOVA does not tell us which groups differ from others and therefore we can better interpret its results by performing a post-hoc test. First, we can look at univariate tests.

```
> summary.aov(manModel2)
 Response 1 :
             Df Sum Sq Mean Sq F value Pr(>F)
indvar2       1      3   2.733  0.0149 0.9028
Residuals   248  45332 182.790               

 Response 2 :
             Df     Sum Sq    Mean Sq F value    Pr(>F)   
indvar2       1 3.4024e+09 3402367231  7.1135 0.008154 **
Residuals   248 1.1862e+11  478294777                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure Z: Univariate summary of MANOVA model

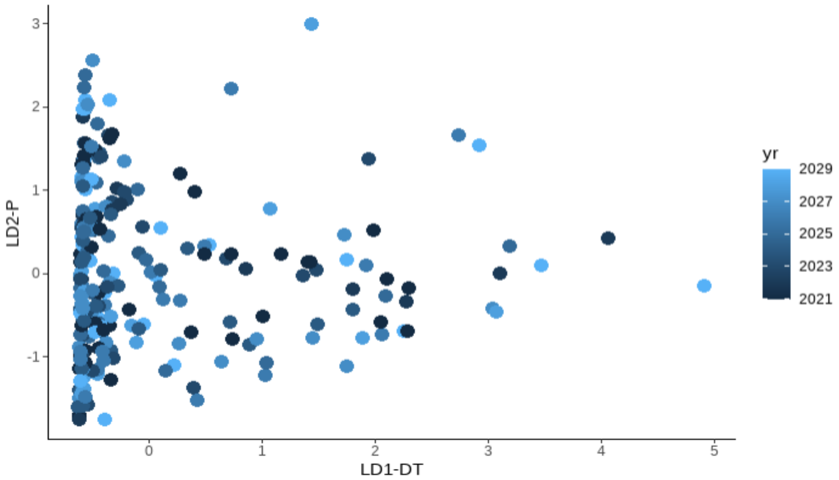
Figure Z informs us that one variable, response 2 or price, is highly significant among years of purchase.

Linear Discriminant analysis finds the linear combination of features that best separates the groups..

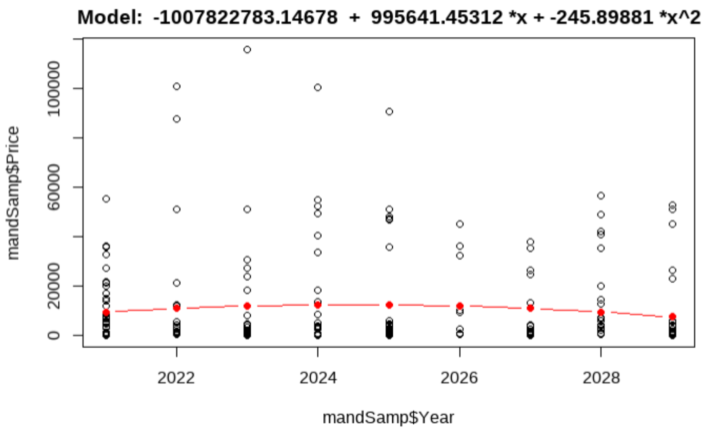
The predict() function allows us to obtain the linear discriminants and integrate them with the independent variable for visualisation purposes.

Figure AA provides a scatterplot of the LD analysis data-frame that has been created.

Figure AA: Linear Discriminant analysis scatterplot and polynomial regression fit



Upon visualisation one would hope to see one or more clear groups form but the colour gradient for the effective variable remains relatively constant. A regression model might be able to provide more information on this relationship.



Here this polynomial regression line indicates a relatively stable relationship between price and purchase years with the most outliers around the years of 204-2025. Although the effect was deemed significant, this relationship will not prove very useful to the firm.

