ECSA Project

23580402

CA Smit

Contents

Introduction

This report will detail the processes and methods used to do statistical analysis on a dataset populated with sales and their respective features. Conclusion on whether processes are in or out of control will be drawn and hypothesis will be tested and significance levels tested.

Data wrangling

The features of the sales dataset are:

X: This is the key identification feature. This describes the sale's position in the dataset. The data type is integer.

ID: This is also an identification feature. It is most likely the ID the point-of-sale system labels each sale with. The data type is integer.

AGE: This is the age of the customer. The data type is integer.

Class: This is the class of the item bought. This is a categorical feature with 7 unique values. They are: "Sweets", "Household", "Gifts", "Technology", "Luxury", "Food" or "Clothing". The data type is character.

Price: This numerical feature is the amount paid for the product. The data type is number.

Year: This feature is the year in which the sale was made. The data type is integer.

Month: This feature is the month in which the sale was made. The data type is integer.

Day: This feature is the day in which the sale was made. The data type is integer.

Delivery.time: This feature is the time it took to deliver the product. The data type is number. This is a categorical feature with 6 unique values. They are: "Recommended", "Website", "Random", "Browsing", "EMail", "Spam". The data type is character.

Summary of descriptive features

AGE

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
18	38	53	54.57	70	108

Price

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	NA's
-588.8	482.3	2259.6	12293.7	15270.7	116619	17

Delivery.time

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
0.5	3	10	14.5	18.5	75

Class

Class	Count
Clothing	26406
Food	24588
Gifts	39154
Household	20067
Luxury	11869
Sweets	21566

Most frequent class: Gifts

Least frequent class: Luxury

Why.Bought

Why.Bought	Count
Browsing	18995
Email	7225
Random	13122
Recommended	107000
Spam	4208

Most frequent class: Recommended

Least frequent class: Spam

Data partition

Some invalid data instances exist in the dataset. To perform statistical process control we must remove these instances from the dataset. There are 2 types of invalid data in this dataset:

Negative value: This is an instance with a negative price. Instance 16320 is an example of this.

Missing data: This is an instance with a missing price. Instance 16321 is an example of this.

Instance 16319 is an example of valid data.

16319	16319	94129	46	Food	654.58	2028	11	29	2.5	Recommended
16320	16320	44142	82	Household	-588.80	2023	10	2	48.0	EMail
16321	16321	81959	43	Technology	*NA*	2029	9	6	22.0	Recommended

Figure 1 Examples of valid and invalid data

Partitioning the data results in a dataset of valid data with 179978 instances of data and a dataset of invalid data with 22 instances of data.

	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
463	463	47101	50	Clothing	1030.86	2021	1	1	9.0	Recommended
2627	2627	88087	21	Clothing	428.03	2021	1	1	10.0	Recommended
3374	3374	25418	68	Household	13184.41	2021	1	1	48.5	Website
5288	5288	13566	94	Household	7021.90	2021	1	1	42.0	Recommended
8182	8182	84692	35	Clothing	475.18	2021	1	1	9.0	Recommended
9272	9272	46305	72	Clothing	580.98	2021	1	1	8.5	Random
9712	9712	92105	45	Household	6877.00	2021	1	1	43.0	Recommended
12163	12163	21614	27	Clothing	513.13	2021	1	1	9.5	Recommended

Figure 2 Extract from the valid data

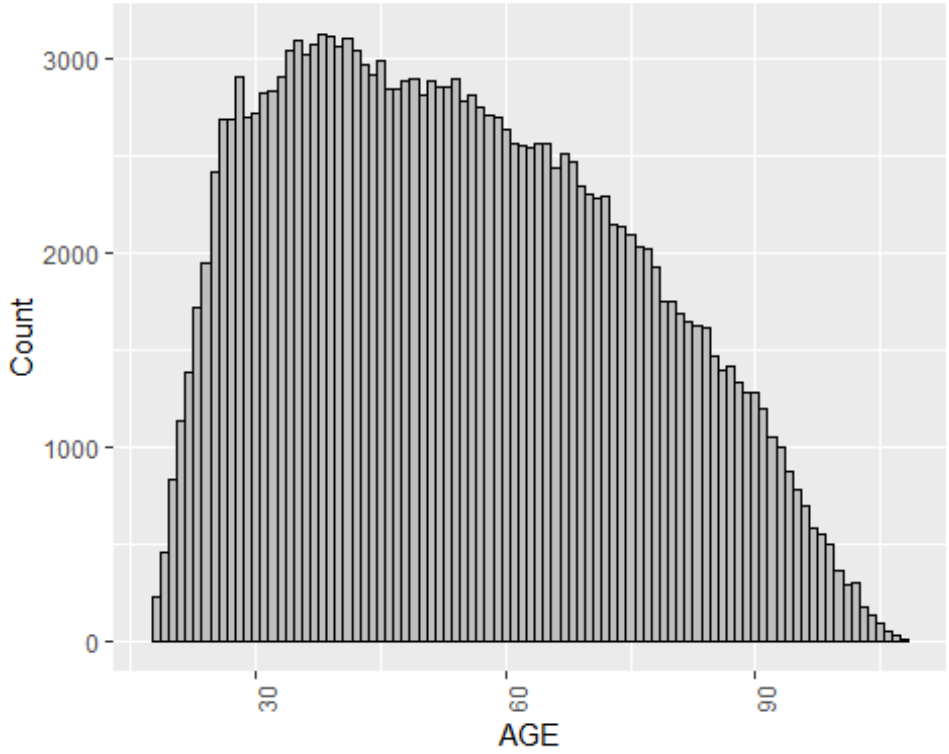
	X	ID	AGE	Class	Price	Year	Month	Day	Delivery.time	Why.Bought
144443	144443	37737	81	Food	-588.8	2022	12	10	2.5	Recommended
16320	16320	44142	82	Household	-588.8	2023	10	2	48.0	EMail
19998	19998	68743	45	Household	-588.8	2024	7	16	45.5	Recommended
155554	155554	36599	29	Luxury	-588.8	2026	4	14	3.5	Recommended
19540	19540	65689	96	Sweets	-588.8	2028	4	7	3.0	Random
98765	98765	64288	25	Clothing	*NA*	2021	1	24	8.5	Browsing
54321	54321	62209	34	Clothing	*NA*	2021	3	24	9.5	Recommended
34567	34567	18748	48	Clothing	*NA*	2021	4	9	8.0	Recommended

Figure 3 Extract from the invalid data

Visualizations of data

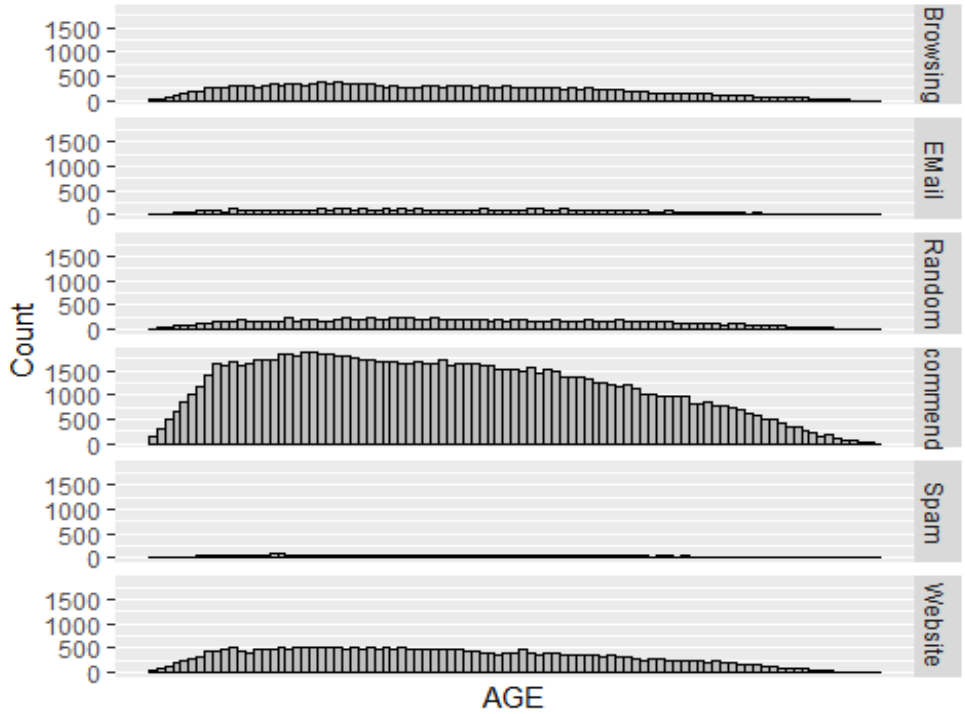
AGE

Age of customer for every sale



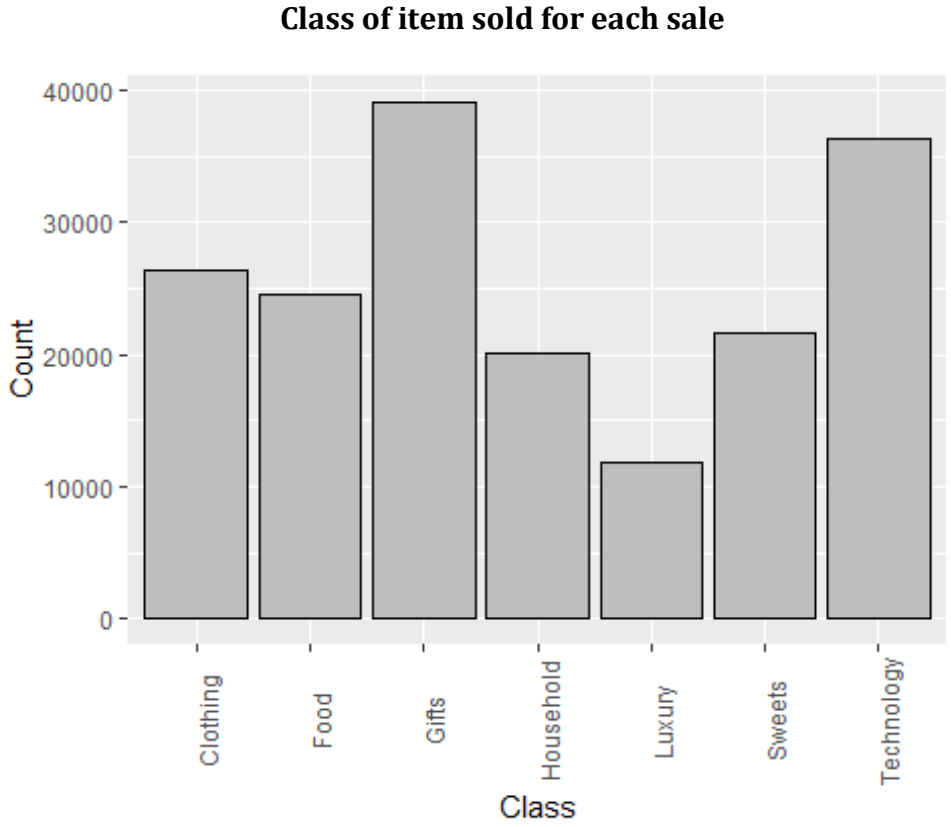
The ages of our customer base are close to normally distributed but is skewed a bit to the right. Our customers are generally younger.

Age of customer for every sale categorized by why they bought



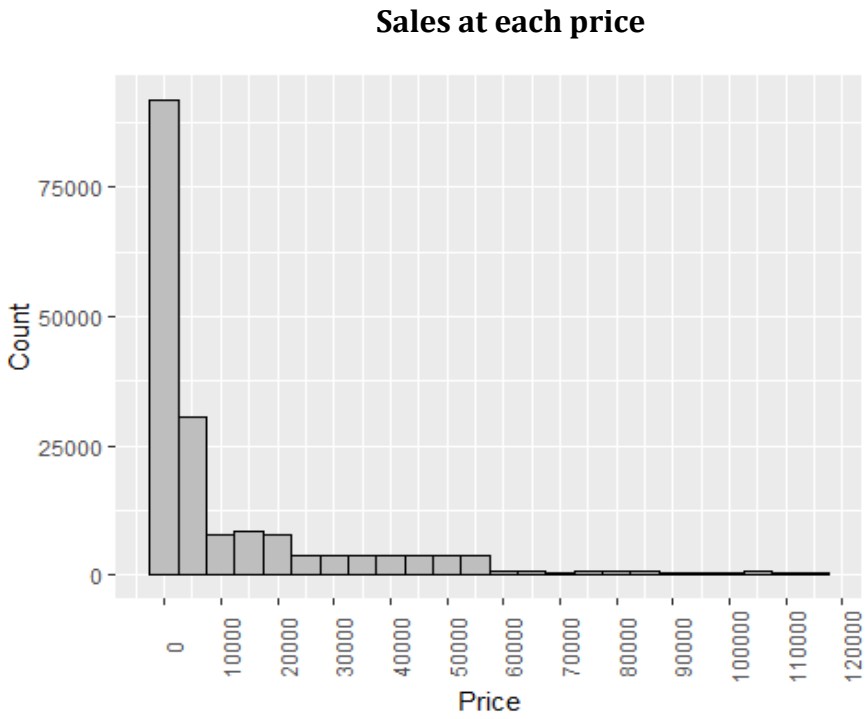
These graphs all mimic the distribution of larger graph. We can therefor conclude that there is not any specific marketing method that is effective for a specific age group.

Class



The class sales are mostly uniform. Gifts and Technology is sold a bit more than the rest and Luxury is sold a bit less. This could be due to the fact Luxury items have higher prices overall.

Price



The graph is unimodal and skewed to the right. Most of the items' prices are near the median of 12294.10

We can more accurately describe the distribution by checking how many items cost more than and less than the mean.

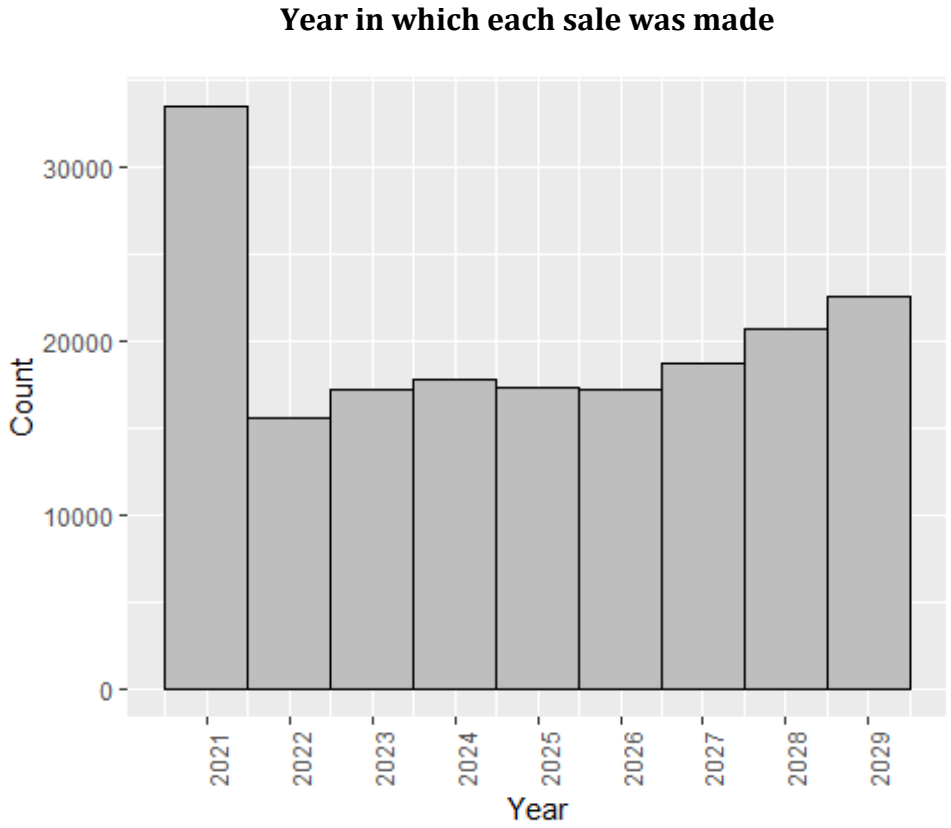
```
More than mean:   49881

Less than mean:   130097
```

Most of the prices are below 10000, but due to a few heavy lifting values the mean ends up at 12294.10.

The different classes can also be contrasted in terms of price. From the table it is clear that the higher value items are classified as Luxury, Technology or Household and lower value items are generally Clothing, Food, Gifts or Sweets.

Class <chr>	Min price <dbl>	Mean price <dbl>	Max price <dbl>	Median price <dbl>
Clothing	127.76	640.5253	1154.02	642.04
Food	127.76	407.8153	691.96	408.37
Gifts	172.61	2961.8414	5774.49	2961.59
Household	127.76	11009.2738	21935.33	10960.88
Luxury	12825.37	64862.6386	116618.97	65342.14
Sweets	35.65	304.0704	576.38	303.25
Technology	935.18	29508.0626	57735.40	29653.90

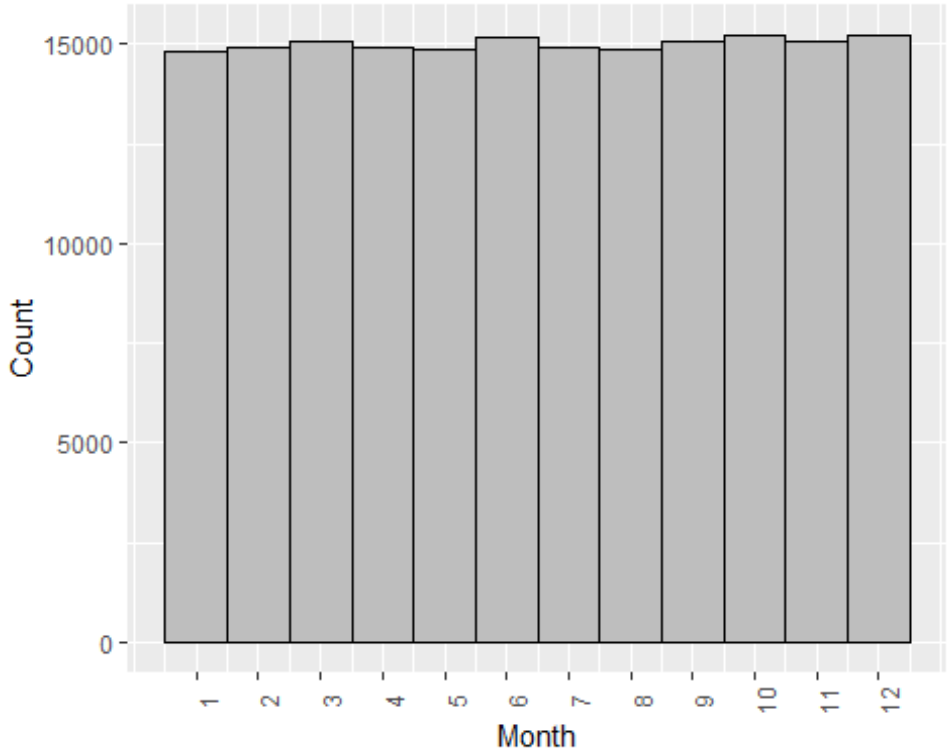
Year



Sales seems to increase with each passing year, but 2021 is an outlier to this being the earliest year and having the most sales.

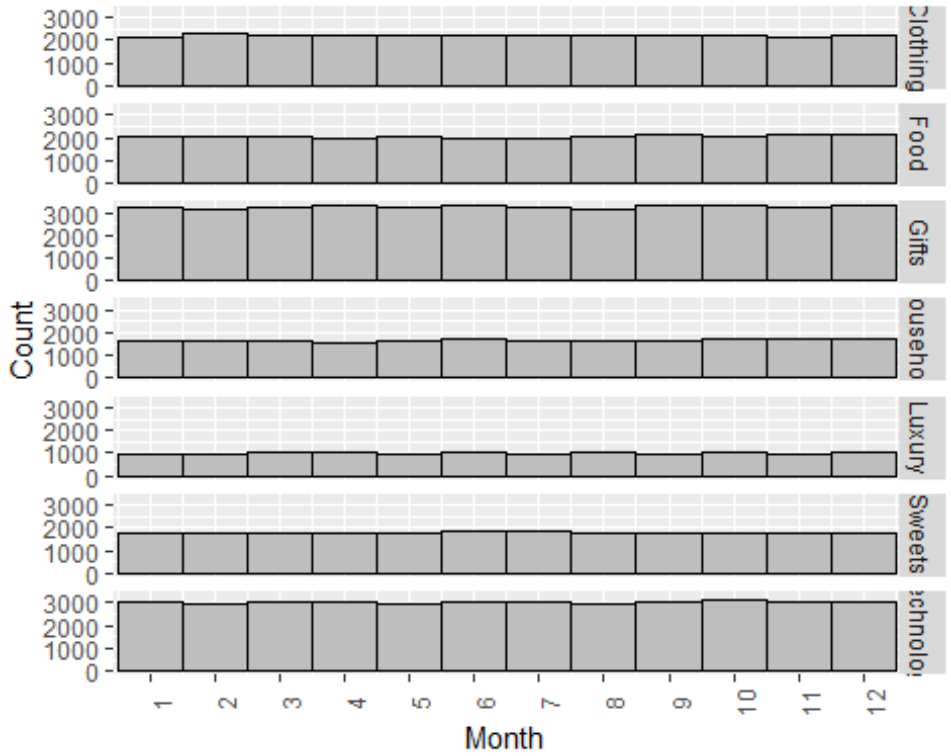
Month

Sales per month



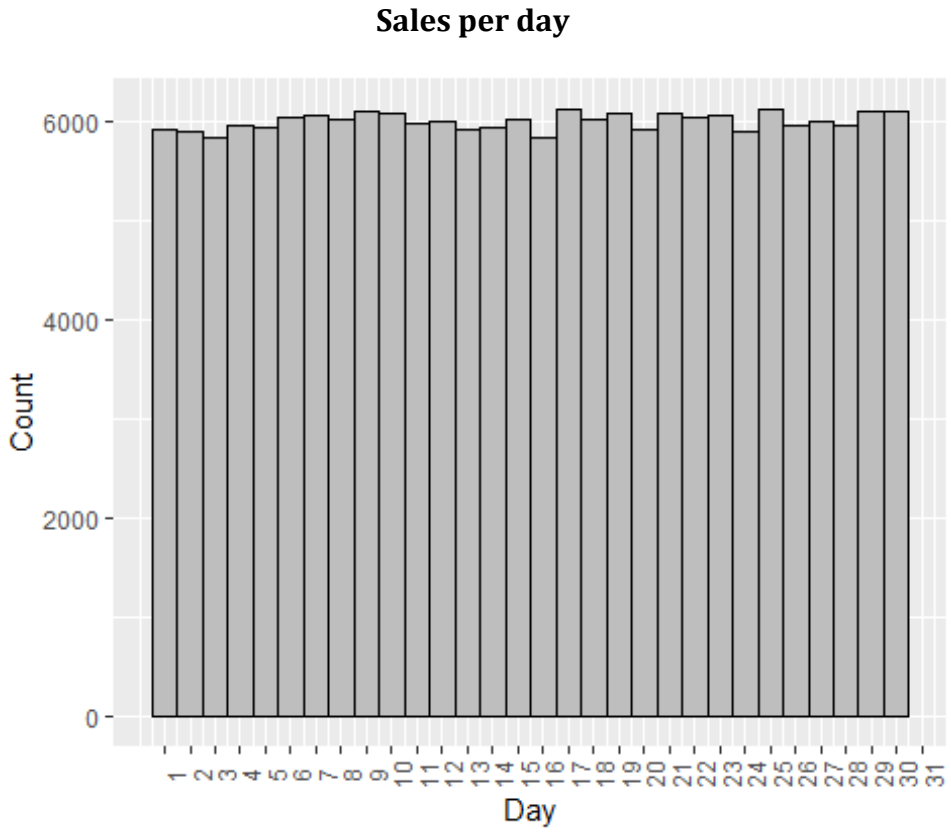
Sales per month is uniform indicating that there is not any seasonality within the year in sales.