Part 6.1: Reliability of the Service & Part

Problem 6&7 Calculations:

```
#6.1
#Let L(X) = $45 (Ass. loss func. of Taguchi) // L(X)=k(x-T)^2
Lx<-45
xminT<-0.04
q6.1_6 <- Lx/((xminT)^2) #L(X)=28125(x-T)^2
Lx2<-35
q6.1_7A<-Lx2/((xminT)^2) #L(X)=21875(x-T)^2
xminT2<-0.027
q6.1_7B<-q6.1_7A*xminT2^2
```

Q6 => L(X)= 28125 (x-T)^2 ; *TAGUCHI LOSS FUNCTION FOR L(X)=45 & x-T=0.04*

Q7A => L(X)=21875(x-T)^2; *TAGUCHI LOSS FUNCTION FOR L(X)=35 & x-T=0.04*

Q7B => 15.94687; *TAGUCHI LOSS FOR x-T=0.027cm & L(X)=35*

Part 6.2: Reliability of the Service & Part

Problem 6.2 Calculations:

```
RA<-0.85
RB<-0.92
RC<-0.90
q6.2_7A<-RA*RB*RC
RA2<-1-(1-RA)^2 #1- Probability of failure of both
RB2<-1-(1-RB)^2
RC2<-1-(1-RC)^2
newRel<-RA2*RB2*RC2
q6.2_7B<-newRel-q6.2_7A #Reliability improvement
```

6.2A=> 0.7038

6.2B=> 0.2577316 => 26% improvement

Management should reconsider flow of operations if breakdowns or stoppages are evident within the system. Parallel machining will allow for a more constant flow and a whopping 26% improvement in overall reliability. The less dependent the operations are on one-another, the better.

Part 6.3: Reliability of the Service & Part

The probabilities for this problem obtained by using the binomial distribution formula (dbinom in r):

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

P(no vehicles avail.)= 0.007064264

P(no drivers avail.)= 0.00306887

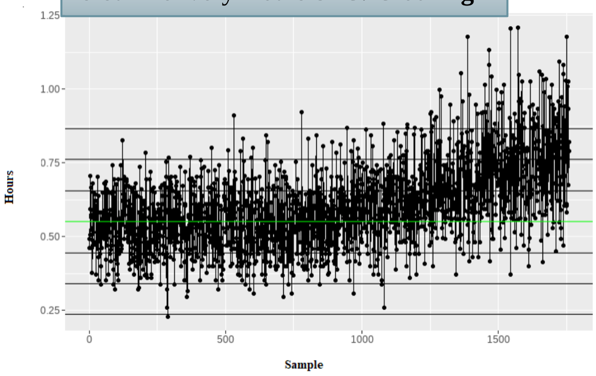
Therefore, we can estimate the number of reliable delivery days per year by multiplying the probability of having available vehicles with the probability of having available drivers.

P(reliable) = (1- P(no vehicles avail.))*(1- P(no drivers avail.) = 0.9898885

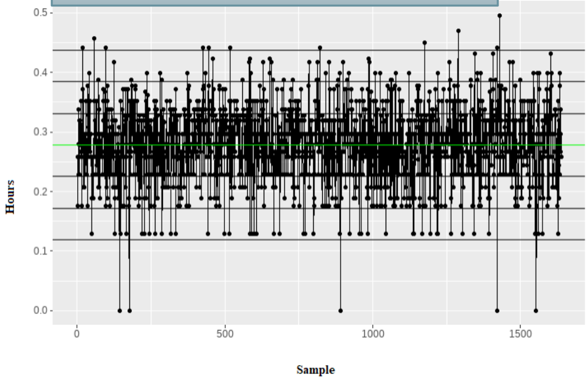
Expected Reliable Days = 0.9898885x365 = 361.3093 = 361 Days

Appendix A S-Charts:

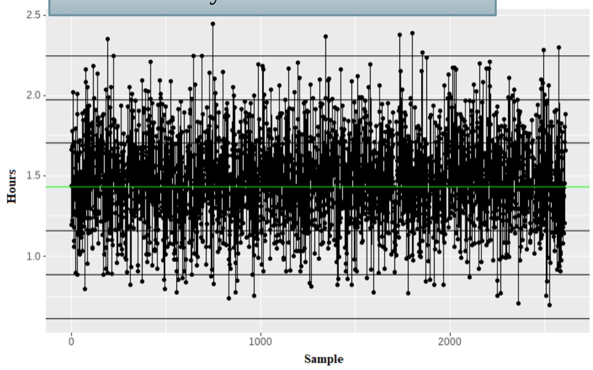
s-bar Delivery Hours SPC: **Clothing**



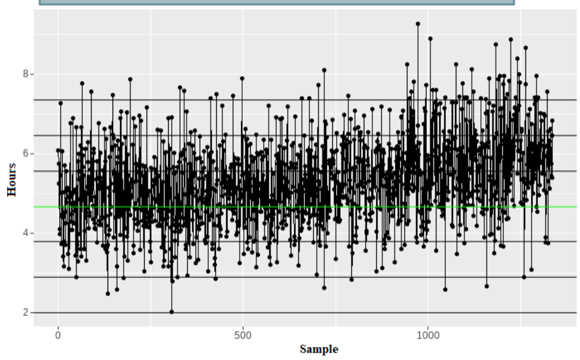
s-bar Delivery Hours SPC: **Food**



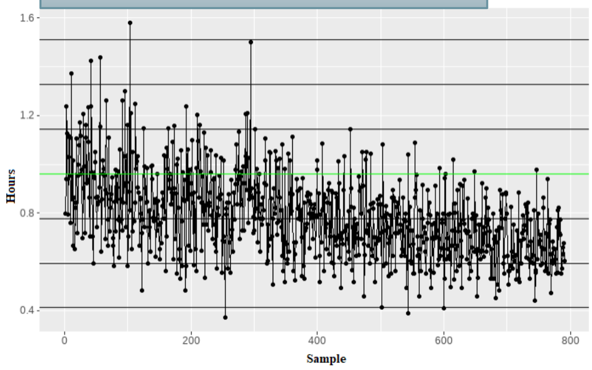
s-bar Delivery Hours SPC: **Gifts**



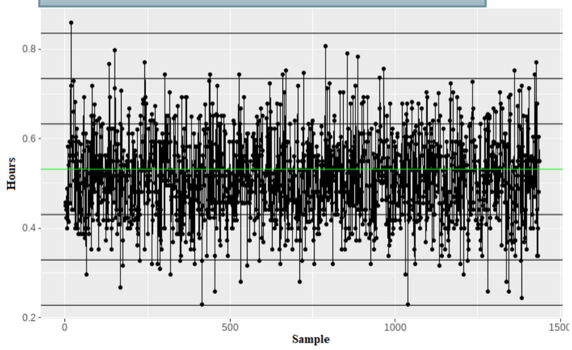