Sales per month per class

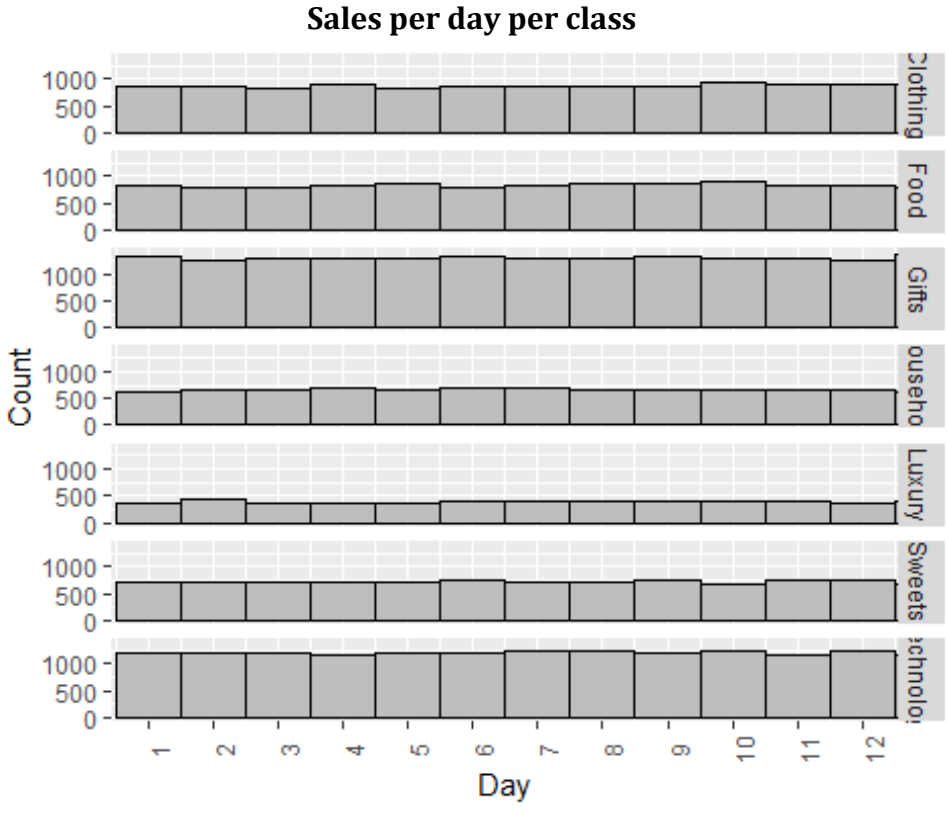


This non-seasonality holds for every class individually as well.

Day



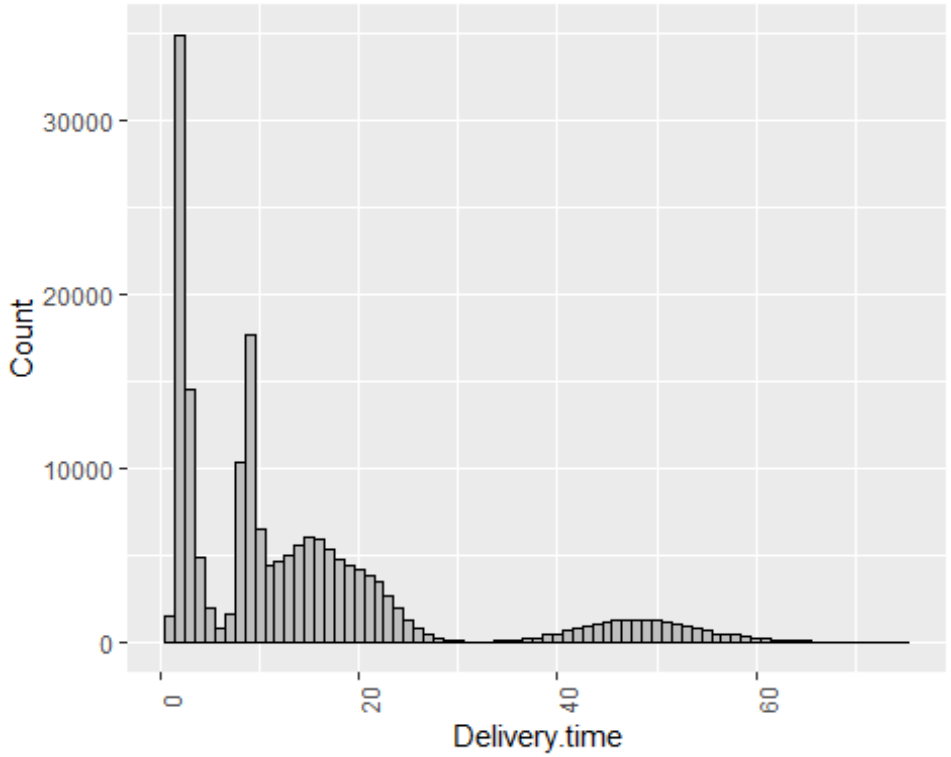
Sales per month is uniform indicating that there is not any seasonality within the month in sales.



This non-seasonality holds for every class individually as well.

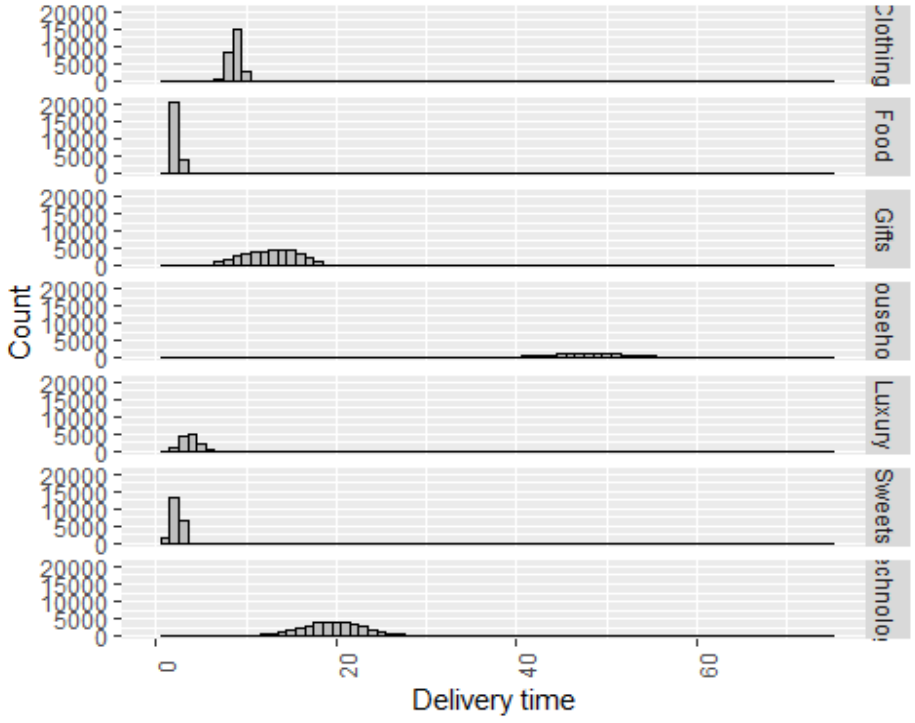
Delivery.Time

Delivery time for each sale



Delivery time is skewed to the right and has 2 somewhat normal bumps near 17 days and 48 days.

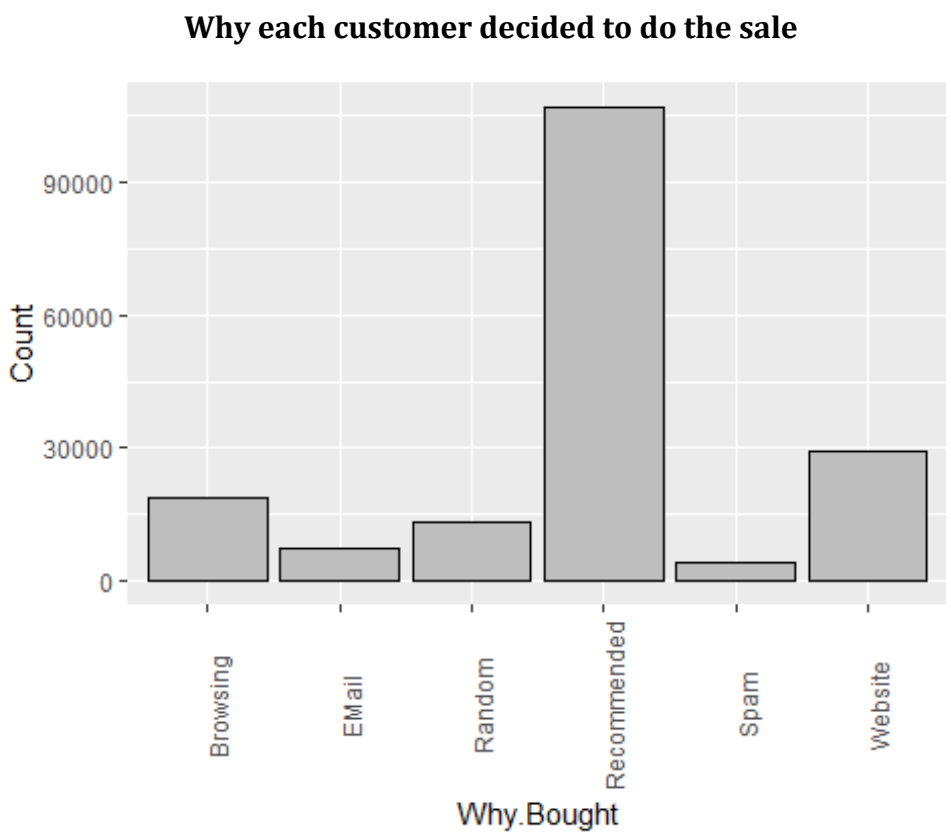
Delivery time for each sale per class



It is clear from these plots that correlation between delivery time and the class of the item exists.

Clothing sales are unimodal and skewed to the right taking an average of 9.00 days to deliver. Food sales are unimodal and skewed to the right taking an average of 2.50 days to deliver. Gifts sales are normally distributed taking an average of 12.89 days to deliver. Household sales are normally distributed taking an average of 48.72 days to deliver. Luxury sales are somewhat normally distributed taking an average of 3.97 days to deliver. Sweets sales are unimodal and skewed to the right taking an average of 2.50 days to deliver. Gifts sales are normally distributed taking an average of 20.01 days to deliver.

Why.Bought



The reason why customer buy items is uniform, but recommendation is an outlier. The most effective way to sell an item is to have someone recommend it.

Process capability indices

The process capability indices for the delivery time of technology class items are calculated using an UCL of 24 and an LCL of 0. An LCL = 0 is logical, because it is the lowest possible amount of days, we can take to deliver an item.

$Cp = (USL-LSL)/(6\sigma)$	$Cpu = (USL - \mu)/(3\sigma)$
$Cpl = (\mu - LSL)/(3\sigma)$	$Cpk = \min(Cpl,Cpu)$

$\sigma = 3.502$

$\mu = 20.011$

$Cp = (24 - 0)/6(3.502)$

$Cp = 1.142$

$Cpu = (24 - 20.011)/3(3.502)$

$Cpu = 0.38$

$Cpl = (20.011-0)/3(3.502)$

$Cpl = 1.905$

$Cpk = 0.38$

Standard Deviation	Mean	Cp	Cpl	Cpu	Cpk
3.502	20.011	1.142	1.905	0.38	0.38

Statistical process control

X-chart limits

Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	9.405	9.26	9.115	8.97	8.825	8.68	8.535
Household	50.248	49.02	47.791	46.562	45.334	44.105	42.876
Food	2.709	2.636	2.563	2.49	2.417	2.344	2.271
Technology	22.975	22.108	21.241	20.374	19.508	18.641	17.774
Sweets	2.897	2.757	2.618	2.478	2.338	2.198	2.059
Gifts	9.489	9.113	8.737	8.361	7.985	7.609	7.234
Luxury	5.494	5.241	4.988	4.736	4.483	4.23	3.977

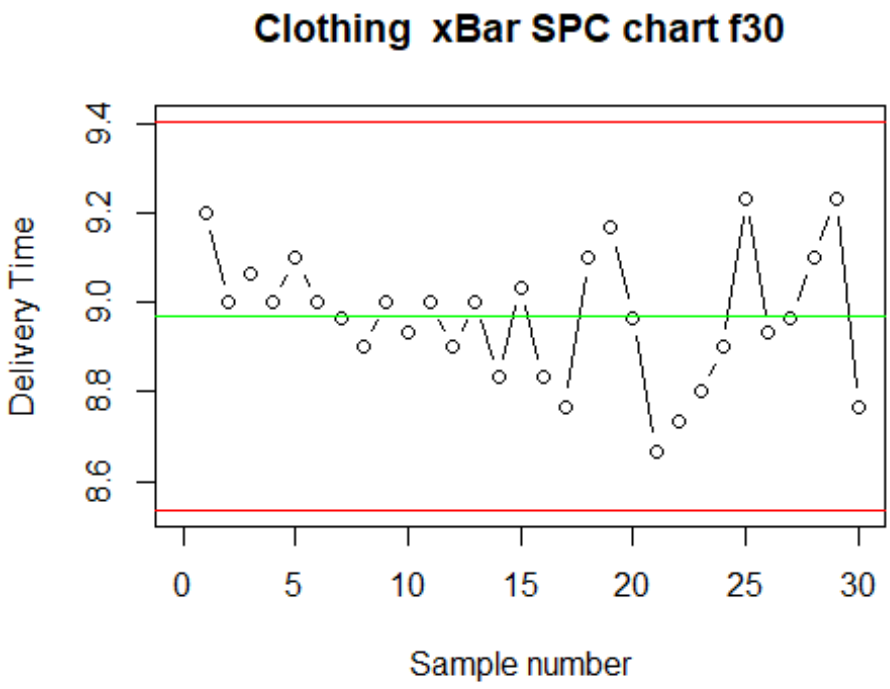
S-chart limits

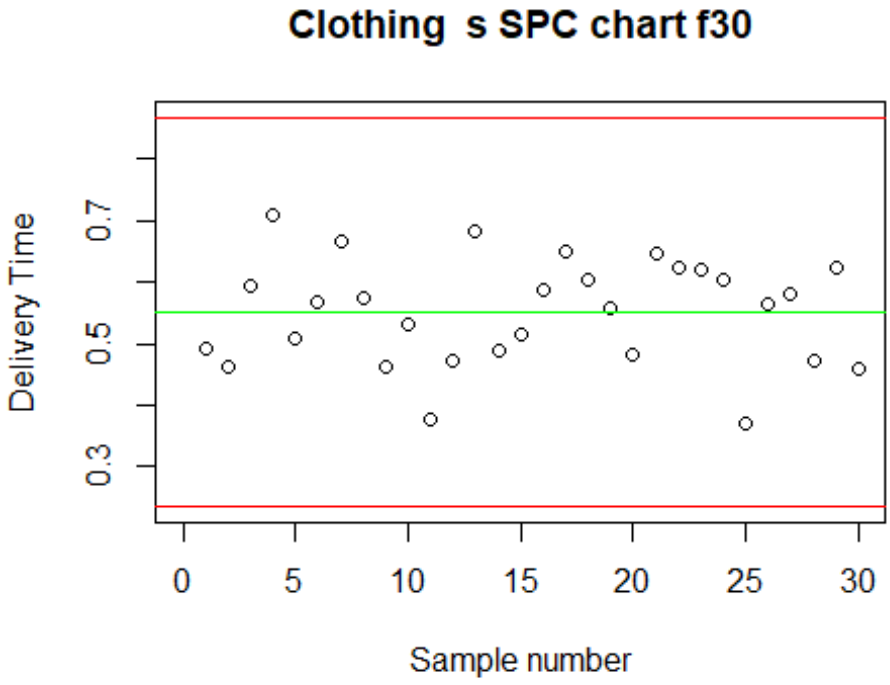
Class	UCL	U2Sigma	U1Sigma	CL	L1Sigma	L2Sigma	LCL
Clothing	0.867	0.761	0.656	0.551	0.446	0.341	0.236
Household	7.344	6.453	5.563	4.672	3.781	2.89	2
Food	0.437	0.384	0.331	0.278	0.225	0.172	0.119
Technology	5.181	4.552	3.924	3.296	2.667	2.039	1.41
Sweets	0.835	0.734	0.633	0.531	0.43	0.329	0.227
Gifts	2.246	1.974	1.701	1.429	1.157	0.884	0.612
Luxury	1.511	1.328	1.145	0.961	0.778	0.595	0.411

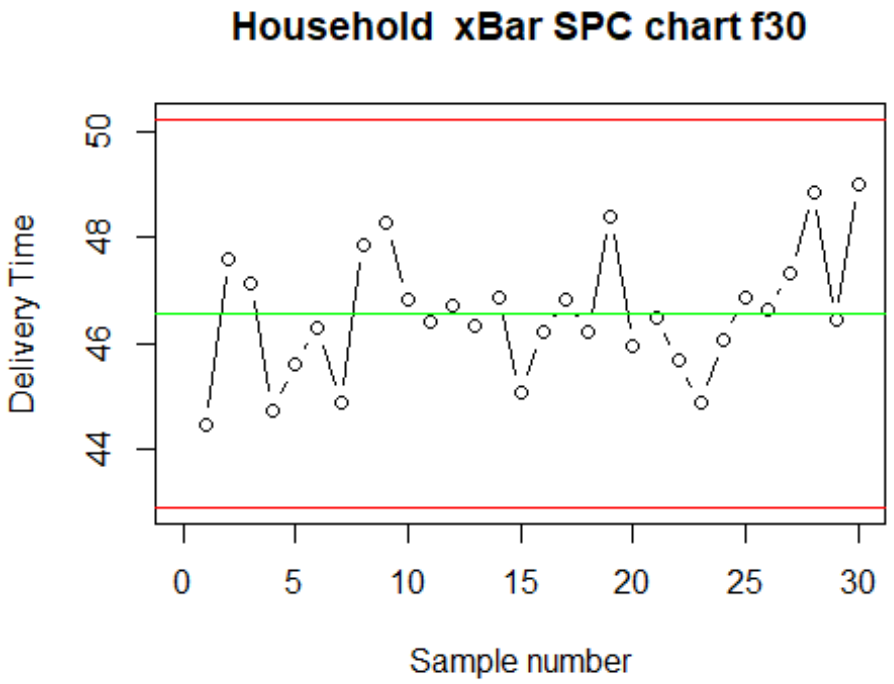
The first 30 samples of each class are used to determine the different control limits. CL is calculated as $\bar{\bar{x}}$ and $\bar{s}$ respectively. $\bar{\bar{x}}$ is the mean of all the sample means and $\bar{s}$ is the mean of all the sample standard deviations. A set of equations is then used to calculate the lower control limit and the upper control limit. These equations are: $LCL = \bar{\bar{x}} - A_3\bar{s}$ and $UCL = \bar{\bar{x}} + A_3\bar{s}$ for the X-charts and $LCL = B_3\bar{s}$ and $UCL = B_4\bar{s}$ for the s-charts. The rest of the table values are calculated by taking the difference between the UCL and the Cl or the CL and the LCL and multiplying it by $1/3$ for U1sigma and $2/3$ for U2sigma. For samples of 15 sales $A_3 = 0.789, B_3 = 0.428$ and $B_4 = 1.572$.