s-bar Delivery Hours SPC: **Household**



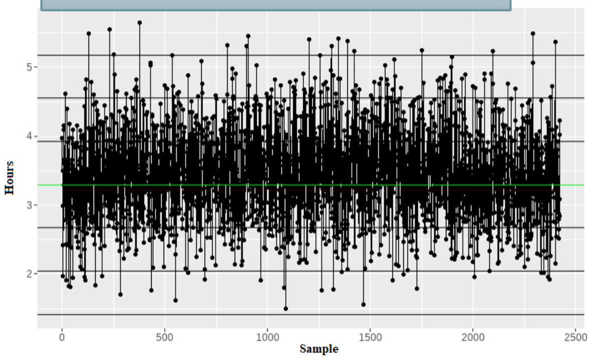
s-bar Delivery Hours SPC: **Luxury**



s-bar Delivery Hours SPC: **Sweets**



s-bar Delivery Hours SPC: **Technology**



Conclusion:

After getting to know the data there are a few things that can be concluded. The company receives the majority of its revenue from sales of Luxury and Technology items. It seems the company has been prioritising improving the delivery times of these items. It is suggested that the firm investigates the effect of a similar initiative for technology items while also keeping track on any potential effects on the performance of the delivery team elsewhere. It can also be concluded that the least effective communications channel is ‘spam’ – this form of communication is somewhat outdated and should be replaced with the implementation of a mailing list for dedicated customers. Finally, it should be reiterated that the optimal mean delivery time given the costs associated with improving infrastructure is 16 hours.

Bibliography:

Bai, D.S. and Choi, I.S. (1995) X and R control charts for skewed populations. Journal of Quality Technology, 27(2), pp.120-131.

Bhandari, P. (2022) Type I & Type II Errors | Differences, Examples, Visualizations. Scribbr. Retrieved October 23, 2022, from https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

Dario Radečić (2022) MANOVA in R – How To Implement and Interpret One-Way MANOVA Available at: https://www.scribbr.co.uk/referencing/harvard-website-reference/#:~:text=To%20reference%20a%20website%20in,which%20you%20accessed%20the%20website.&text=Author%20surname%2C%20initial.,Accessed%3A%20Day%20Month%20Year). (Accessed: 20 10 2022).

Evans, J.R. and Lindsay, W.M. (2013) Managing for quality and performance excellence. Cengage Learning.

French, A., Macedo, M., Poulsen, J., Waterson, T. and Yu, A.(2008) Multivariate analysis of variance (MANOVA).

Ipek, H.A.L.İ.L., Ankara, H.Ü.S.E.Y.İ.N. and Ozdag, H. (1999) The application of statistical process control. Minerals Engineering, 12(7), pp.827-835.