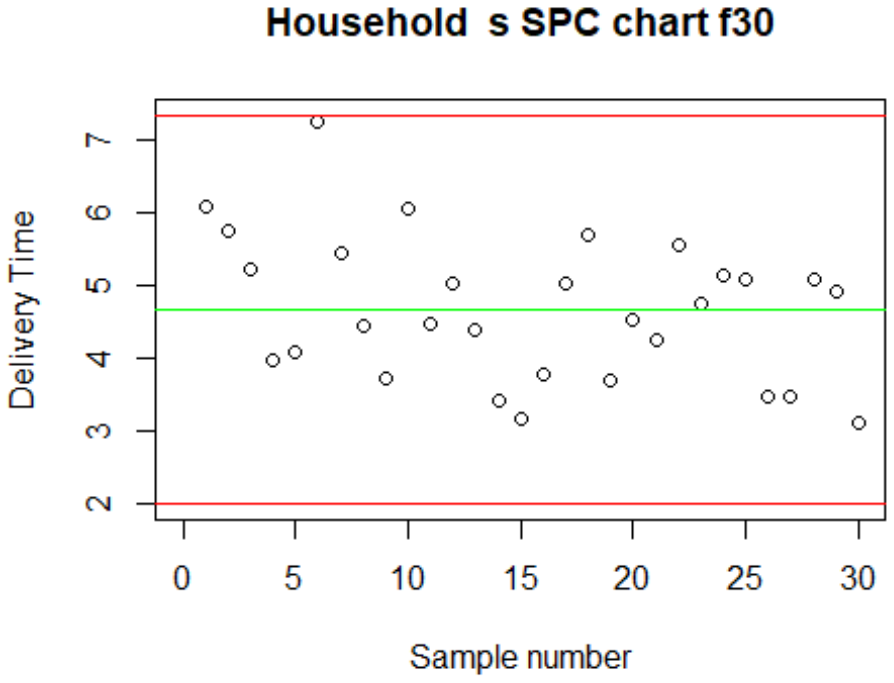
Chart analysis

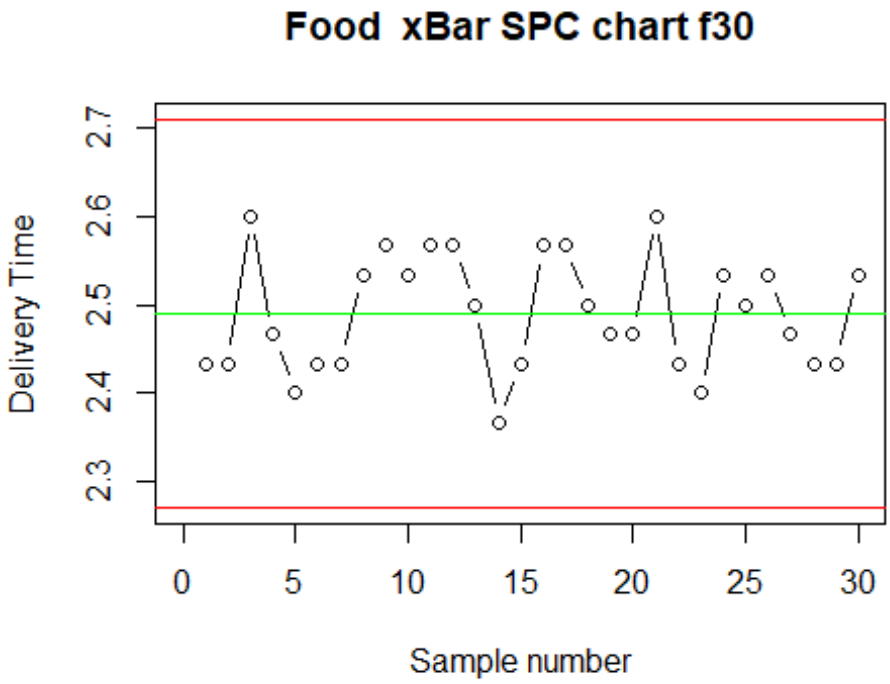
First 30 samples

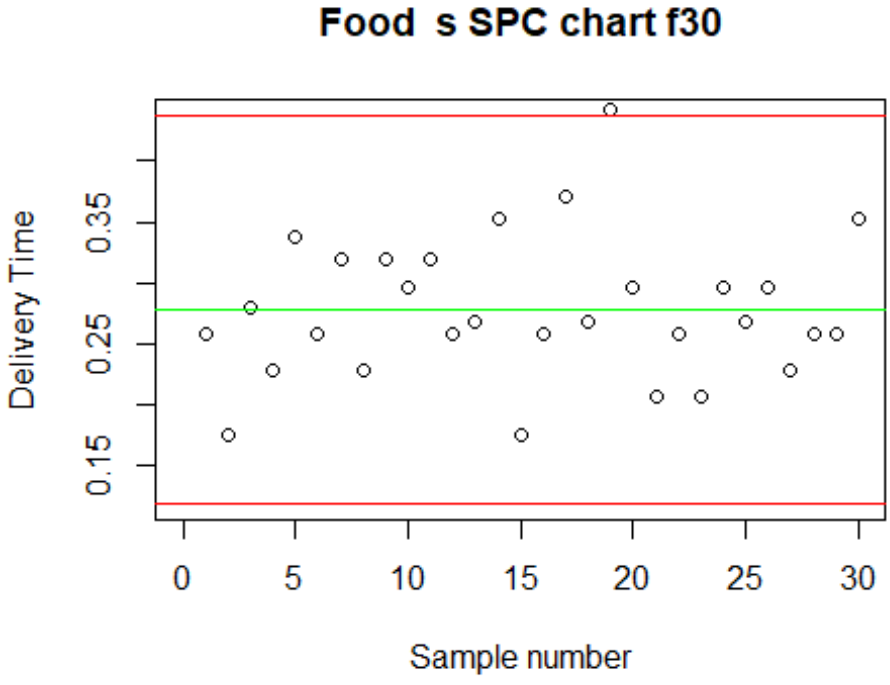




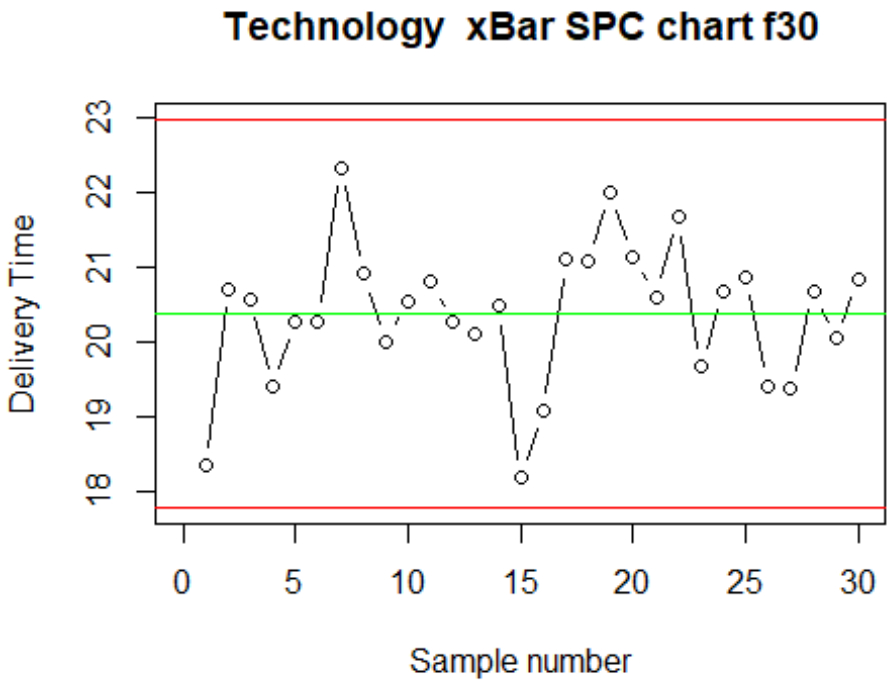

Household xBar SPC chart f30
Delivery Time
44
46
48
50
0
5
10
15
20
25
30
Sample number

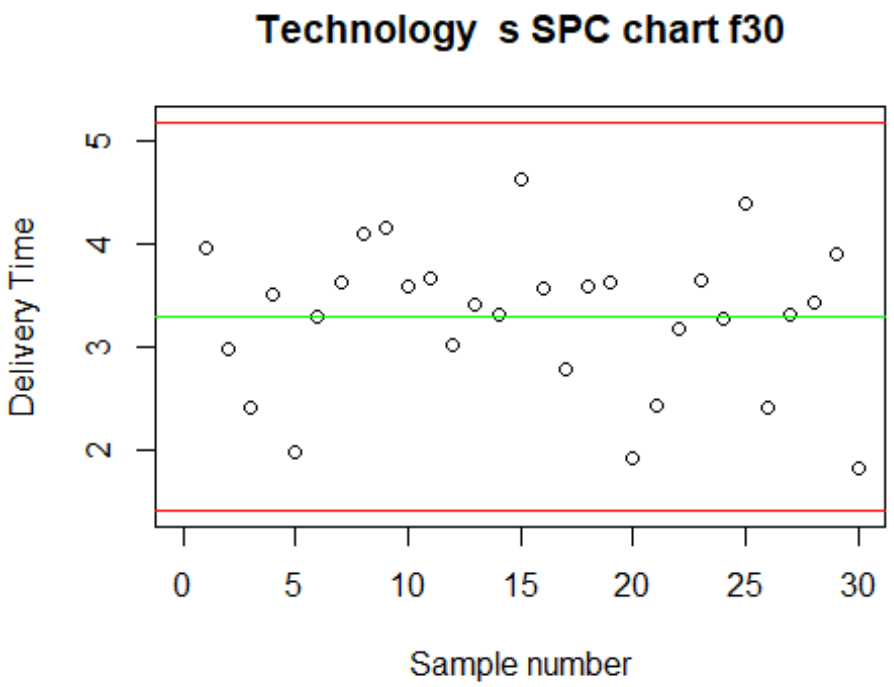

Household s SPC chart f30
Delivery Time
2
3
4
5
6
7
0
5
10
15
20
25
30
Sample number

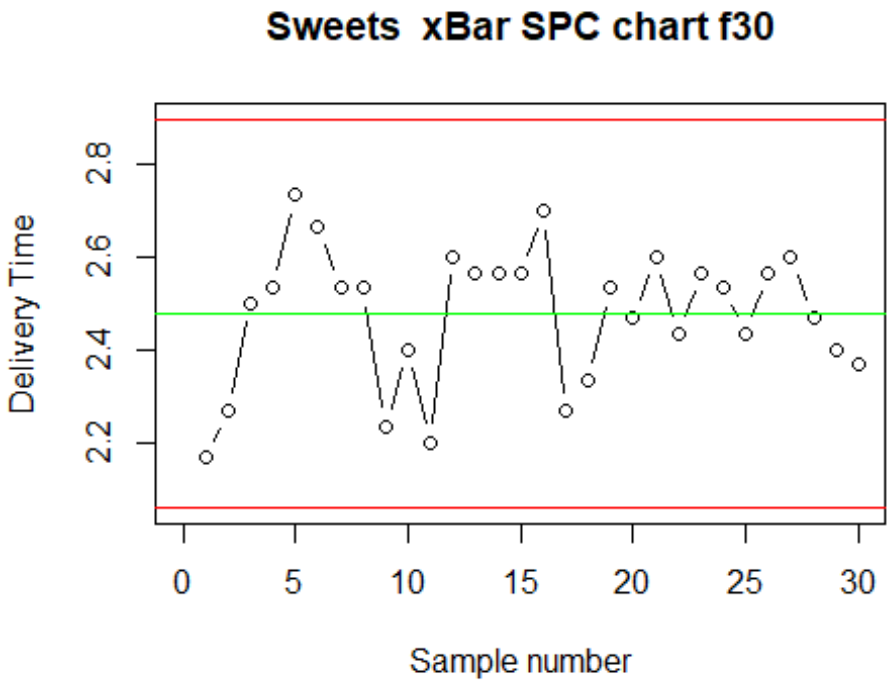


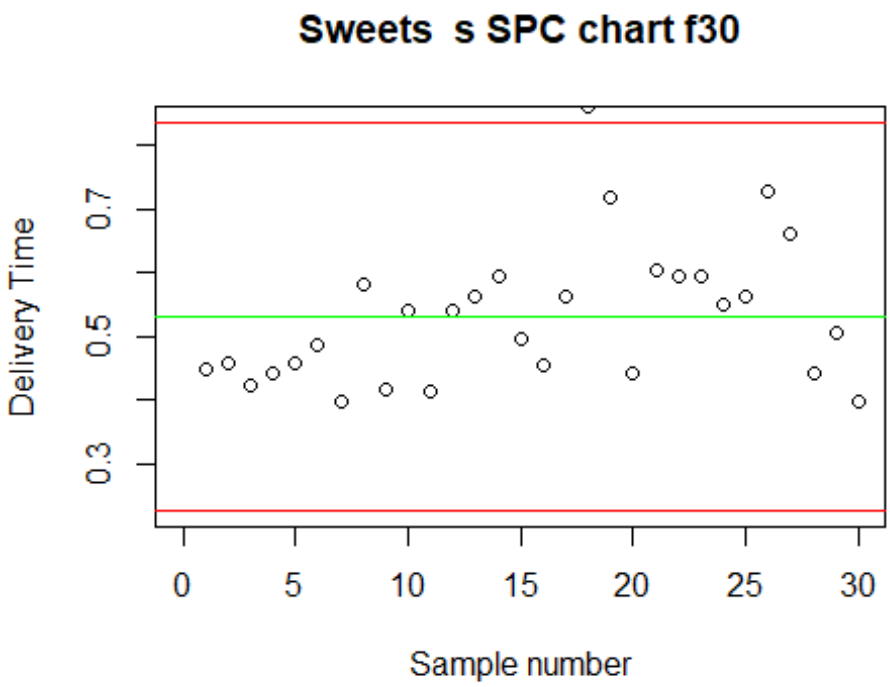


One instance out of control limits. Removed and then recalculated

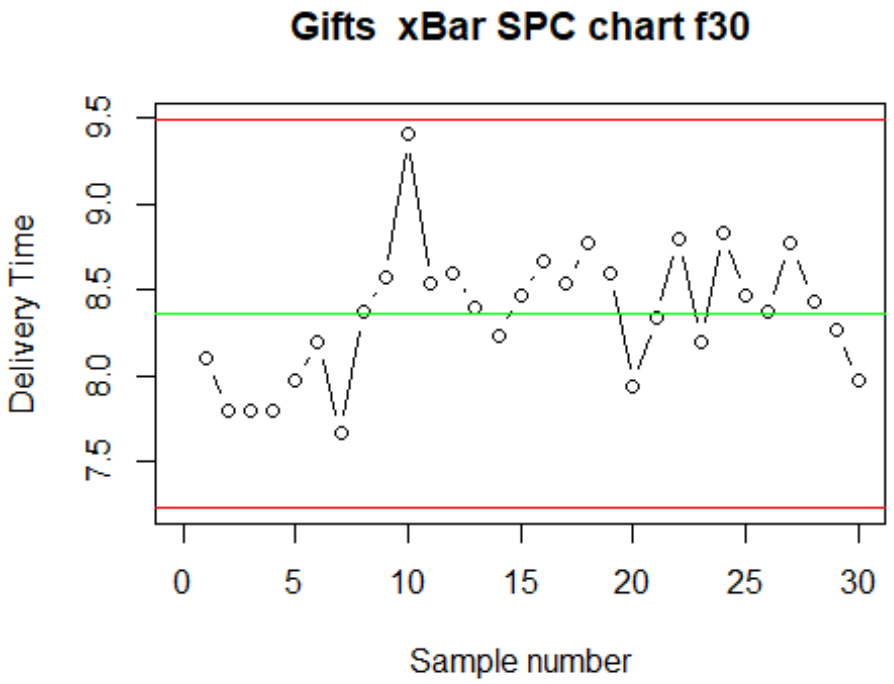

Technology xBar SPC chart f30
Delivery Time
18
19
20
21
22
23
0
5
10
15
20
25
30
Sample number

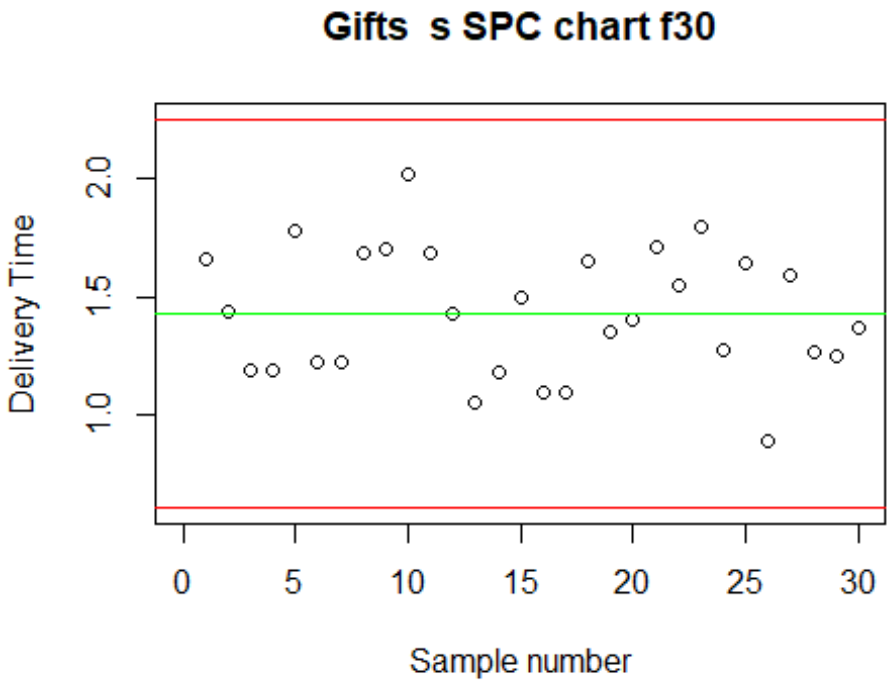

Technology s SPC chart f30
Delivery Time
2
3
4
5
0
5
10
15
20
25
30
Sample number

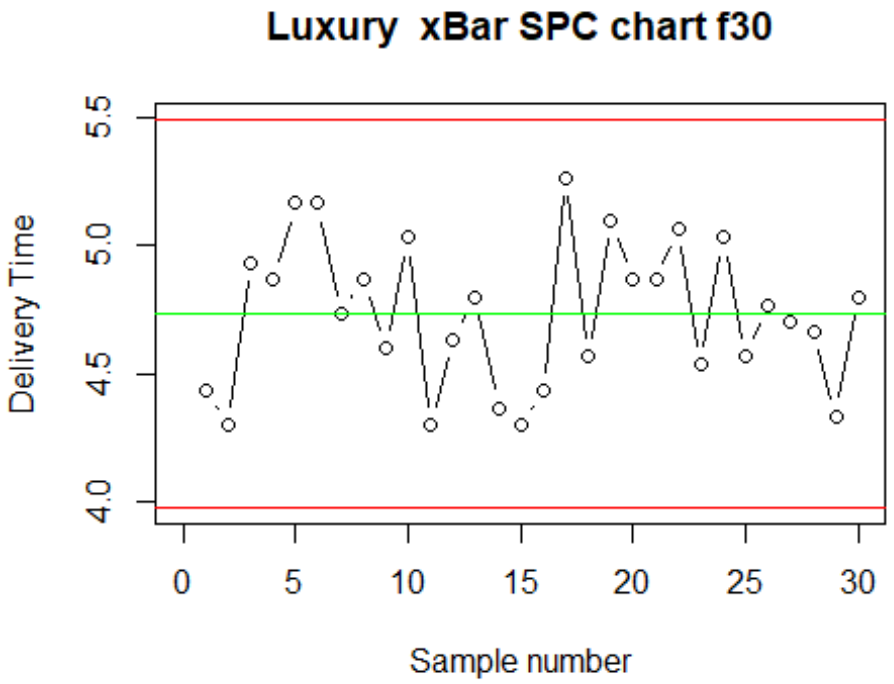


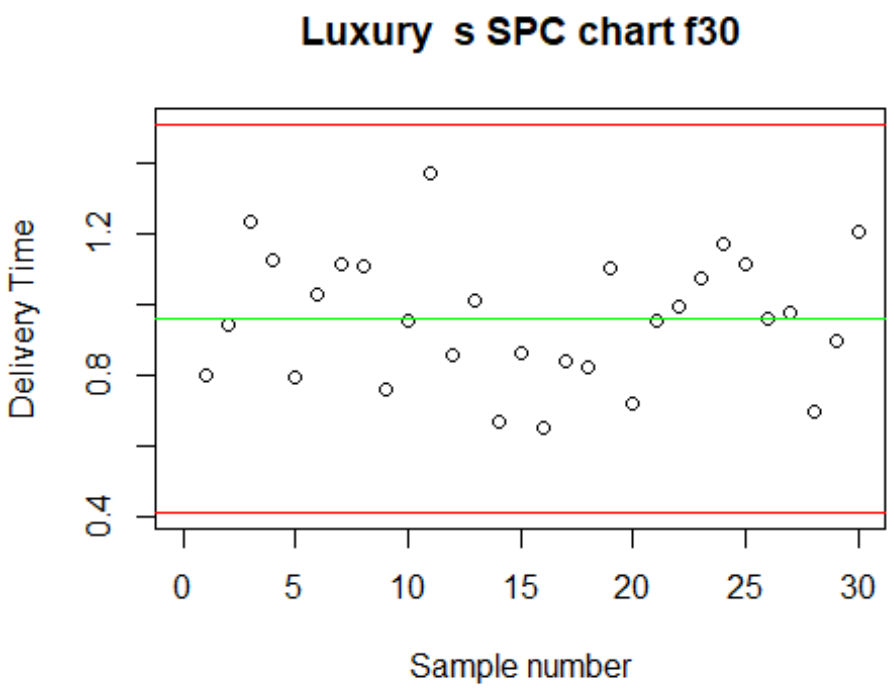


One instance out of control limits. Removed and recalculated


Gifts xBar SPC chart f30
Delivery Time
7.5
8.0
8.5
9.0
9.5
0
5
10
15
20
25
30
Sample number

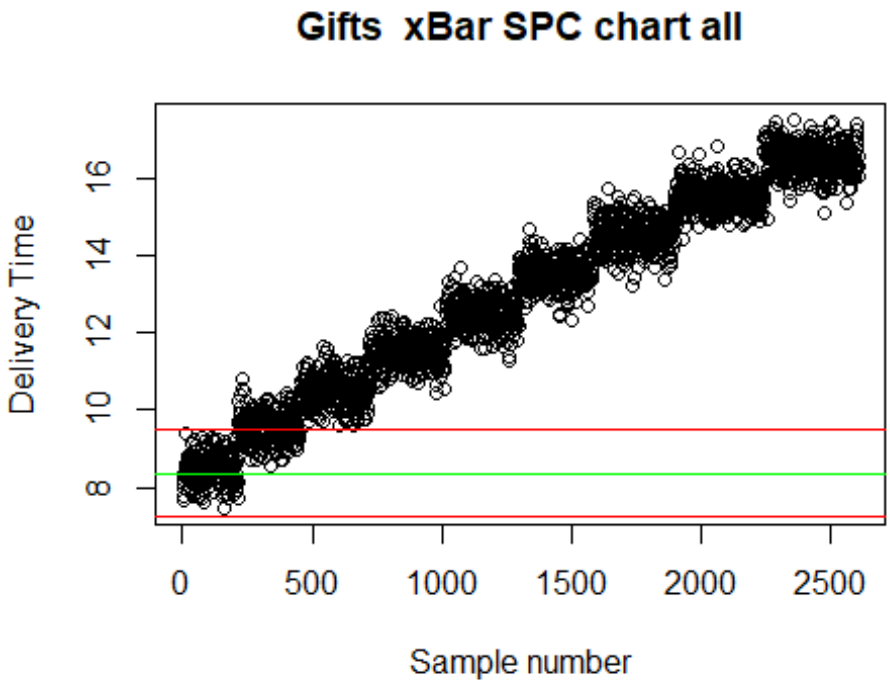

Gifts s SPC chart f30
Delivery Time
1.0
1.5
2.0
0
5
10
15
20
25
30
Sample number

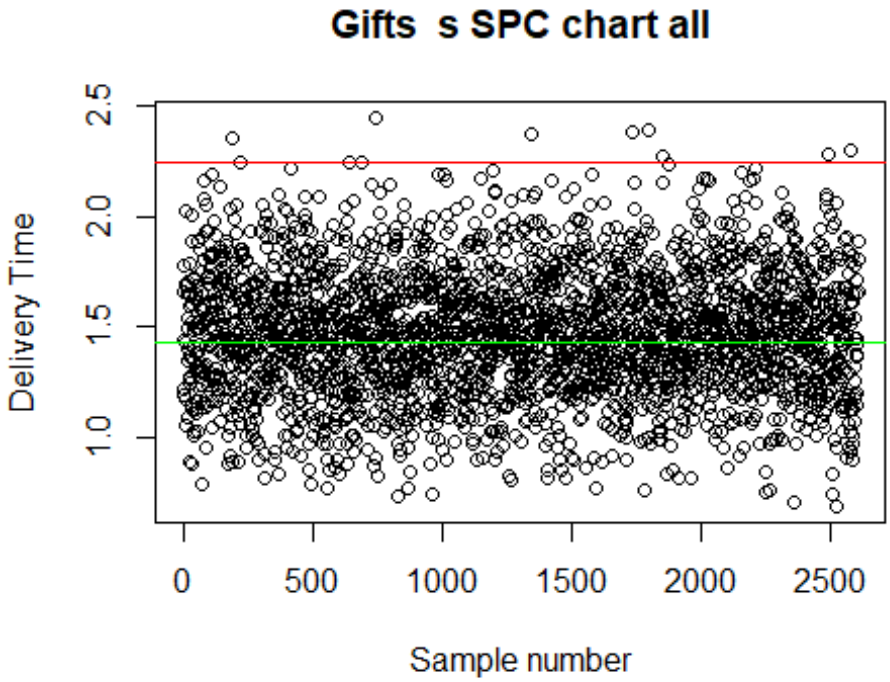

Luxury xBar SPC chart f30
Delivery Time
4.0
4.5
5.0
5.5
0
5
10
15
20
25
30
Sample number

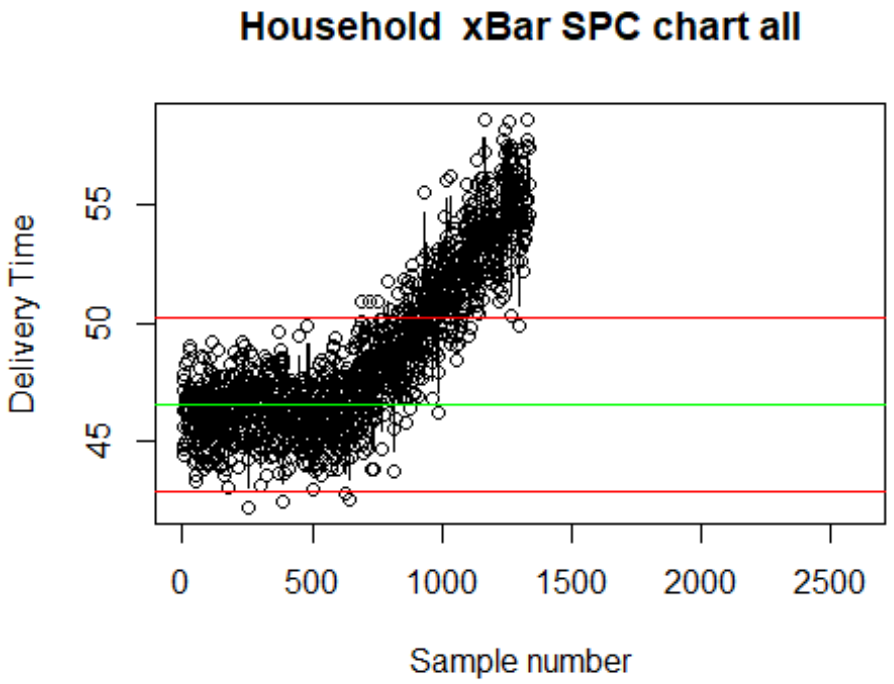

Luxury s SPC chart f30
Delivery Time
0.4
0.8
1.2
0
5
10
15
20
25
30
Sample number

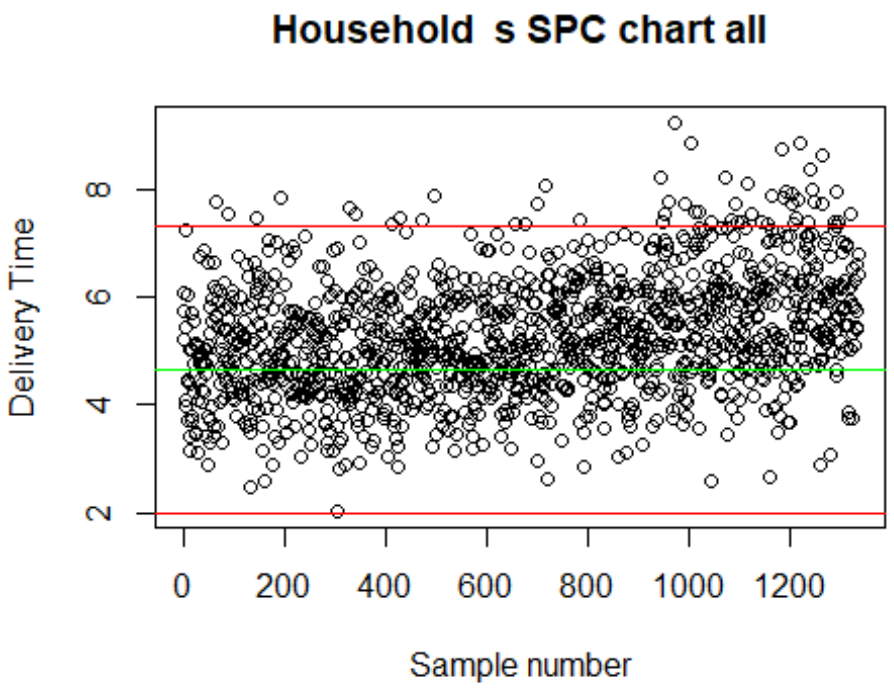
All samples

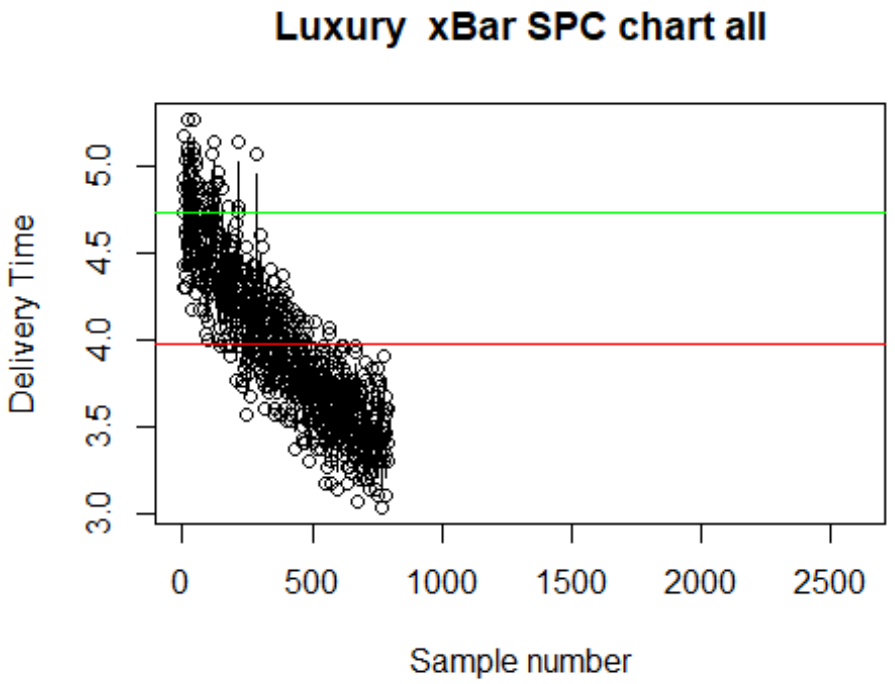
Out of control processes

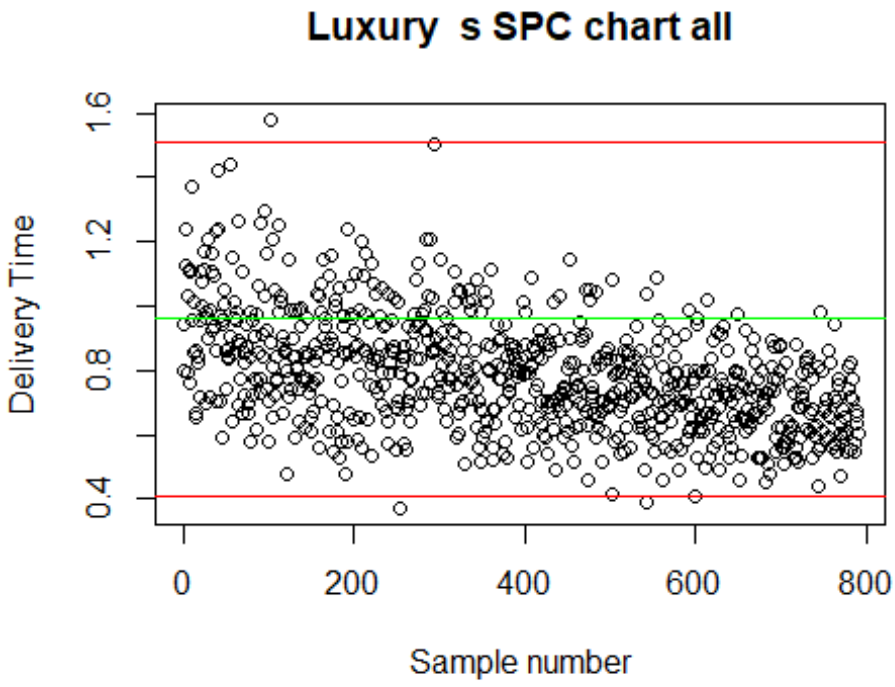





Household xBar SPC chart all
Delivery Time
45
50
55
0
500
1000
1500
2000
2500
Sample number

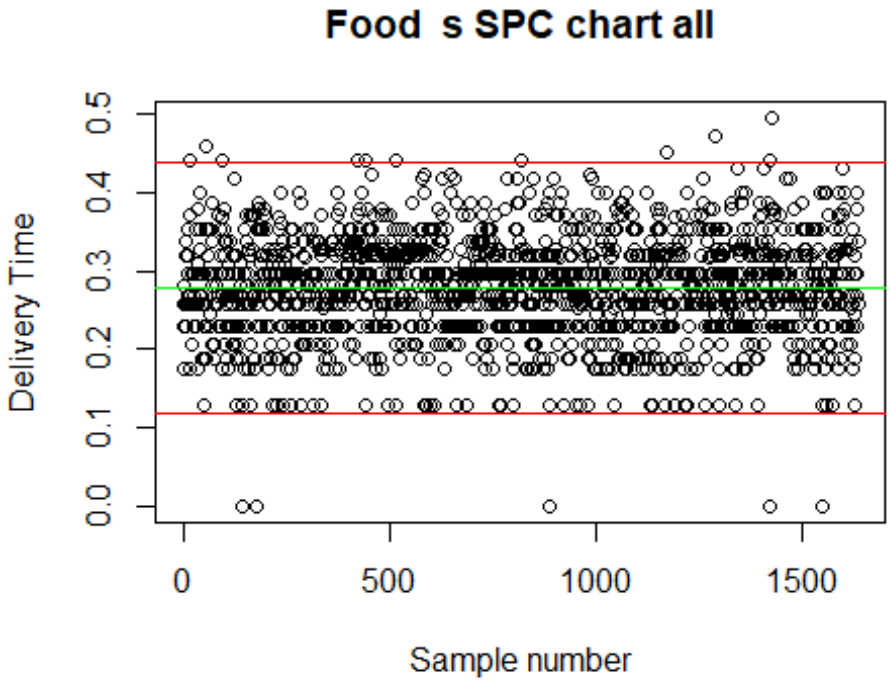

Household s SPC chart all
Delivery Time
2
4
6
8
0
200
400
600
800
1000
1200
Sample number

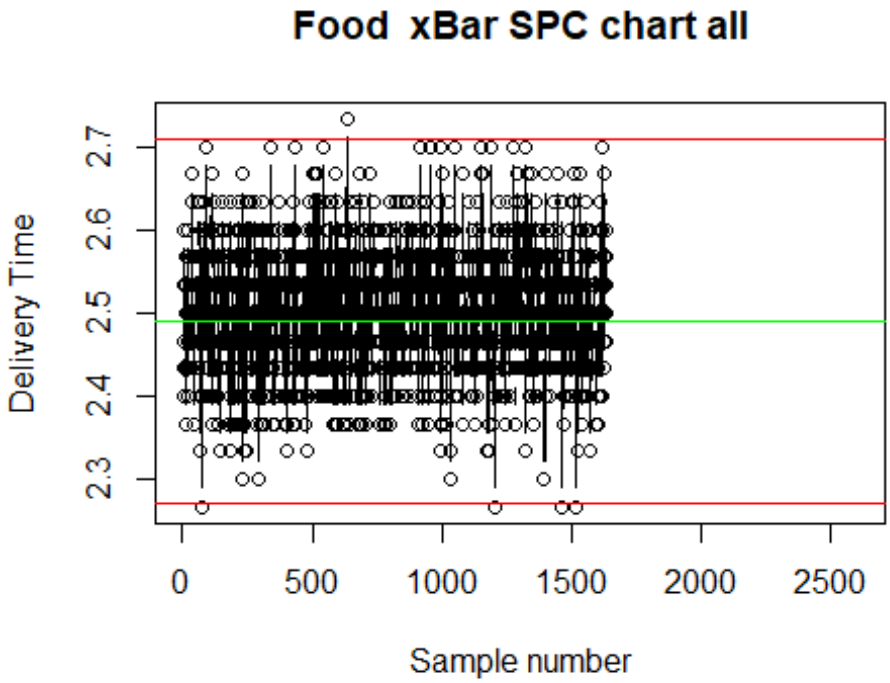


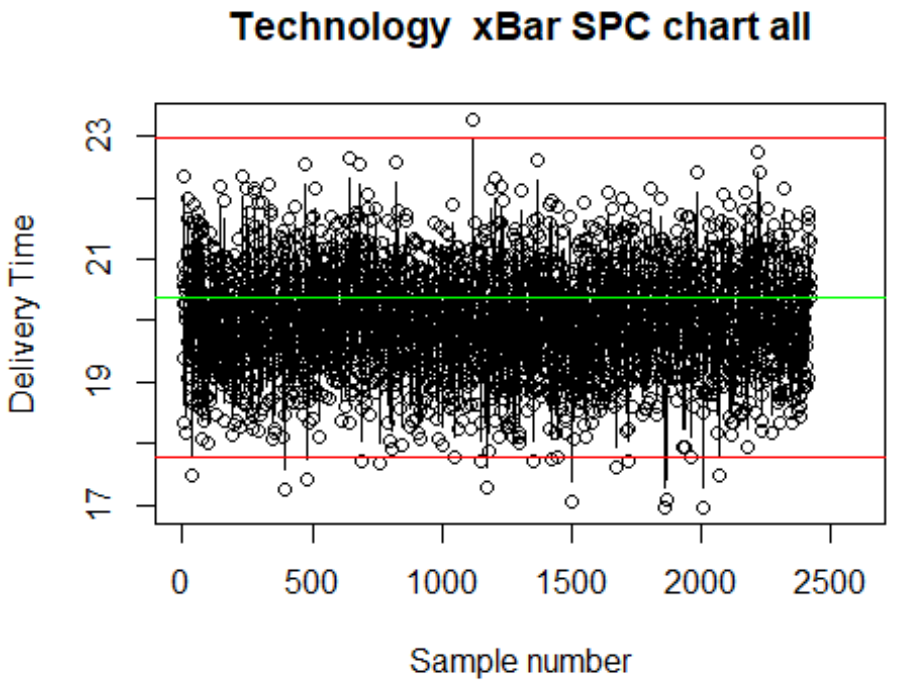


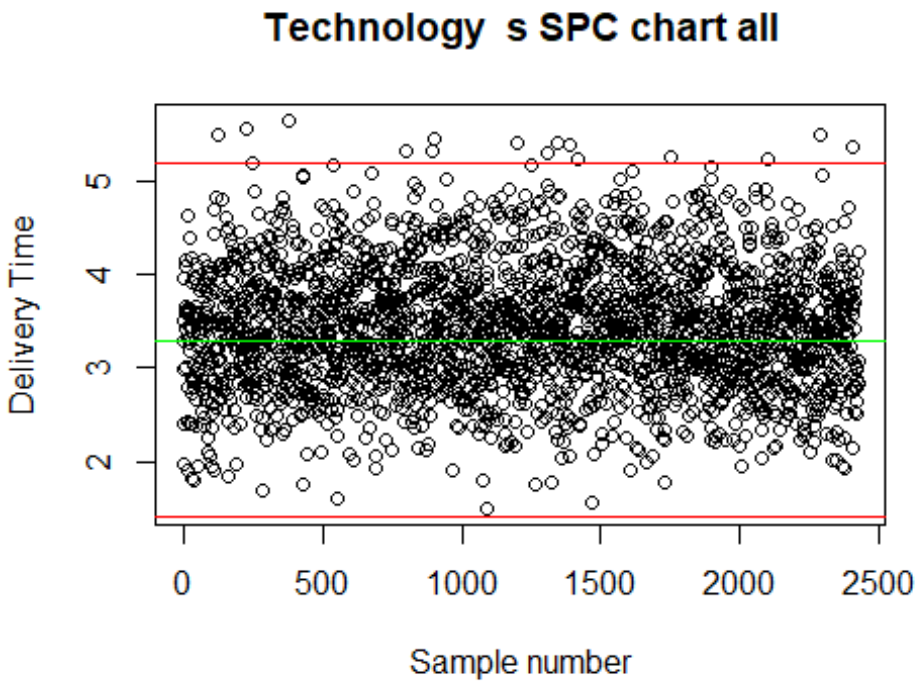
Luxury and Household items had a sudden shift in the process average. This could be due to a new process operator, a new inspector, a new machine setting, or a change in the production setup or method (R and William, 2010). Gifts seem to have a continuous shift in process average.

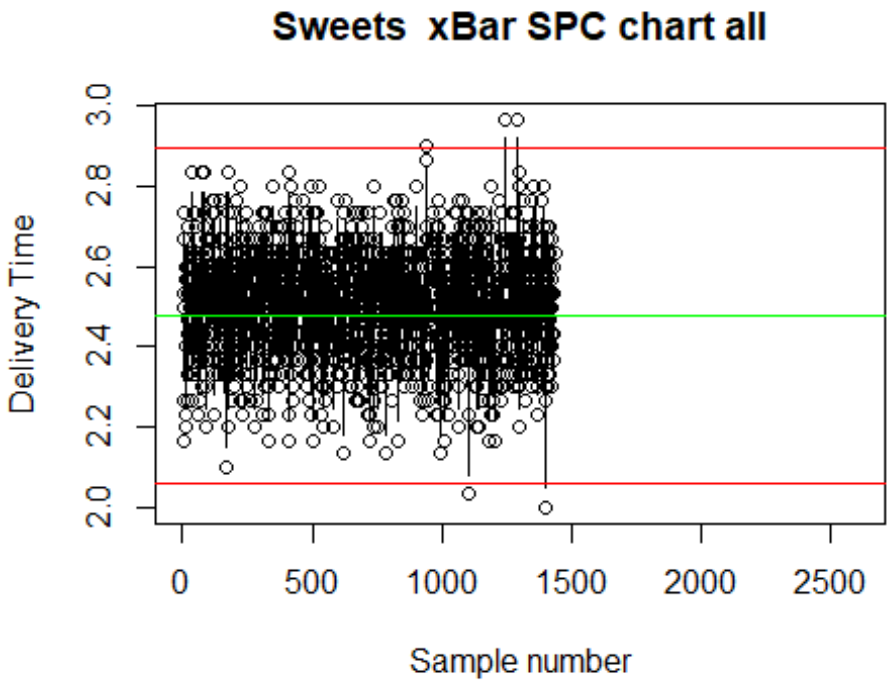
In control processes

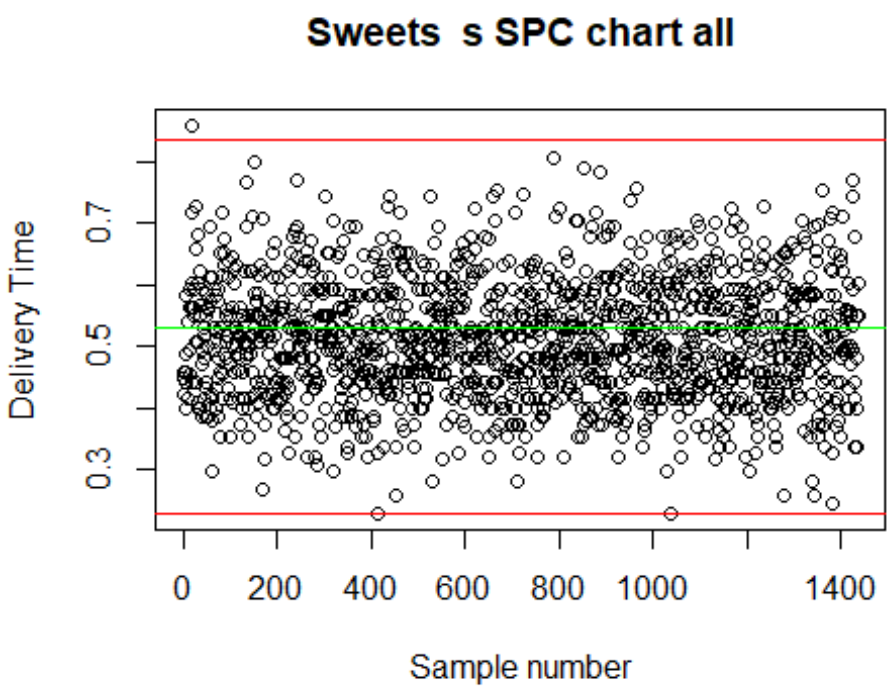


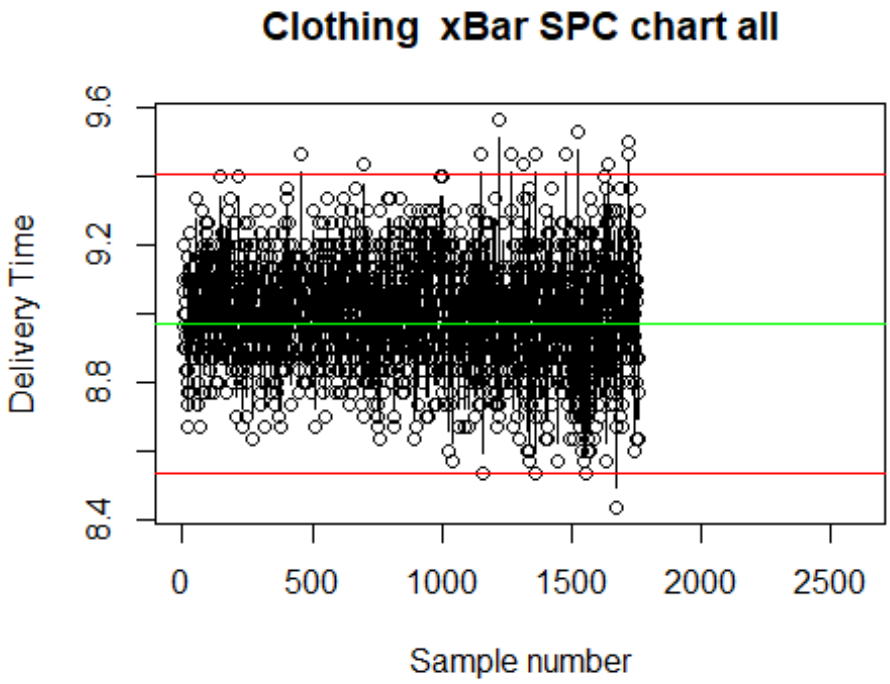


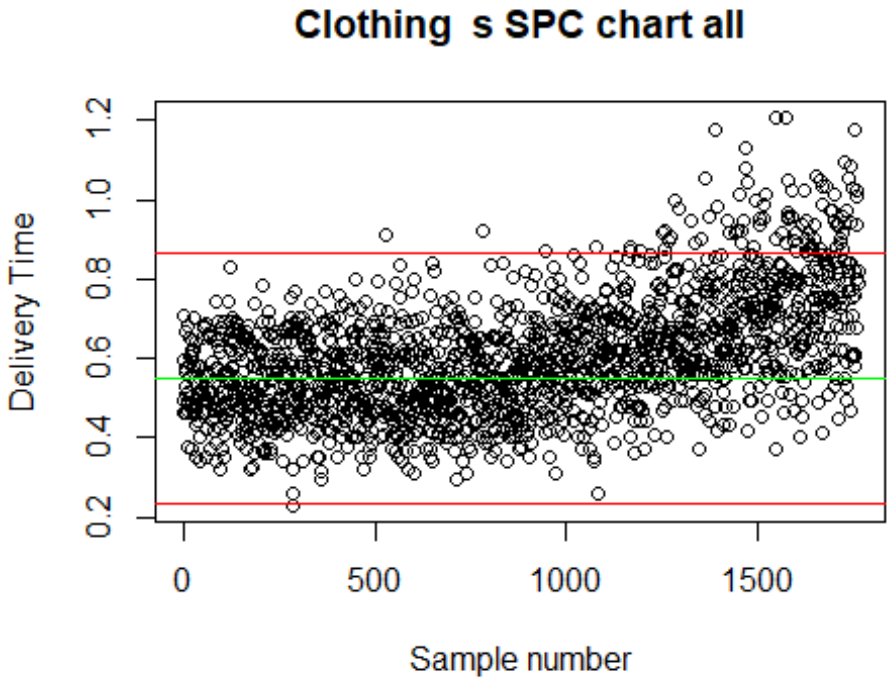

Technology xBar SPC chart all
Delivery Time
17
19
21
23
0
500
1000
1500
2000
2500
Sample number


Technology s SPC chart all
Delivery Time
2
3
4
5
0
500
1000
1500
2000
2500
Sample number


Sweets xBar SPC chart all
Delivery Time
2.0
2.2
2.4
2.6
2.8
3.0
0
500
1000
1500
2000
2500
Sample number


Sweets s SPC chart all
Delivery Time
0.3
0.5
0.7
0
200
400
600
800
1000
1400
Sample number





Food, Technology, Sweets, and clothing seem to be in control. Clothing should be monitored for a sudden shift in process average however due to more than 8 consecutive points fell on the top side of the chart.

Optimizing the delivery process

4.1 Sample means outside Control limits

	Total	1st	2nd	3rd	3rd last	2nd last	last
Clothing	17	9.467	9.433	9.467	8.533	8.533	8.433
Household	2156	50.867	50.867	50.933	NA	NA	NA
Food	1	2.733	NA	NA	NA	NA	NA
Technology	1761	23.267	18.333	20.7	19.867	19.6	19.133
Sweets	3	2.9	2.967	2.967	NA	NA	NA
Gifts	2290	10.233	9.667	9.7	16.567	16.3	16.033
Luxury	0	NA	NA	NA	NA	NA	NA

4.2 The most consecutive samples of "s-bar or sample standard deviations" between -0.3 and +0.4 sigma-control limits is **6** and the ending sample number is 1746.

Type I (Manufacturer's) Error:

$$H_0: \mu = CL$$

$$H_a: \mu \neq CL$$

We can convert our samples to the standard normal distribution where mean = 0 and standard deviation = 1. The natural limits will then be LCL = -3 and UCL = 3.

$$Z = \frac{X - \mu}{sd}$$

$$Z = \frac{3}{1} = 1$$

$$P(Z) = 0.9987$$

Type I error = (1-P(Z))*2 = 0.0027

Manova

For the manova test the significance of age and delivery.time in regard to price was tested.

```
 Response 1 :
               Df   Sum Sq Mean Sq F value    Pr(>F)
Price           1  1485839 1485839  3646.7 < 2.2e-16 ***
Residuals   179976 73331231     407
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 2 :
               Df   Sum Sq Mean Sq F value    Pr(>F)
Price           1   304294  304294    1576 < 2.2e-16 ***
Residuals   179976 34748723     193
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the results we can conclude that they both have high significance. This means that more expensive items take longer to deliver and is usually bought by older people.

Reliability of the service and products.

Problem 6: $L = k(y-m)^2 = k(0.04)^2 = 45$

$k = 18750$

Now Taguchi loss function: $L = 18750(y-m)^2$

Problem 7a:

$L = k(y-m)^2 = k(0.04)^2 = 35$

$k = 21875$

Now Taguchi loss function: $L = 21875(y-m)^2$

Problem 7b:

$L = k(y-m)^2$

$L = 21875(0.027)^2$

= $15.95

Problem 27a:

Rseries = $R_A \times R_B \times R_C$

Rseries = $0.85 \times 0.92 \times 0.9 = 0.7038$

= 70.38% reliability

Problem 27b:

$R_{parallel} = [1 - (1 - R_A)(1 - R_A)] \times [1 - (1 - R_B)(1 - R_B)] \times [1 - (1 - R_C)(1 - R_C)]$

$= [1 - (0.15)(0.15)] \times [1 - (0.08)(0.08)] \times [1 - (0.1)(0.1)]$

$= 0.9775 \times 0.9936 \times 0.99$

$= 0.9615$

= 96.15% reliability

Conclusion

This report will detail the processes and methods used to do statistical analysis on a dataset populated with sales and their respective features. These methods include statistical process control charts and control limit calculations. Processes were classified as in control or out of control. A manova hypothesis test was also conducted.

Reference list

R, J. and William, L. (2010). *Managing for Quality and Performance Excellence (Book Only)*. South Western Educational Publishing.